

## **CHƯƠNG 5**

### **PHÂN TÍCH TƯƠNG QUAN VÀ HỒI QUI**

#### **5.1 Những khái niệm mở đầu**

Trong thực tế nghiên cứu khí tượng, khí hậu có không ít những vấn đề được đặt ra trong đó cần phải xác định được qui luật biến đổi của các hiện tượng khí quyển. Tuy nhiên, hiện tượng khí quyển lại được phản ánh thông qua các đặc trưng yếu tố khí quyển mà chúng, đến lượt mình, lại phụ thuộc vào sự biến đổi của các nhân tố bên ngoài. Muốn nắm được qui luật biến đổi của các hiện tượng khí quyển cần thiết phải xác định sự liên hệ giữa các đặc trưng yếu tố khí quyển (được xem là biến phụ thuộc) với tập hợp các nhân tố ảnh hưởng mà người ta gọi là các biến độc lập. Điều đó cũng có nghĩa là, về phương diện thống kê, thông thường ta cần phải giải quyết một số vấn đề sau đây:

- 1) Xác định sự phân bố không gian của các đặc trưng yếu tố khí tượng, khí hậu, tức là nghiên cứu qui luật phụ thuộc vào toạ độ không gian của các biến khí quyển.
- 2) Xác định qui luật, tính chất diễn biến theo thời gian của các đặc trưng yếu tố khí quyển.
- 3) Xác định mối quan hệ ràng buộc để từ đó tìm qui luật liên hệ giữa các đặc trưng yếu tố khí quyển với nhau theo không gian và thời gian.

Một trong những phương pháp giải quyết các vấn đề đó là phương pháp phân tích tương quan và hồi qui mà nội dung của nó có thể được chia thành:

- 1) Tương quan và hồi qui theo không gian: Là xét mối quan hệ giữa hai hay nhiều biến khí quyển với nhau của cùng một yếu tố, cùng thời gian (đồng thời) nhưng khác nhau về vị trí không gian.
- 2) Tương quan và hồi qui theo thời gian: Là xét mối quan hệ giữa hai hay nhiều biến khí quyển với nhau của cùng một yếu tố, cùng một địa điểm nhưng khác nhau về thời gian.
- 3) Tương quan và hồi qui phổ biến: Là xét mối quan hệ giữa hai hay nhiều biến khí quyển của một hoặc nhiều yếu tố, có thể khác nhau về không gian, thời gian hoặc cả không-thời gian.

Về phương diện toán học, căn cứ vào dạng thức của biểu thức biểu diễn, người ta chia sự quan hệ tương quan làm bốn dạng:

- 1) Tương quan và hồi qui tuyến tính một biến: Xét mối quan hệ tương quan và hồi qui tuyến tính giữa một bên là biến phụ thuộc với một bên là một biến độc lập.
- 2) Tương quan và hồi qui phi tuyến một biến: Xét mối quan hệ tương quan và hồi qui phi tuyến giữa một bên là biến phụ thuộc với một bên là một biến độc lập.
- 3) Tương quan và hồi qui tuyến tính nhiều biến: Xét mối quan hệ tương quan và hồi qui tuyến tính giữa một bên là biến phụ thuộc với một bên là tập hợp nhiều biến độc lập.
- 4) Tương quan và hồi qui phi tuyến nhiều biến: Xét mối quan hệ tương quan và hồi qui phi tuyến giữa một bên là biến phụ thuộc với một bên là tập hợp nhiều biến độc lập.

Thông thường để giải quyết các bài toán tương quan và hồi qui trong khí tượng, khí hậu cần phải tiến hành các bước sau:

- 1) Xác lập được dạng thức của mối liên hệ tương quan, tức là tìm ra dạng hồi qui thích hợp: Tuyến tính hay phi tuyến, nếu là phi tuyến thì cụ thể là dạng nào.
- 2) Đánh giá được mức độ chặt chẽ của các mối liên hệ theo nghĩa quan hệ tương quan.
- 3) Bằng phương pháp nào đó, xác lập biểu thức giải tích của phương trình hồi qui xấp xỉ mối liên hệ tương quan, tức là xây dựng hàm hồi qui. Trong khí tượng, khí hậu phương pháp phổ biến để xây dựng hàm hồi qui là phương pháp bình phương tối thiểu.
- 4) Đánh giá độ chính xác và khả năng sử dụng của phương trình hồi qui.

## 5.2 Tương quan tuyến tính

### 5.2.1 Hệ số tương quan tổng thể

Xét hai biến ngẫu nhiên  $X_1$  và  $X_2$ . Khi đó phương sai của tổng (hiệu) hai biến được xác định bởi:

$$\begin{aligned} D[X_1 \pm X_2] &= M[(X_1 \pm X_2) - M(X_1 \pm X_2)]^2 = M[(X_1 - MX_1) \pm (X_2 - MX_2)]^2 = \\ &= M[(X_1 - MX_1)^2] + M[(X_2 - MX_2)^2] \pm 2M[(X_1 - MX_1)(X_2 - MX_2)] = \\ &= D[X_1] + D[X_2] \pm 2 M[(X_1 - MX_1)(X_2 - MX_2)] = \\ &= \mu_{11} + \mu_{22} \pm 2\mu_{12} \end{aligned}$$

trong đó  $\mu_{12}$  là mômen tương quan giữa  $X_1$  và  $X_2$ ,  $\mu_{11}$  và  $\mu_{22}$  tương ứng là phương sai của  $X_1$  và  $X_2$ . Nếu  $X_1$  và  $X_2$  không tương quan với nhau thì:

$$D[X_1 \pm X_2] = D[X_1] + D[X_2], \text{ suy ra } \mu_{12} = 0.$$

Do vậy, người ta dùng  $\mu_{12}$  làm thước đo mức độ tương quan giữa  $X_1$  và  $X_2$ . Vì  $\mu_{12}$  là một đại lượng có thứ nguyên (bằng tích thứ nguyên của  $X_1$  và  $X_2$ ) nên để thuận tiện trong việc so sánh, phân tích thay cho  $\mu_{12}$  người ta dùng đại lượng vô thứ nguyên:

$$\rho_{12} = \frac{\mu_{12}}{\sqrt{\mu_{11}\mu_{22}}} \quad (5.2.1)$$

và được gọi là hệ số tương quan giữa hai biến  $X_1$  và  $X_2$ . Người ta gọi  $\rho_{12}$  là hệ số tương quan tổng thể hay hệ số tương quan lý thuyết và là một hằng số.

Hệ số tương quan có các tính chất sau đây:

1) Hệ số tương quan nhận giá trị trên đoạn  $[-1;1]$ :  $-1 \leq \rho_{12} \leq 1$ .

Thật vậy, ta có:

$$\begin{aligned} D\left[\frac{X_1}{\sqrt{DX_1}} \pm \frac{X_2}{\sqrt{DX_2}}\right] &= \left[\left(\frac{X_1}{\sqrt{DX_1}} - M\left[\frac{X_1}{\sqrt{DX_1}}\right]\right) \pm \left(\frac{X_2}{\sqrt{DX_2}} - M\left[\frac{X_2}{\sqrt{DX_2}}\right]\right)\right]^2 = \\ &= D\left[\frac{X_1}{\sqrt{DX_1}}\right] + D\left[\frac{X_2}{\sqrt{DX_2}}\right] \pm 2M\left[\left(\frac{X_1}{\sqrt{DX_1}} - M\left[\frac{X_1}{\sqrt{DX_1}}\right]\right)\left(\frac{X_2}{\sqrt{DX_2}} - M\left[\frac{X_2}{\sqrt{DX_2}}\right]\right)\right] \\ &= \frac{1}{DX_1} DX_1 + \frac{1}{DX_2} DX_2 \pm 2 \frac{1}{\sqrt{DX_1 DX_2}} \mu_{12} = 2 \pm 2 \frac{\mu_{12}}{\sqrt{\mu_{11}\mu_{22}}} = 2(1 \pm \rho_{12}) \geq 0 \end{aligned}$$

Hay  $1 \pm \rho_{12} \geq 0 \Rightarrow \text{đpcm}$

2) Điều kiện cần và đủ để  $|\rho_{12}|=1$  là  $X_1$  và  $X_2$  có quan hệ hàm tuyến tính.

*Điều kiện đủ:*

Giả sử ta có quan hệ hàm tuyến tính giữa  $X_1$  và  $X_2$ :  $X_2 = a + bX_1$ , với  $a, b$  là các hệ số hằng số. Khi đó:

$$\begin{aligned} \mu_{12} &= M[(X_1 - MX_1)(X_2 - MX_2)] = M[(X_1 - MX_1)(a + bX_1 - a - bMX_1)] = \\ &= M[b(X_1 - MX_1)^2] = b\mu_{11} \end{aligned}$$

$$\mu_{22} = M[(X_2 - MX_2)^2] = M[(a + bX_1 - a - bMX_1)^2] = b^2 M[(X_1 - MX_1)^2] = b^2 \mu_{11}$$

$$\text{Vậy } \rho_{12} = \frac{\mu_{12}}{\sqrt{\mu_{11}\mu_{22}}} = \frac{b\mu_{11}}{\sqrt{b^2\mu_{11}^2}} = \frac{b}{|b|} = \begin{cases} 1 & \text{kh } b > 0 \\ -1 & \text{kh } b < 0 \end{cases}$$

*Điều kiện cần:*

$$\text{Từ hệ thức } D\left[\frac{X_1}{\sqrt{DX_1}} \pm \frac{X_2}{\sqrt{DX_2}}\right] = 2(1 \pm \rho_{12}) \text{ ta có:}$$

$$\text{Nếu } (1 \pm \rho_{12}) = 0 \text{ thì } \left[ \frac{X_1}{\sqrt{DX_1}} \pm \frac{X_2}{\sqrt{DX_2}} \right] = C = \text{Const}$$

Từ đó suy ra  $X_2 = \pm \sqrt{\frac{\mu_{22}}{\mu_{11}}} X_1 + C \sqrt{\mu_{22}}$ , tức là giữa  $X_2$  và  $X_1$  tồn tại quan hệ hàm tuyến tính.

Do tính chất này nên hệ số tương quan được xem là đại lượng đặc trưng cho mức độ tương quan tuyến tính giữa hai biến.

### 5.2.2 Hệ số tương quan mẫu

Cho hai biến ngẫu nhiên  $X_1, X_2$  với  $n$  cặp trị số quan sát:

$$\{x_{t1}, x_{t2}\} = \{(x_{11}, x_{12}), (x_{21}, x_{22}), \dots, (x_{n1}, x_{n2})\}$$

Khi đó mômen tương quan mẫu – ước lượng của mômen tương quan tổng thể  $\mu_{12}$  – giữa  $X_1$  và  $X_2$  được xác định bởi:

$$R_{12} = \frac{1}{n} \sum_{t=1}^n (x_{t1} - \bar{x}_1)(x_{t2} - \bar{x}_2) = \overline{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)} \quad (5.2.2)$$

và hệ số tương quan mẫu:

$$r_{12} = \frac{\frac{1}{n} \sum_{t=1}^n (x_{t1} - \bar{x}_1)(x_{t2} - \bar{x}_2)}{\sqrt{\frac{1}{n} \sum_{t=1}^n (x_{t1} - \bar{x}_1)^2} \sqrt{\frac{1}{n} \sum_{t=1}^n (x_{t2} - \bar{x}_2)^2}} = \frac{l_{12}}{\sqrt{l_{11}l_{22}}} \quad (5.2.3)$$

trong đó:

$$l_{12} = \sum_{t=1}^n (x_{t1} - \bar{x}_1)(x_{t2} - \bar{x}_2) = nR_{12} \text{ là tổng của tích các độ lệch của } X_1 \text{ và } X_2 \text{ so với}$$

trung bình của chúng.

$$l_{11} = \sum_{t=1}^n (x_{t1} - \bar{x}_1)^2 = n s_1^2 - \text{tổng bình phương các độ lệch của } X_1 \text{ so với trung bình}$$

của nó.

$$l_{22} = \sum_{t=1}^n (x_{t2} - \bar{x}_2)^2 = n s_2^2 - \text{tổng bình phương các độ lệch của } X_2 \text{ so với trung bình}$$

của nó.

$$\bar{x}_1 = \frac{1}{n} \sum_{t=1}^n x_{t1}, \quad \bar{x}_2 = \frac{1}{n} \sum_{t=1}^n x_{t2} - \text{trung bình của } X_1 \text{ và } X_2$$

Hệ số tương quan mẫu  $r_{12}$  là ước lượng của hệ số tương quan tổng thể  $\rho_{12}$ . Nếu  $\rho_{12}$  là một hằng số thì trái lại  $r_{12}$  là một đại lượng ngẫu nhiên. Năm 1915

R.A.Fisher [3,5,6] đã tìm ra biểu thức chính xác của hàm mật độ xác suất của hệ số tương quan mẫu  $r_{12}$  trong trường hợp phân bố đồng thời của  $X_1$  và  $X_2$  là chuẩn:

$$f_n(r) = \frac{2^{n-3}}{\pi \Gamma(n-2)} (1-\rho^2)^{\frac{n-1}{2}} (1-r^2)^{\frac{n-4}{2}} \sum_{i=0}^{\infty} \left( \Gamma\left(\frac{n+i-1}{2}\right) \right)^2 \frac{(2\rho r)^i}{i!}, \quad (5.2.4)$$

với  $-1 \leq r \leq 1$ . Ở đây, để tiện biểu diễn ta đã thay ký hiệu  $r_{12}$  bằng ký hiệu  $r$ .

Bằng phép biến đổi chuỗi lũy thừa vế phải của biểu thức  $f_n(r)$  người ta đã thu được dạng khác đối với mật độ xác suất của  $r$ :

$$f_n(r) = \frac{n-2}{\pi} (1-\rho^2)^{\frac{n-1}{2}} (1-r^2)^{\frac{n-4}{2}} \int_0^1 \frac{x^{n-2}}{(1-\rho r x)^{n-1}} \frac{dx}{\sqrt{1-x^2}} \quad (5.2.5)$$

Ta thấy rằng phân bố của  $r$  chỉ phụ thuộc vào dung lượng mẫu  $n$  và hệ số tương quan tổng thể  $\rho$ . Khi  $n = 2$  thì  $f_n(r) = 0$ , điều đó phù hợp với sự kiện hệ số tương quan được tính từ tập mẫu chỉ có 2 quan trắc phải bằng  $\pm 1$ .

Kỳ vọng của hệ số tương quan mẫu  $r$ :  $M[r] = \rho$

Phương sai của hệ số tương quan mẫu  $r$ :

$$D[r] = \frac{\rho^2}{4n} \left( \frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2} + \frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{4\mu_{22}}{\mu_{11}^2} - \frac{4\mu_{31}}{\mu_{11}\mu_{20}} - \frac{4\mu_{13}}{\mu_{11}\mu_{02}} \right)$$

trong đó  $\mu_{ij} = M[(X_1 - MX_1)^i (X_2 - MX_2)^j]$  – các mômen trung tâm bậc  $i+j$ .

Để thuận tiện trong tính toán thực hành, nhất là việc ước lượng khoảng cho  $\rho$ , người ta thường dùng phép biến đổi sau đây của Fisher:

$$z = \frac{1}{2} \log \frac{1+r}{1-r}, \quad \zeta = \frac{1}{2} \log \frac{1+\rho}{1-\rho} \quad (5.2.6)$$

Fisher đã chứng minh được rằng ngay cả với những giá trị  $n$  không lớn lắm biến  $z$  cũng phân bố xấp xỉ chuẩn với giá trị trung bình và phương sai được cho bởi biểu thức gần đúng sau:

$$M[z] = \zeta + \frac{\rho}{2(n-1)}, \quad D[z] = \frac{1}{n-3} \quad (5.2.7)$$

Vì vậy khoảng tin cậy của  $\zeta$  với độ tin cậy  $1-\alpha$  là:

$$\left( z - \frac{r}{2(n-1)} - u_\alpha \frac{1}{\sqrt{n-3}}, z - \frac{r}{2(n-1)} + u_\alpha \frac{1}{\sqrt{n-3}} \right) \quad (5.2.8)$$

trong đó  $u_\alpha$  nhận được từ phân bố chuẩn  $N(0,1)$  bởi hệ thức:  $P(|u| \geq u_\alpha) = \alpha$ . Từ đó ta nhận được khoảng tin cậy của  $\rho$ .

Trong trường hợp  $\rho = 0$  thì biến  $t = r \sqrt{\frac{n-2}{1-r^2}}$  có phân bố Student với  $n-2$  bậc tự do. Hệ số tương quan mẫu  $r$  là ước lượng vững nhưng chệch của hệ số tương quan tổng thể  $\rho$  với độ chệch bằng  $\frac{-\rho(1-\rho^2)}{2n}$ . Do đó khi tính toán thực hành nếu nhận được  $r = 0$  thì điều đó không có nghĩa là  $\rho$  bằng 0. Và ngược lại, nếu  $r \neq 0$  thì cũng không hẳn là  $\rho$  khác 0. Nếu dung lượng mẫu nhỏ thì mặc dù  $\rho = 0$  nhưng giá trị của  $r$  lại có thể có ý nghĩa. Vì vậy ta cần kiểm tra xem độ lớn của  $r$  có ý nghĩa thực sự hay không, hay nói cách khác cần kiểm nghiệm độ rõ rệt của  $r$ .

Để kiểm nghiệm, ta đặt giả thiết  $H_0: \rho = 0$ . Thay  $\rho \approx r$ , với giới hạn tin cậy ban đầu  $d$  thì khi  $H_0$  đúng ta có  $P(|r| \geq d) = \alpha$ .

$$\text{Đặt} \quad t = \frac{r}{\sqrt{1-r^2} / \sqrt{n-2}}, \quad t_\alpha = \frac{d}{\sqrt{1-r^2} / \sqrt{n-2}} \quad (5.2.9)$$

Khi đó nếu  $H_0$  đúng thì:  $P(|t| \geq t_\alpha) = \alpha$ . Biến  $t$  trong (5.2.9) có phân bố Student ( $t$ ) với  $n-2$  bậc tự do. Từ đó ta xác định được  $t_\alpha$ . Và chỉ tiêu kiểm nghiệm sẽ là:

Nếu  $|t| \geq t_\alpha$  thì bác bỏ  $H_0$  và đưa ra kết luận  $r$  lớn rõ rệt

Nếu  $|t| < t_\alpha$  thì chấp nhận  $H_0$  và kết luận  $r$  không lớn rõ rệt.

**Ví dụ 5.2.1** Từ tập mẫu  $\{x_t, y_t, t=1..11\}$  ta tính được hệ số tương quan  $r_{xy}=0.76$ . Hãy cho biết với giá trị nhận được như vậy thì hệ số tương quan có lớn rõ rệt không nếu lấy mức ý nghĩa  $\alpha=0.01$ ?

Để trả lời câu hỏi đặt ra ta cần kiểm nghiệm giả thiết:  $H_0: r_{xy}=0$ . Muốn vậy, ta tính đại lượng  $t = \frac{r_{xy}}{\sqrt{1-r^2} / \sqrt{n-2}} = \frac{0.76}{\sqrt{1-0.76^2} / \sqrt{11-2}} = 3.51$ . Từ  $\alpha=0.01$  ta xác định được

$t_\alpha$  từ phân bố Student:  $t_\alpha = \text{St}(11-2, 0.01) = 3.25$ .

Vì  $|t|=3.51 > 3.25=t_\alpha$  do đó ta bác bỏ giả thiết  $H_0$  và đưa ra kết luận  $r_{xy}$  lớn rõ rệt.

Ngoài việc kiểm tra độ rõ rệt của hệ số tương quan, trong thực tế người ta còn đánh giá sự có nghĩa của nó. Để xác định sự có nghĩa của  $r$  trước hết ta tính giá trị  $H = |r| \sqrt{n-1} \equiv H(n, r)$ . Tương ứng với các giá trị dung lượng mẫu  $n$  khác nhau, khi cho trước độ tin cậy  $p$ , tra bảng ta sẽ tính được trị số tới hạn  $H_0$  của  $H$ :  $H_0 = H(p, n)$ . Trong bảng 5.1 đã cho các giá trị tới hạn  $H_0$  ứng với các độ tin cậy  $p$  và dung lượng mẫu  $n$  khác nhau.

Từ đó chỉ tiêu kiểm nghiệm sự có nghĩa của  $r$  sẽ là:

Nếu  $H(n, r) > H_0(p, n)$  thì kết luận  $r$  có nghĩa với độ tin cậy  $p$

Nếu  $H(n, r) \leq H_0(p, n)$  thì kết luận  $r$  không có nghĩa với độ tin cậy  $p$ .

**Bảng 5.1** Giá trị tới hạn  $H_0(p,n)$ 

| n  | p    |      |      |       | n        | p     |       |       |
|----|------|------|------|-------|----------|-------|-------|-------|
|    | 0.90 | 0.95 | 0.99 | 0.999 |          | 0.95  | 0.99  | 0.999 |
| 10 | 1.65 | 1.90 | 2.29 | 2.62  | 25       | 1.941 | 2.475 | 3.026 |
| 11 | 1.65 | 1.90 | 2.32 | 2.68  | 26       | 1.941 | 2.479 | 3.037 |
| 12 | 1.65 | 1.92 | 2.35 | 2.73  | 27       | 1.492 | 2.483 | 3.047 |
| 13 | 1.65 | 1.92 | 2.37 | 2.77  | 28       | 1.943 | 2.487 | 3.056 |
| 14 | 1.65 | 1.92 | 2.39 | 2.81  | 29       | 1.493 | 2.490 | 3.064 |
| 15 | 1.65 | 1.92 | 2.40 | 2.85  | 30       | 1.944 | 2.492 | 3.071 |
| 16 | 1.65 | 1.93 | 2.41 | 2.87  | 35       | 1.947 | 2.505 | 3.102 |
| 17 | 1.65 | 1.93 | 2.42 | 2.90  | 40       | 1.949 | 2.514 | 3.126 |
| 18 | 1.65 | 1.93 | 2.43 | 2.92  | 45       | 1.950 | 2.521 | 3.145 |
| 19 | 1.65 | 1.93 | 2.44 | 2.94  | 50       | 1.951 | 2.527 | 3.161 |
| 20 | 1.65 | 1.94 | 2.45 | 2.96  | 60       | 1.953 | 2.535 | 3.830 |
| 21 | 1.65 | 1.94 | 2.45 | 2.98  | 70       | 1.954 | 2.541 | 3.190 |
| 22 | 1.65 | 1.94 | 2.46 | 2.99  | 80       | 1.955 | 2.546 | 3.209 |
| 23 | 1.65 | 1.94 | 2.47 | 3.00  | 90       | 1.956 | 2.550 | 3.219 |
| 24 | 1.65 | 1.94 | 2.47 | 3.02  | 100      | 1.956 | 2.553 | 3.226 |
|    |      |      |      |       | $\infty$ | 1.960 | 2.576 | 3.291 |

### 5.2.3 Cách tính hệ số tương quan mẫu

Cho hai biến ngẫu nhiên  $X_1, X_2$  với  $n$  cặp trị số quan sát:

$$\{x_{t1}, x_{t2}\} = \{(x_{11}, x_{12}), (x_{21}, x_{22}), \dots, (x_{n1}, x_{n2})\}$$

Từ tập mẫu này có thể tính hệ số tương quan giữa  $X_1, X_2$  theo các phương pháp sau đây.

#### 5.2.3.1 Phương pháp tính trực tiếp

Phương pháp trực tiếp tính hệ số tương quan mẫu là tính theo công thức (5.2.3). Thế nhưng, trong thực hành người ta thường biến đổi và đưa nó về dạng khác.

$$\begin{aligned}
 R_{12} &= \overline{(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)} = \overline{x_1 x_2 - \bar{x}_1 x_2 + x_2 x_1 - \bar{x}_1 x_2} = \overline{x_1 x_2 - \bar{x}_1 x_2} \\
 &= \overline{x_1 x_2} - \bar{x}_1 \bar{x}_2 = \frac{1}{n} \sum_{t=1}^n x_{t1} x_{t2} - \frac{1}{n} \sum_{t=1}^n x_{t1} \frac{1}{n} \sum_{t=1}^n x_{t2} \quad (5.2.10)
 \end{aligned}$$

$$\begin{aligned}
 s_1^2 &= \overline{(x_1 - \bar{x}_1)^2} = \overline{(x_1)^2 - 2x_1 \bar{x}_1 + (\bar{x}_1)^2} = \overline{(x_1)^2} - (\bar{x}_1)^2 \\
 &= \frac{1}{n} \sum_{t=1}^n (x_{t1})^2 - \left(\frac{1}{n} \sum_{t=1}^n x_{t1}\right)^2 \quad (5.2.11)
 \end{aligned}$$

Tương tự ta có:

$$s_2^2 = \frac{1}{n} \sum_{t=1}^n (x_{t2})^2 - \left( \frac{1}{n} \sum_{t=1}^n x_{t2} \right)^2 \quad (5.2.12)$$

$$\text{Kết hợp (5.2.10)–(5.2.12) ta nhận được: } r_{12} = \frac{R_{12}}{s_1 s_2} \quad (5.2.13)$$

Hoặc có thể tính theo công thức:

$$r_{12} = \frac{\sum_{t=1}^n x_{t1} x_{t2} - \frac{1}{n} \sum_{t=1}^n x_{t1} \sum_{t=1}^n x_{t2}}{\sqrt{\sum_{t=1}^n (x_{t1})^2 - \frac{1}{n} \left( \sum_{t=1}^n x_{t1} \right)^2} \sqrt{\sum_{t=1}^n (x_{t2})^2 - \frac{1}{n} \left( \sum_{t=1}^n x_{t2} \right)^2}} \quad (5.2.14)$$

**Ví dụ 5.2.2** Trong bảng 5.2 dẫn ra số liệu quan trắc tổng lượng mưa tháng 1 của hai trạm mà ta đặt chúng là hai biến  $X_1$ ,  $X_2$  và kết quả các bước tính trung gian theo công thức (5.2.14). Cột thứ nhất chỉ số thứ tự năm (t). Hai cột tiếp theo của bảng chứa số liệu hai chuỗi  $\{x_{t1}\}$  và  $\{x_{t2}\}$ . Cột thứ tư là tích từng cặp  $(x_{t1}, x_{t2})$ , hai cột cuối cùng chứa bình phương các giá trị  $x_{t1}$  và  $x_{t2}$ . Dòng cuối cùng của bảng là tổng theo từng cột.

**Bảng 5.2** Số liệu lượng mưa tháng 1 và những kết quả tính trung gian

| t    | $x_{t1}$ | $x_{t2}$ | $x_{t1}x_{t2}$ | $(x_{t1})^2$ | $(x_{t2})^2$ |
|------|----------|----------|----------------|--------------|--------------|
| 1    | 10.6     | 19.1     | 202.46         | 112.36       | 364.81       |
| 2    | 0.9      | 11.8     | 10.62          | 0.81         | 139.24       |
| 3    | 9.6      | 86.9     | 834.24         | 92.16        | 7551.61      |
| 4    | 2.0      | 16.4     | 32.80          | 4.00         | 268.96       |
| 5    | 38.3     | 12.4     | 474.92         | 1466.89      | 153.76       |
| 6    | 0.9      | 9.6      | 8.64           | 0.81         | 92.16        |
| 7    | 46.7     | 26.8     | 1251.56        | 2180.89      | 718.24       |
| 8    | 142.5    | 48.7     | 6939.75        | 20306.25     | 2371.69      |
| 9    | 68.2     | 28.9     | 1970.98        | 4651.24      | 835.21       |
| 10   | 54.1     | 87.4     | 4728.34        | 2926.81      | 7638.76      |
| 11   | 25.9     | 66.1     | 1711.99        | 670.81       | 4369.21      |
| 12   | 41.3     | 42.7     | 1763.51        | 1705.69      | 1823.29      |
| 13   | 11.8     | 37.7     | 444.86         | 139.24       | 1421.29      |
| 14   | 5.0      | 55.1     | 275.50         | 25.00        | 3036.01      |
| 15   | 30.0     | 104.1    | 3123.00        | 900.00       | 10836.81     |
| 16   | 21.8     | 33.9     | 739.02         | 475.24       | 1149.21      |
| 17   | 26.0     | 39.0     | 1014.00        | 676.00       | 1521.00      |
| 18   | 6.0      | 38.0     | 228.00         | 36.00        | 1444.00      |
| 19   | 15.0     | 116.0    | 1740.00        | 225.00       | 13456.00     |
| Tổng | 556.6    | 880.6    | 27494.19       | 36595.20     | 59191.26     |

Đối sánh với từng thành phần trong (5.2.14) ta có:  $n=19$

$$\sum_{t=1}^n x_{t1}x_{t2} = 27494.19, \quad \frac{1}{n} \sum_{t=1}^n x_{t1} \sum_{t=1}^n x_{t2} = 556.6 * 880.6 / 19 = 25796,$$

$$\sum_{t=1}^n (x_{t1})^2 = 36595.20, \quad \frac{1}{n} (\sum_{t=1}^n x_{t1})^2 = 16305.45$$

$$\sum_{t=1}^n (x_{t2})^2 = 59191.26, \quad \frac{1}{n} (\sum_{t=1}^n x_{t2})^2 = 40813.49$$

Sau khi thay vào và tính ra ta được  $r_{12}=0.087894$ .

### 5.2.3.2 Phương pháp biến đổi tương đương.

Khi giá trị của các thành phần trong chuỗi khá lớn việc tính toán trực tiếp theo các công thức (5.2.10)–(5.2.14) thường gặp trở ngại, phức tạp và dễ gây sai số, nhất là quá trình tính toán được tiến hành thủ công. Do đó, trong nhiều trường hợp, để đơn giản ta sử dụng phép biến đổi sau đây:

$$y_{t1} = d_1 x_{t1} - C_1 \quad (*)$$

$$y_{t2} = d_2 x_{t2} - C_2 \quad (**)$$

trong đó  $d_1, d_2, C_1, C_2$  là những hằng số nào đó, mà trong những trường hợp cụ thể, sẽ được chọn sao cho thích hợp. Chẳng hạn, khi xử lý chuỗi số liệu nhiệt độ ta thấy chúng thường dao động xung quanh trị số 20 ( $^{\circ}\text{C}$ ), vậy có thể chọn  $C=20$ ; các giá trị khí áp thường lên xuống quanh giá trị 1000 (mb) thì chọn  $C=1000, \dots$

Với phép biến đổi (\*), (\*\*) ta có:

$$x_{t1} = \frac{y_{t1} + C_1}{d_1}, \quad x_{t2} = \frac{y_{t2} + C_2}{d_2}$$

Hay 
$$\bar{x}_1 = \frac{\bar{y}_1 + C_1}{d_1}, \quad \bar{x}_2 = \frac{\bar{y}_2 + C_2}{d_2}$$

Suy ra 
$$l_{12} = \sum \left( \frac{y_{t1} + C_1}{d_1} - \frac{\bar{y}_1 + C_1}{d_1} \right) \left( \frac{y_{t2} + C_2}{d_2} - \frac{\bar{y}_2 + C_2}{d_2} \right)$$
  

$$= \frac{1}{d_1 d_2} \sum (y_{t1} - \bar{y}_1)(y_{t2} - \bar{y}_2) = \frac{l'_{12}}{d_1 d_2}$$

Tương tự ta được: 
$$l_{11} = \frac{l'_{11}}{d_1^2}, \quad l_{22} = \frac{l'_{22}}{d_2^2}$$

Do đó: 
$$r_{12} = \frac{l_{12}}{\sqrt{l_{11} l_{22}}} = \frac{\frac{1}{d_1 d_2} l'_{12}}{\frac{1}{d_1 d_2} \sqrt{l'_{11} l'_{22}}} = \frac{l'_{12}}{\sqrt{l'_{11} l'_{22}}} = r'_{12} \quad (5.2.15)$$

Như vậy, qua phép biến đổi (\*) và (\*\*) hệ số tương quan vẫn không bị thay đổi.

#### 5.2.4 Ma trận tương quan

Trong thực tế ta thường gặp những bài toán mà ở đó đòi hỏi phải khảo sát mối quan hệ tương quan giữa các biến khác nhau của một tập nhiều hơn hai biến. Khi đó ta không chỉ có một hệ số tương quan mà là một ma trận tương quan.

Xét tập hợp  $m$  biến ngẫu nhiên  $X_1, X_2, \dots, X_m$ . Hệ số tương quan tổng thể giữa các biến  $X_j$  và  $X_k$  được xác định bởi hệ thức:

$$\rho_{jk} = \frac{\mu_{jk}}{\sqrt{\mu_{jj}\mu_{kk}}}, j, k=1..m \quad (5.2.16)$$

trong đó  $\mu_{jk}$  là mômen tương quan giữa  $X_j$  và  $X_k$ ,  $\mu_{jj}$  là phương sai của  $X_j$ . Tập hợp các hệ số tương quan  $\rho_{jk}$  lập thành ma trận tương quan:

$$(\rho_{jk}) = \begin{pmatrix} \rho_{11} & \dots & \rho_{1m} \\ \dots & \dots & \dots \\ \rho_{m1} & \dots & \rho_{mm} \end{pmatrix} \quad (5.2.16')$$

Ma trận tương quan là một ma trận đối xứng có các phần tử trên đường chéo chính bằng 1.

Nếu  $X_{tj}$ ,  $j=1..m$ ,  $t=1..n$  là số liệu thực nghiệm của các biến  $X_j$  thì ước lượng  $r_{jk}$  của  $\rho_{jk}$  được xác định bởi:

$$r_{jk} = \frac{\frac{1}{n} \sum_{t=1}^n (x_{tj} - \bar{x}_j)(x_{tk} - \bar{x}_k)}{\sqrt{\frac{1}{n} \sum_{t=1}^n (x_{tj} - \bar{x}_j)^2} \sqrt{\frac{1}{n} \sum_{t=1}^n (x_{tk} - \bar{x}_k)^2}} \quad (5.2.17)$$

trong đó  $\bar{x}_j = \frac{1}{n} \sum_{t=1}^n x_{tj}$  là trung bình của biến  $X_j$ ,  $j=1..m$ .

Tập hợp các hệ số tương quan  $r_{jk}$  cũng lập thành một ma trận đối xứng:

$$(r_{jk}) = \begin{pmatrix} r_{11} & \dots & r_{1m} \\ \dots & \dots & \dots \\ r_{m1} & \dots & r_{mm} \end{pmatrix} \quad (5.2.17')$$

#### 5.2.5 Khảo sát mối quan hệ tương quan giữa hai biến

Việc đánh giá mối quan hệ tương quan giữa hai biến có thể được tiến hành thông qua việc xem xét hệ số tương quan giữa chúng tính được từ tập mẫu. Giá trị tuyệt đối của hệ số tương quan càng lớn thì mối quan hệ tuyến tính giữa hai biến càng chặt chẽ. Hệ số tương quan dương phản ánh mối quan hệ cùng chiều (đồng biến), ngược lại, hệ số tương quan âm biểu thị mối quan hệ ngược (nghịch biến)

giữa hai biến. Tuy nhiên, như đã chỉ ra trong mục 5.2.1, khái niệm hệ số tương quan được trình bày trên đây mới chỉ cho phép ta đánh giá được mối quan hệ tuyến tính giữa hai tập mẫu.

Thực tế trong nhiều trường hợp, khi khảo sát mối quan hệ giữa hai biến, người ta chưa cần hoặc thậm chí không cần những kết quả tính toán chính xác của hệ số tương quan, mà trước hết muốn biết bức tranh khái quát về quan hệ giữa hai tập mẫu để từ đó đưa ra quyết định cho những bước xử lý tiếp theo. Đa số trong những trường hợp như vậy người ta thường quan tâm đến khả năng tồn tại mối quan hệ tương quan tuyến tính giữa các biến khảo sát. Khi đó thay cho việc tính hệ số tương quan trên đây, người ta có thể xây dựng các đồ thị điểm biểu diễn sự phụ thuộc hoặc tính các hệ số tương quan giản lược.

Ngày nay nhờ có phương tiện máy tính, việc biểu diễn đồ thị điểm để khảo sát sơ bộ sự phụ thuộc tương quan giữa các biến đã trở nên phổ biến và rất có hiệu quả. Đồ thị điểm thông thường được biểu diễn trên hệ tọa độ vuông góc trong mặt phẳng, với hai trục tọa độ biểu thị sự biến thiên của hai biến  $X$ ,  $Y$  (hay  $X_1$ ,  $X_2$ ). Mỗi một cặp quan trắc  $\{x_i, y_i\}$  được biểu diễn bởi một điểm trên mặt phẳng. Căn cứ vào sự phân bố của tập hợp các điểm này ta có thể đánh giá được quan hệ giữa các biến.

Hình 5.1 dẫn ra một ví dụ đồ thị điểm biểu diễn mối quan hệ giữa nhiệt độ tối cao ( $T_x$ ) và nhiệt độ tối thấp ( $T_m$ ) trong những ngày tháng 1 ở một trạm. Từ đồ thị ta có thể thấy sự phân bố “hỗn loạn” của tập hợp các điểm trên mặt phẳng. Có những chỗ các điểm qui tụ khá dày đặc nhưng cũng có những chỗ chỉ rải rác 1–2 điểm. Sự phân bố tản mạn đó của các điểm biểu thị mối quan hệ “kém chặt chẽ” giữa hai yếu tố  $T_x$  và  $T_m$ . Tuy vậy, xét một cách tổng thể ta thấy giữa hai yếu tố này tồn tại sự phụ thuộc lẫn nhau: Đường như nhiệt độ tối thấp bé có liên quan tới giá trị của nhiệt độ tối cao bé, và nhiệt độ tối thấp lớn có xu hướng kéo theo nhiệt độ tối cao lớn. Ngoài ra, đồ thị còn cho thấy trong khoảng nhiệt độ  $T_m$  từ 12–18°C mỗi liên hệ giữa  $T_m$  và  $T_x$  có vẻ yếu hơn nhiều so với trường hợp giá trị  $T_m$  nằm ngoài khoảng đó.

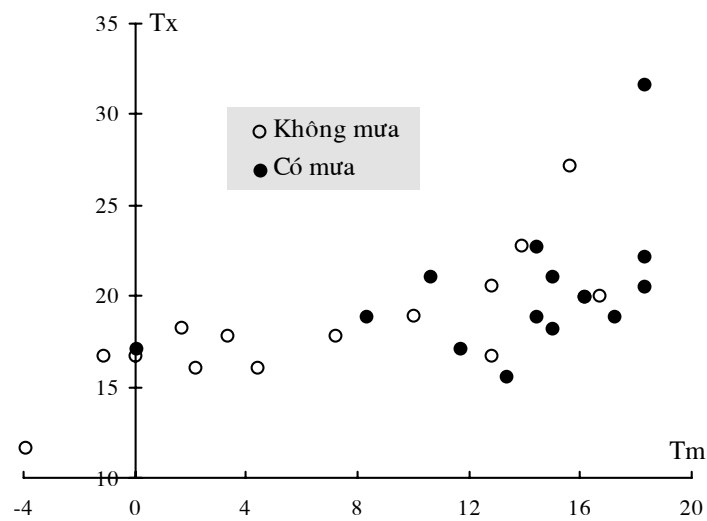
Việc chia tập số liệu ra làm hai trường hợp có mưa và không mưa sẽ làm đa dạng hóa đồ thị, cho phép khảo sát tỷ mỉ hơn mối quan hệ giữa hai biến. Hiện tượng các điểm ứng với trường hợp có mưa qui tụ vào khoảng nhiệt độ tối thấp từ 12–18°C gợi cho ta một nhận định rằng trong những ngày có mưa mối quan hệ giữa hai biến trở nên “kém chặt chẽ” hơn. Mặt khác, điều đó làm cho ta liên tưởng đến xác suất có điều kiện đã xét trước đây.

Với mục đích đánh giá mức độ tương quan tuyến tính giữa hai biến một cách nhanh chóng nhưng không cần độ chính xác cao ngoài việc sử dụng phương pháp đồ thị điểm đôi khi người ta còn tính hệ số tương quan hạng (*range correlation*

*coefficient*). Khác với hệ số tương quan mà ta đã xét, hệ số tương quan hạng được tính không phải với chính các giá trị của số liệu mà với *thứ hạng lớn bé* của chúng trong toàn tập mẫu. Nghĩa là từ tập mẫu ban đầu  $\{x_t, y_t, t=1..n\}$  ta biến đổi thành tập mới  $\{u_t, v_t, t=1..n\}$  trong đó  $u_t, v_t$  tương ứng chỉ các thành phần  $x_t, y_t$  *được xếp thứ bao nhiêu* trong bảng xếp hạng từ nhỏ nhất đến lớn nhất của mỗi chuỗi. Rõ ràng, các tập các thành phần của tập mới phải thỏa mãn  $1 \leq u_t, v_t \leq n$ . Hệ số tương quan hạng được tính bởi công thức:

$$r_{\text{range}} = 1 - \frac{6 \sum_{t=1}^n D_t^2}{n(n-1)(n+1)} \quad (5.2.18)$$

trong đó  $D_t = u_t - v_t$  là hiệu giữa các thứ hạng của  $x_t$  và  $y_t$  trong từng chuỗi.



Hình 5.1 Đồ thị điểm biểu diễn sự phụ thuộc giữa  $T_x$  và  $T_m$

**Ví dụ 5.2.3** Bảng 5.3 dẫn ra kết quả tính hệ số tương quan hạng cho tập mẫu nhiệt độ tối thấp ( $T_m$ ) và nhiệt độ tối cao ( $T_x$ ). Cột thứ nhất và cột thứ hai chứa số liệu ban đầu. Cột 3, 4, 5 chứa các giá trị tương ứng của  $T_m, T_x$  trong tập ban đầu và kết quả xếp hạng chúng. Cột 6 và cột 7 chứa giá trị hạng của từng thành phần tương ứng trong cột 1 và cột 2. Cột cuối cùng là hiệu giữa các hạng. Chẳng hạn,  $u_1=4$  có nghĩa là ứng với  $T_{m1}=12.8$  ở cột 1, khi đối chiếu giá trị này ở kết quả xếp hạng (cột 3 và cột 5) ta nhận được hạng của  $T_{m1}$  bằng 4. Tương tự như vậy với  $v_1=8$  (giá trị  $T_{x1}=20.6$ , tìm giá trị này ở cột 4 rồi đối chiếu sang cột 5 ta có hạng bằng 8). Hiệu  $D_1 = 4-8=-4$ .

Sử dụng kết quả tính trung gian ở bảng 5.3 kết hợp với công thức (5.2.18) với  $n=10$  ta nhận được  $r_{\text{range}} = 0.4546$ .

**Bảng 5.3** Tính hệ số tương quan hạng

| Số liệu ban đầu |                | Kết quả xếp hạng |                |      | Số liệu xếp hạng |                | D <sub>t</sub> |
|-----------------|----------------|------------------|----------------|------|------------------|----------------|----------------|
| T <sub>m</sub>  | T <sub>x</sub> | T <sub>m</sub>   | T <sub>x</sub> | Hạng | u <sub>t</sub>   | v <sub>t</sub> |                |
| (1)             | (2)            | (3)              | (4)            | (5)  | (6)              | (7)            | (8)            |
| 12.8            | 20.6           | 1.7              | 16.1           | 1    | 4                | 8              | -4             |
| 16.1            | 20.0           | 4.4              | 18.0           | 2    | 9                | 7              | 2              |
| 14.4            | 18.6           | 10.0             | 18.3           | 3    | 6                | 5              | 1              |
| 1.7             | 18.0           | 12.8             | 18.4           | 4    | 1                | 2              | -1             |
| 4.4             | 16.1           | 13.9             | 18.6           | 5    | 2                | 1              | 1              |
| 10.0            | 18.4           | 14.4             | 18.9           | 6    | 3                | 4              | -1             |
| 13.9            | 22.8           | 14.8             | 20.0           | 7    | 5                | 9              | -4             |
| 14.8            | 23.0           | 15.0             | 20.6           | 8    | 7                | 10             | -3             |
| 15.0            | 18.3           | 16.1             | 22.8           | 9    | 8                | 3              | 5              |
| 17.2            | 18.9           | 17.2             | 23.0           | 10   | 10               | 6              | 4              |

### 5.3 Hồi qui tuyến tính một biến

#### 5.3.1 Khái niệm về hồi qui

Xét mối quan hệ giữa hai biến ngẫu nhiên X và Y. Khi đó có thể xảy ra hai trường hợp sau đây:

- Giữa chúng có mối quan hệ phụ thuộc hàm nếu tồn tại một hàm f nào đó sao cho có thể biểu diễn được  $X = f(Y)$ .
- Giữa chúng có mối quan hệ phụ thuộc thống kê nếu mỗi giá trị x của X tương ứng với một hàm phân bố (hoặc hàm mật độ) có điều kiện  $F(y/x)$  (hoặc  $f(y/x)$ ) của Y. Ta gọi mối quan hệ phụ thuộc này là sự phụ thuộc tương quan giữa hai biến ngẫu nhiên.

Để nghiên cứu mối phụ thuộc tương quan giữa hai biến X và Y trên cơ sở tập mẫu quan trắc  $\{(x_t, y_t), t=1..n\}$  ta cần phải chọn dạng lý thuyết của phân bố đồng thời  $F(x, y)$ , hoặc dạng hàm mật độ đồng thời  $f(x, y)$ , sau đó phải ước lượng các tham số này. Từ đó ta tìm được mật độ phân bố có điều kiện:

$$f(y/x) = \frac{f(x, y)}{f_1(x)}, \quad f(x/y) = \frac{f(x, y)}{f_2(y)} \quad (5.3.1)$$

trong đó  $f_1(x)$ ,  $f_2(y)$  là các hàm mật độ riêng của X và Y.

(Chú ý rằng, trong mục này và một số mục tiếp theo ta đã thay đổi một cách tự nhiên ký hiệu các biến ngẫu nhiên X, Y thay cho ký hiệu trước đây vẫn dùng là  $X_1$ ,  $X_2$ . Sự thay đổi này hoàn toàn không ảnh hưởng tới bản chất của vấn đề. Tuy

nhien, do thói quen cố hữu trong toán học, nếu ta dùng ký hiệu mới này thì khái niệm hàm (Y) và đối số (X) tỏ ra dễ chấp nhận khi trình bày ?!. Sau này, ta sẽ quay lại ký hiệu trước đây).

Như vậy việc nghiên cứu sự phụ thuộc tương quan như trên là hết sức công kênh và phức tạp. Do đó trong thực tế người ta chỉ giới hạn xét mối quan hệ phụ thuộc giữa X và một số đặc trưng có điều kiện của Y, như kỳ vọng, trung vị, mốt,... trong đó phổ biến hơn cả là nghiên cứu mối quan hệ giữa X và kỳ vọng có điều kiện  $M[Y/X]$ :

$$m_y(x) = M[X/Y=x] = \int_{-\infty}^{+\infty} yf(y/x)dy \quad (5.3.2)$$

Và người ta gọi sự phụ thuộc này là phụ thuộc hồi qui: Hồi qui của Y lên X. Hệ thức (5.3.2) thông thường được biểu diễn dưới dạng:

$$y = m_y(x) \quad (5.3.3)$$

Quan hệ (5.3.3) được gọi là phương trình hồi qui I hay đường hồi qui I. Nếu quan hệ này là một hàm tuyến tính thì hồi qui được gọi là hồi qui tuyến tính. Tuy nhiên, trong trường hợp tổng quát (5.3.3) là một hàm bất kỳ.

Một tính chất quan trọng của hồi qui I là tính cực tiểu:

Nếu ta tìm được một hàm  $g(X)$  sao cho  $M[Y - g(X)]^2 \rightarrow \min$

$$\text{thì} \quad g(X) = M[Y/X], \text{ hay } g(x) = m_y(x). \quad (5.3.4)$$

Vì quan hệ (5.3.3) là một đường bất kỳ mà việc biểu diễn giải tích nó nói chung rất khó khăn, thậm chí không thể được cho nên trong thực tế thay cho (5.3.3) người ta xấp xỉ nó trong một lớp hàm  $f$  xác định nào đó đã biết:

$$y \approx \hat{y} = f(x) \quad (5.3.5)$$

Trong trường hợp này hàm hồi qui tìm được gọi là hồi qui II. Nếu hàm hồi qui II được xác định bằng phương pháp bình phương tối thiểu thì nó được gọi là hồi qui bình phương trung bình. Trường hợp đơn giản nhất của hồi qui bình phương trung bình là hồi qui bình phương trung bình tuyến tính— $f(x)$  là hàm bậc nhất.

Từ nay trở đi, nếu không nói gì thêm, ta sẽ hiểu hồi qui II là hồi qui bình phương trung bình và được gọi một cách đơn giản là hồi qui II.

Nếu hồi qui II (5.3.5) là tuyến tính, khi đó ta có thể viết:

$$Y = f(X) = \alpha + \beta X$$

$$\text{Hay} \quad \hat{y} = f(x) = \alpha + \beta x$$

Ta có thể chứng minh được rằng để  $f(x)$  xấp xỉ tốt nhất theo nghĩa bình phương tối thiểu của hồi qui I thì các hệ số  $\alpha$  và  $\beta$  sẽ được xác định bởi:

$$\alpha = M[Y] - \beta M[X], \quad \beta = \mu_{12}/\mu_{11}$$

trong đó  $\mu_{12}$  là mômen tương quan giữa  $X$  và  $Y$  còn  $\mu_{11} = D[X]$ . Ta sẽ quay trở lại vấn đề này khi trình bày cách xác định các hệ số hồi qui thực nghiệm mà chúng là ước lượng của  $\alpha$  và  $\beta$  trong mục sau.

### 5.3.2 Xây dựng phương trình hồi qui tuyến tính một biến từ số liệu thực nghiệm

Cho hai biến khí quyển  $X$  và  $Y$  với  $n$  cặp trị số quan sát  $\{(x_t, y_t), t=1..n\}$ . Xét sự phụ thuộc hồi qui II của  $Y$  lên  $X$  là hồi qui tuyến tính, tức là:

$$y \approx \hat{y} = a_0 + a_1 x \quad (5.3.6)$$

trong đó  $a_0$  và  $a_1$  là các hệ số phải tìm. Chúng là các giá trị ước lượng của tham số lý thuyết  $\alpha$  và  $\beta$  trong phương trình  $\hat{y} = \alpha + \beta x$ .

Với các trị số quan sát  $x_t$  của  $X$  ta có các giá trị của  $Y$  tính được theo (5.3.6) là:

$$\hat{y}_t = a_0 + a_1 x_t, \quad (t=1..n) \quad (5.3.6')$$

Các trị số quan trắc thực nghiệm  $y_t$  và giá trị tính toán (ước lượng) của  $Y$  theo (5.3.6') sai khác nhau một lượng bằng  $\delta_t = y_t - \hat{y}_t$ , chúng được gọi là sai số của phép xấp xỉ  $y = m_y(x)$  bởi (5.3.6). Để phép xấp xỉ này là tốt nhất theo nghĩa bình phương tối thiểu các hệ số  $a_0$  và  $a_1$  phải được xác định sao cho tổng bình phương các sai số  $\delta_t$  phải đạt nhỏ nhất:

$$\sum_{t=1}^n \delta_t^2 = \sum_{t=1}^n (y_t - \hat{y}_t)^2 \rightarrow \min$$

Xem rằng tổng các bình phương sai số như là hàm của các hệ số  $a_0, a_1$ , khi đó chúng phải thỏa mãn điều kiện:

$$R(a_0, a_1) = \sum_{t=1}^n (y_t - \hat{y}_t)^2 \rightarrow \min \quad (5.3.7)$$

Người ta đã chứng minh được rằng, để  $R(a_0, a_1)$  đạt cực tiểu thì các đạo hàm riêng của  $R(a_0, a_1)$  theo  $a_0$  và  $a_1$  phải đồng thời triệt tiêu:

$$\frac{\partial R(a_0, a_1)}{\partial a_0} = \frac{\partial R(a_0, a_1)}{\partial a_1} = 0$$

Từ đó ta nhận được hệ phương trình với các ẩn số  $a_0$  và  $a_1$ :

$$\begin{cases} \frac{\partial R(a_0, a_1)}{\partial a_0} = -2 \sum_{t=1}^n (y_t - a_0 - a_1 x_t) = 0 \\ \frac{\partial R(a_0, a_1)}{\partial a_1} = -2 \sum_{t=1}^n (y_t - a_0 - a_1 x_t) x_t = 0 \end{cases}$$

Hay:

$$\begin{cases} \sum_{t=1}^n (y_t - a_0 - a_1 x_t) = 0 \\ \sum_{t=1}^n (y_t - a_0 - a_1 x_t) x_t = 0 \end{cases} \quad (5.3.8)$$

Từ phương trình thứ nhất trong hệ (5.3.8) ta có:

$$\sum_{t=1}^n (y_t - a_0 - a_1 x_t) = 0.$$

Suy ra:  $a_0 = \bar{y} - a_1 \bar{x}$  (5.3.9)

Thay (5.3.9) vào phương trình thứ hai của (5.3.8) ta nhận được:

$$\sum_{t=1}^n (y_t - a_0 - a_1 x_t) x_t = \sum_{t=1}^n (y_t - \bar{y} + a_1 \bar{x} - a_1 x_t) x_t = 0$$

Hay 
$$\sum_{t=1}^n (y_t - \bar{y}) x_t - a_1 \sum_{t=1}^n (x_t - \bar{x}) x_t = 0$$

Do đó: 
$$a_1 = \frac{\sum_{t=1}^n (y_t - \bar{y}) x_t}{\sum_{t=1}^n (x_t - \bar{x}) x_t}$$

Vì  $\sum_{t=1}^n (y_t - \bar{y}) \bar{x} = 0$  và  $\sum_{t=1}^n (x_t - \bar{x}) \bar{x} = 0$  nên ta có:

$$a_1 = \frac{\sum_{t=1}^n (y_t - \bar{y}) x_t - \sum_{t=1}^n (y_t - \bar{y}) \bar{x}}{\sum_{t=1}^n (x_t - \bar{x}) x_t - \sum_{t=1}^n (x_t - \bar{x}) \bar{x}} = \frac{\sum_{t=1}^n (y_t - \bar{y}) (x_t - \bar{x})}{\sum_{t=1}^n (x_t - \bar{x})^2} = \frac{l_{xy}}{l_{xx}} \quad (5.3.10)$$

Hay: 
$$a_1 = \frac{l_{xy}}{l_{xx}} = \frac{l_{xy} \sqrt{l_{yy}}}{\sqrt{l_{xx} l_{yy}} \sqrt{l_{xx}}} = \frac{r_{xy} \sqrt{l_{yy}}}{\sqrt{l_{xx}}} = r_{xy} \frac{s_y}{s_x} \quad (5.3.11)$$

Như vậy, phương trình (5.3.6) với các hệ số  $a_0$  và  $a_1$  được tính theo (5.3.9) và (5.3.10) hoặc (5.3.11) xác định mối quan hệ hồi qui II của Y lên X. Nó được gọi là phương trình hồi qui tuyến tính một biến (một biến độc lập). Người ta gọi Y (hay y) là biến phụ thuộc, còn X (hay x) là biến độc lập.

Nếu không xét trực tiếp tập số liệu  $\{(x_t, y_t), t=1..n\}$  mà thay cho nó ta sử dụng tập số liệu chuẩn hoá  $\{(x'_t, y'_t), t=1..n\}$ :

$$x'_t = \frac{x_t - \bar{x}}{s_x}, \quad y'_t = \frac{y_t - \bar{y}}{s_y}$$

thì, bằng các phép biến đổi tương tự trên đây ta nhận được:

$$a'_0 = 0 \quad \text{và} \quad a'_1 = r_{xy}$$

*Ví dụ 5.3.1:* Từ số liệu nhiệt độ tháng 5 trạm A (biến Y – cột 1) và trạm B (biến X – cột 2) cho trong bảng 5.4, sau khi tiến hành các bước tính trung gian (ở các cột tiếp theo) ta nhận được:

$$\bar{x} = 25,9; \quad \bar{y} = 22,9; \quad l_{xy} = 7,588; \quad l_{xx} = 18,624;$$

$$a_1 = l_{xy}/l_{xx} = 7,588/18,624 = 0,407;$$

$$a_0 = \bar{y} - a_1 \cdot \bar{x} = 22,9 - 0,407 \times 25,9 = 12,361;$$

Vậy phương trình hồi qui tuyến tính giữa Y và X có dạng:

$$y = 12,361 + 0,407.X$$

**Bảng 5.4** Các bước tính hệ số hồi qui giữa y và x

| y                | x                | $y - \bar{y}$ | $x - \bar{x}$ | $(y - \bar{y})(x - \bar{x})$ | $(x - \bar{x})^2$  |
|------------------|------------------|---------------|---------------|------------------------------|--------------------|
| 22,7             | 27,7             | -0,2          | 1,8           | -0,4048                      | 3,0976             |
| 23,8             | 26,0             | 0,9           | 0,1           | 0,0522                       | 0,0036             |
| 23,7             | 26,5             | 0,8           | 0,6           | 0,4312                       | 0,3136             |
| 21,3             | 24,3             | -1,6          | -1,6          | 2,6732                       | 2,6896             |
| 22,5             | 28,0             | -0,4          | 2,1           | -0,8858                      | 4,2436             |
| 25,1             | 27,4             | 2,2           | 1,5           | 3,1682                       | 2,1316             |
| 23,3             | 25,9             | 0,4           | 0,0           | -0,0148                      | 0,0016             |
| 23,8             | 24,4             | 0,9           | -1,5          | -1,3398                      | 2,3716             |
| 21,2             | 24,3             | -1,7          | -1,6          | 2,8372                       | 2,6896             |
| 21,9             | 24,9             | -1,0          | -1,0          | 1,0712                       | 1,0816             |
| $\bar{y} = 22,9$ | $\bar{x} = 25,9$ |               |               | $l_{xy} = 7,5880$            | $l_{xx} = 18,6240$ |

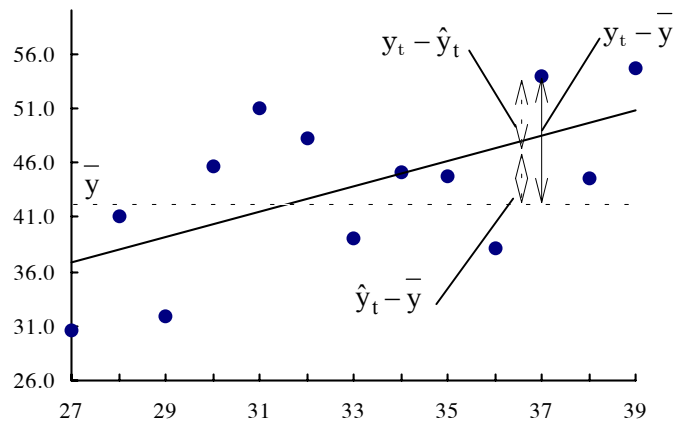
### 5.3.3 Phân tích phương sai phương trình hồi qui tuyến tính một biến

Phương trình hồi qui  $\hat{y} = a_0 + a_1 x$  là hệ thức biểu thị mối quan hệ tuyến tính giữa hai biến Y và X. Tuy nhiên, do những dao động ngẫu nhiên mà các điểm thực

nghiệm  $(x_t, y_t)$  nói chung thường phân bố xoay quanh đường thẳng hồi qui, tức là có sự sai khác giữa  $y_t$  và  $\hat{y}_t$ . Mặt khác, các giá trị quan trắc  $y_t$  của  $Y$  cũng dao động biến đổi xung quanh giá trị trung bình  $\bar{y}$  (hình 5.2). Những dao động của  $y_t$  xung quanh  $\bar{y}$  thường do nhiều nguyên nhân gây nên. Phân tích phương sai là xem xét vai trò của các nguyên nhân tạo nên những biến đổi của  $Y$ .

Mức độ biến động của  $Y$  được đánh giá thông qua tổng bình phương các độ lệch của  $y_t$  khỏi giá trị trung bình của nó:

$$l_{yy} = \sum_{t=1}^n (y_t - \bar{y})^2.$$



**Hình 5.2** Sơ đồ phân tích phương sai

Từ hình 5.2 ta thấy, mỗi một thành phần  $y_t - \bar{y}$  có thể được tách thành tổng 2 thành phần: Sự sai lệch của  $y_t$  so với đường hồi qui và sự sai lệch của giá trị hồi qui  $\hat{y}_t$  so với trung bình  $\bar{y}$ :

$$y_t - \bar{y} = (y_t - \hat{y}_t) + (\hat{y}_t - \bar{y})$$

$$\begin{aligned} \text{Do đó:} \quad l_{yy} &= \sum_{t=1}^n [(y_t - \hat{y}_t) + (\hat{y}_t - \bar{y})]^2 = \\ &= \sum_{t=1}^n (y_t - \hat{y}_t)^2 + \sum_{t=1}^n (\hat{y}_t - \bar{y})^2 + 2 \sum_{t=1}^n (y_t - \hat{y}_t)(\hat{y}_t - \bar{y}) \end{aligned}$$

$$\begin{aligned} \text{Vì} \quad \sum_{t=1}^n (y_t - \hat{y}_t)(\hat{y}_t - \bar{y}) &= \sum_{t=1}^n (y_t - a_0 - a_1 x_t)(a_0 + a_1 x_t - \bar{y}) = \\ &= \sum_{t=1}^n (y_t - \bar{y} - a_1 \bar{x} - a_1 x_t)(\bar{y} + a_1 \bar{x} + a_1 x_t - \bar{y}) = \\ &= n(a_1(\overline{xy} - \bar{x}\bar{y}) - a_1^2(\overline{x^2} - \bar{x}^2)) = a_1 r_{xy} s_x s_y - a_1^2 s_x^2 = 0 \end{aligned}$$

$$\text{Nên} \quad l_{yy} = \sum_{t=1}^n (y_t - \hat{y}_t)^2 + \sum_{t=1}^n (\hat{y}_t - \bar{y})^2 = Q + U \quad (5.3.12)$$

$$\text{trong đó} \quad Q = \sum_{t=1}^n (y_t - \hat{y}_t)^2, \quad U = \sum_{t=1}^n (\hat{y}_t - \bar{y})^2 \quad (5.3.13)$$

Người ta gọi  $U$  là tổng bình phương các biến sai hồi qui, còn  $Q$  là tổng bình phương các biến sai thặng dư. Như vậy tổng bình phương các độ lệch của  $y$  khỏi giá trị trung bình là sự đóng góp của tổng bình phương các biến sai hồi qui và tổng bình phương các biến sai thặng dư.

Ta thấy đối với một tập mẫu thì  $\bar{y}$  không đổi, do đó sự biến đổi  $\hat{y}_t$  là nguyên nhân gây nên sự biến đổi của  $U$ . Đại lượng  $U$  đặc trưng cho mức đóng góp của nhân tố hồi qui trong độ phân tán của  $Y$ . Còn  $Q$  đặc trưng cho sự đóng góp ngoài hồi qui.

Ta có:

$$\begin{aligned} U &= \sum_{t=1}^n (\hat{y}_t - \bar{y})^2 = \sum_{t=1}^n (a_0 + a_1 x_t - a_0 - a_1 \bar{x})^2 = a_1^2 \sum_{t=1}^n (x_t - \bar{x})^2 = \\ &= a_1^2 l_{xx} = a_1 \frac{l_{xy}}{l_{xx}} l_{xx} = a_1 l_{xy} \end{aligned}$$

$$Q = l_{yy} - U = l_{yy} - a_1 l_{xy}$$

$$\text{Do đó} \quad \frac{U}{l_{yy}} = \frac{a_1 l_{xy}}{l_{yy}} = \frac{l_{xy}^2}{l_{xx} l_{yy}} = r_{xy}^2. \quad (5.3.14)$$

Như vậy,  $U$  càng lớn khi  $r_{xy}$  càng lớn. Tức là  $U$  càng lớn thì mức độ tương quan tuyến tính giữa  $X$  và  $Y$  càng chặt chẽ.

$$\frac{Q}{l_{yy}} = \frac{l_{yy} - U}{l_{yy}} = 1 - \frac{U}{l_{yy}} = 1 - r_{xy}^2 \quad (5.3.15)$$

Từ đó suy ra rằng,  $r_{xy}$  càng lớn thì  $Q$  càng bé. Hồi qui được gọi là tốt nhất (lý tưởng) nếu tổng bình phương các biến sai thặng dư  $Q = 0$ . Khi đó  $r_{xy}^2 = 1$ , tất cả các điểm thực nghiệm đều nằm trên đường hồi qui. Nếu  $Q$  càng bé thì hồi qui càng tốt, điều đó cũng có nghĩa là nếu  $U$  càng lớn thì hồi qui càng có hiệu quả.

#### 5.3.4 Sự dao động của các điểm thực nghiệm xung quanh đường hồi qui

Từ (5.3.15) ta thấy rằng khi  $r_{xy}^2 = 1$  thì  $Q = 0$ . Như vậy ta có thể dùng đại lượng  $Q$  để đo mức độ dao động của các điểm thực nghiệm xung quanh đường hồi qui. Tuy nhiên, theo (5.3.13) thứ nguyên của  $Q$  bằng bình phương thứ nguyên của  $Y$ . Hơn

nữa, số bậc tự do của  $l_{yy}$  là  $n-1$ , của  $U$  là 1 (1 nhân tố), do đó số bậc tự do của  $Q$  là  $n-2$ . Chính vì vậy thay cho  $Q$ , trong thực tế người ta sử dụng đại lượng:

$$s = \sqrt{\frac{Q}{n-2}} \quad (5.3.16)$$

làm thước đo mức độ dao động của các giá trị thực nghiệm xung quanh trị số hồi qui. Giá trị của  $s$  càng nhỏ thì các điểm thực nghiệm càng nằm sát đường hồi qui. Đại lượng  $s$  được gọi là chuẩn sai thặng dư. Vậy chuẩn sai thặng dư là thước đo phần đóng góp trung bình của nhân tố ngoài hồi qui đối với sai số của phép hồi qui. Nói cách khác,  $s$  là chỉ tiêu phản ánh độ chính xác của hồi qui.

Khi  $|r_{xy}| \neq 1$  thì các điểm thực nghiệm không nằm trùng hoàn toàn trên đường hồi qui  $\hat{y} = a_0 + a_1x$  và sự tản mạn này có thể thấy được thông qua số liệu thực tế (hình 5.2). Vậy một vấn đề đặt ra là ứng với mỗi giá trị  $x_t$  xác định, quan hệ giữa  $y_t$  và  $\hat{y}_t$  sẽ như thế nào?

Theo (5.3.16), nói chung các trị số  $y_t$  của  $Y$  dao động xung quanh  $\hat{y}_t$  với mức trung bình là  $s$  và người ta đã xác định được rằng sự phân bố của  $y_t$  xung quanh  $\hat{y}_t$  gần với phân bố chuẩn. Tức là:

$$y_t \in N(\hat{y}_t, s)$$

Hay 
$$y'_t = \frac{y_t - \hat{y}_t}{s} \in N(0, 1)$$

Từ đó ta có: 
$$P(|y_t - \hat{y}_t| < s) = P\left(\left|\frac{y_t - \hat{y}_t}{s}\right| < 1\right) = \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-\frac{1}{2}t^2} dt \approx 0.68$$

Như vậy, xác suất để các giá trị  $y_t$  dao động xung quanh  $\hat{y}_t$  trong khoảng  $1s$  bằng 68%. Hay nói cách khác, có khoảng 68% số điểm thực nghiệm nằm trong phạm vi  $\pm 1s$  kể từ đường hồi qui.

Bằng cách tính tương tự, ta có:

$$P(|y_t - \hat{y}_t| < 2s) \approx 0.95 \text{ và } P(|y_t - \hat{y}_t| < 3s) \approx 0.997$$

Tức là có khoảng 95% số điểm thực nghiệm rơi vào miền  $\hat{y}_t \pm 2s$  và 99.7% số điểm rơi vào miền  $\hat{y}_t \pm 3s$ . Vậy hầu như tất cả các giá trị  $y_t$  đều nằm trong khoảng  $\hat{y}_t \pm 3s$ .

### 5.3.5 Đánh giá chất lượng phương trình hồi qui

Có thể nhận thấy rằng, việc đánh giá chất lượng phương trình hồi qui (5.3.6) là "tốt" hay "không tốt" hoặc "xấu" căn cứ vào hệ số tương quan  $r_{xy}$  hoặc theo giá trị