Chương 6

Tương quan và hồi quy

Trong chương này chúng ta sẽ xem xét mối quan hệ giữa hai biến định lượng được khảo sát đồng thời trên một đám đông, điều này có nghĩa là khi ta lấy ngẫu nhiên một cá thể của đám đông ra xem xét thì phải cân đo, phân tích, thử nghiệm đồng thời hai đặc tính sinh học định lượng X và Y.

Ví dụ cân và đo đường kính của trứng gà, cân và đo vòng ngực của bò, cân khối lượng buồng trứng và đo chiều dài của cá, nhiệt độ môi trường và lượng thức ăn thu nhận; hàm lượng lysin và protein trong thức ăn, độ dày mỡ lưng và tỷ lệ nạc ở lợn . . .

Sau khi khảo sát một mẫu gồm n cá thể ta thu được n cặp số (x_i, y_i), một câu hỏi rất tự nhiên là hai biến X và Y có quan hệ với nhau hay không ? nếu có thì khi X thay đổi Y sẽ thay đổi theo như thế nào?

Câu hỏi đầu: X và Y có quan hệ với nhau hay không được trình bầy ở mục hệ số tương quan, câu hỏi sau khi X thay đổi Y sẽ thay đổi theo như thế nào được trình bầy ở mục hồi quy.

6.1. Sắp xếp số liệu

Khi có ít số liệu có thể để dãy n cặp số dưới dạng cột hay hàng, nếu nhiều hơn thì có thể sắp dưới dạng có tần số, nếu nhiều nữa thì chia khoảng cả X và Y để sắp thành bảng hai chiều.

1) Sắp thành hàng

X	x_1	x_2	. . .	x_n
Y	y_1	y_2	. . .	y_n

2) Sắp thành hàng có tần số

X	x_1	x_2	. . .	x_k	
Y	y_1	y_2	. . .	y_k	
m	m_1	m_2	. . .	m_k	n

3) Sắp thành cột hoặc thành cột có tần số

X	Y
x_1	y_1
x_2	y_2
. . .	. . .
x_n	y_n

X	Y	m
x_1	y_1	m_1
x_2	y_2	m_2
. . .	. . .	. . .
x_k	y_k	m_k
	Tổng	n

4) Sắp thành bảng, X gồm k lớp, Y gồm l lớp với các điểm giữa x_i và y_j

	y_1	y_2	. . .	y_l
x_1	m_{11}	m_{12}	. . .	m_{1l}
x_2	m_{21}	m_{22}	. . .	m_{2l}
. . .	. . .	. . .	. . .	. . .
x_k	m_{k1}	m_{k2}	. . .	m_{kl}

Từ dạng bảng có thể dễ dàng chuyển thành dạng cột hay hàng có tần số và ngược trở lại chuyển từ dạng cột hay hàng có tần số thành bảng.

Ở phần sau các công thức tính toán đưa ra chỉ đúng khi số liệu viết dưới dạng hai cột không có tần số, khi có tần số thì phải thêm tần số vào các công thức.

6.2. Hệ số tương quan.

Trong toán học khi có hai dãy số x_i và y_i người ta có thể khảo sát mối quan hệ giữa X và Y bằng khái niệm hàm số.

Trong thống kê x_i và y_i là các giá trị thu được trong mẫu quan sát của hai biến ngẫu nhiên X, Y và người ta muốn đưa ra một con số để đánh giá hai biến ngẫu nhiên X và Y có quan hệ với nhau hay không.

Có khá nhiều con số được dùng để đánh giá X và Y có quan hệ hay không nhưng không có con số nào thoả mãn được mọi mong muốn của chúng ta. Trong thực tế, các nhà nghiên cứu thường quan tâm đến mối quan hệ tuyến tính giữa 2 tính trạng. Mức độ quan hệ này được thể hiện bằng hệ số tương quan. Hệ số tương quan được đánh giá là đơn giản, dễ dùng và có nhiều ưu điểm, nhưng chỉ thể hiện được mối quan hệ tuyến tính giữa X và Y chứ không thể dùng để đánh giá mối quan hệ nói chung của hai biến.

6.2.1. Tính hệ số tương quan

Dựa trên lý thuyết xác suất về hệ số tương quan chúng ta có công thức sau để tính hệ số tương quan mẫu r_{XY} giữa hai biến ngẫu nhiên X và Y

$$r_{XY} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_1^n (x_i - \bar{x})^2 \sum_1^n (y_i - \bar{y})^2}} \tag{6.1}$$

Khai triển công thức này được công thức (6.2) thuận tiện hơn về mặt tính toán

$$r_{XY} = \frac{\sum_1^n x_i y_i - \frac{\sum_1^n x_i \sum_1^n y_i}{n}}{\sqrt{(x_i^2 - \frac{(\sum_1^n x_i)^2}{n})(y_1^2 - \frac{(\sum_1^n y_i)^2}{n})}} = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(x_i^2 - n(\bar{x})^2)(y_1^2 - n(\bar{y})^2)}} \tag{6.2}$$

Nếu tính tuần tự các tham số thì có thể lần lượt tính phương sai mẫu của biến X, phương sai mẫu của biến Y, hiệp phương sai mẫu của X và Y.

$$r_{XY} = \frac{Cov_{XY}}{s_X s_Y} \qquad (6.3)$$

Trong đó: $s_x^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{(n-1)}$; $s_y^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{(n-1)}$; $\text{cov}_{XY} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$

6.2.2. Tính chất của hệ số tương quan mẫu

1) Là một số nằm giữa -1 và + 1, nói cách khác $|r_{XY}| \le 1$

2) Nếu Y và X có quan hệ tuyến tính Y = a + bX thì $|r_{XY}| = 1$ và ngược lại nếu $|r_{XY}| = 1$ thì Y và X có quan hệ tuyến tính Y = a + bX

3) Nếu X và Y độc lập về xác suất thì $r_{XY} = 0$ nhưng ngược lại không đúng, nếu $r_{XY} = 0$ (gọi là không tương quan) thì chưa thể kết luận X và Y độc lập về xác suất. (Như vậy độc lập về xác suất suy ra không tương quan nhưng không tương quan không suy ra độc lập về xác suất).

4) Nếu thực hiện hai phép biến đổi tuyến tính

U= aX + b; V = cY + d thì $r_{UV} = r_{XY}$

Tính chất này được phát biểu dưới dạng: Hệ số tương quan bất biến đối với phép biến đổi tuyến tính.

Trong thống kê thường dùng cách chọn gốc đo mới và đơn vi đo mới. Nếu gọi xo là gốc mới, h là đơn vị mới, số đo x của biến X bây giờ là u:

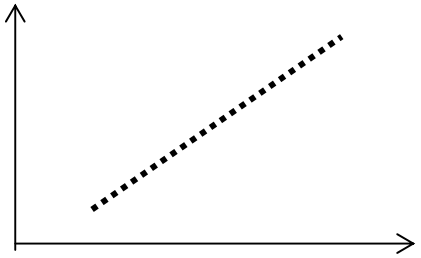
$$u = \frac{(x - xo)}{h} \quad \text{hay} \quad x = xo + hu$$

như vậy ta đã thực hiện phép biến đổi tuyến tính X = xo + hU. Tương tự đối với Y ta biến đổi Y = yo + kV

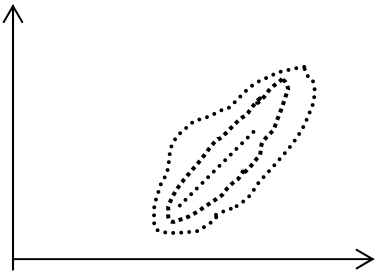
Bốn tính chất này có thể chứng minh chặt chẽ nhờ các bất đẳng thức toán học đối với 2 dãy số nhưng ở đây chúng ta thừa nhận không chứng minh.

Hệ số tương quan được coi là một số đo mối quan hệ hay liên hệ tuyến tính giữa X và Y vì khi $|r_{XY}|$ gần về phía 1 (thường gọi là tương quan mạnh) thì có thể kết luận X và Y có quan hệ gần với quan hệ tuyến tính, còn nếu $|r_{XY}|$ gần về phía 0 (thường gọi là tương quan yếu) thì không kết luận được gì vì có thể X và Y độc lập hoặc có thể có quan hệ, nhưng nếu có thì quan hệ này không thể là quan hệ tuyến tính.

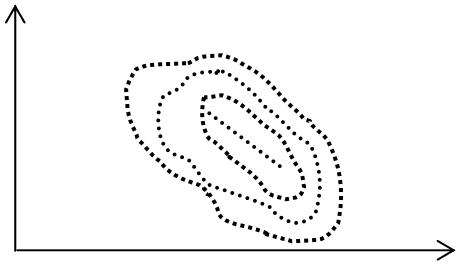
Về dấu thì nếu $r_{XY} > 0$ ta có tương quan dương, nếu < 0 thì tương quan âm



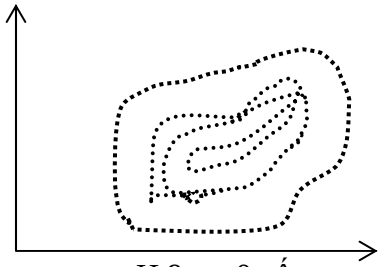
H6 r = 1



H 7 r > 0 mạnh



H 8 r < 0



H 9 r > 0 yếu

Ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (gram) của một loại trứng gà. Tiến hàn đo đường kính lớn và cân khối lượng của 10 quả trứng. Số liệu thu thập được như sau:

Quả trứng	1	2	3	4	5	6	7	8	9	10
Đường kính lớn (x)	57	54	55	52	55	60	56	56	57	58
Khối lượng (y)	61	59	58	56	57	59	56	58	56	60

Dựa vào công thức 6.1 ta có thể tính được hệ số tương quan như sau:

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
57	61	1	3	1	9	3
54	59	-2	1	4	1	-2
55	58	-1	0	1	0	0
52	56	-4	-2	16	4	8
55	57	-1	-1	1	1	1
60	59	4	1	16	1	4
56	56	0	-2	0	4	0
56	58	0	0	0	0	0
57	56	1	-2	1	4	-2
58	60	2	2	4	4	4
560	580	0	0	44	28	16

Ta có: n = 10; $\Sigma x_i = 560$; $\Sigma y_i = 580$; $\bar{x} = 56$; $\bar{y} = 58$.

Nếu tính theo (6.1)

$$r_{XY} = \frac{16}{\sqrt{44 \times 28}} = 0{,}4558$$

Nếu tính theo (6.2) thì

$$\Sigma x_i^2 = 31404;\ \Sigma y_i^2 = 33668;\ (\bar{x})^2 = 3136;\ \Sigma x_i^2 - n(\bar{x})^2 = 44$$

$$\Sigma x_i y_i = 32496;\ \Sigma x_i y_i - n \times \bar{x} \times \bar{y} = 16;\ \Sigma y_i^2 - n(\bar{y})^2 = 28$$

$$r_{XY} = \frac{16}{\sqrt{44 \times 28}} = 0{,}4558$$

Nếu tính tuần tự theo (8.3) thì:

$$s_X^2 = \frac{44}{9} = 4{,}8889\ ;\quad s_Y^2 = \frac{28}{9} = 3{,}1111\ ;\quad \mathrm{cov}_{XY} = \frac{16}{9} = 1{,}7778$$

$$r_{XY} = \frac{1{,}7778}{\sqrt{4{,}8889 \times 3{,}1111}} = 0{,}4558$$

6.3. Hồi quy tuyến tính

Vẽ các điểm quan sát $M_i(x_i,y_i)$ trên hệ toạ độ vuông góc, các điểm này họp thành một đám mây quan sát nhìn chung có dạng một elíp (trừ một vài điểm tách ra xa gọi là điểm ngoại lai), nếu $|r_{XY}|$ gần bằng 1 thì elíp rất dẹt, nếu $|r_{XY}|$ vừa phải thì elíp bầu bĩnh, nếu $|r_{XY}|$ gần bằng không thì có 2 khả năng: hoặc đám mây quan sát tản mạn trên một phạm vi rộng (không quan hệ), hoặc đám mây quan sát không còn dạng elíp mà tập trung thành một hình cong (phi tuyến).

Trường hợp $|r_{XY}|$ gần 1 elíp đám mây quan sát khá dẹt. Để giải thích sự thay đổi của Y khi cho X thay đổi người ta thường đưa ra mô hình hồi quy tuyến tính Y = a + bX.

Có thể tìm hiểu mô hình hồi quy tuyến tính theo hai cách sau đây:

6.3.1. Đường trung bình của biến ngẫu nhiên Y theo X trong phân phối chuẩn 2 chiều

Khảo sát đồng thời 2 biến ngẫu nhiên định lượng (như đã làm từ đầu chương này). Cặp biến X,Y thường tuân theo luật chuẩn hai chiều, khi ấy nếu theo dõi biến X trước thì ứng với mỗi giá trị x của biến ngẫu nhiên X có vô số giá trị của biến Y, các giá trị này có giá trị trung bình lý thuyết là kỳ vọng M(Y/ x).

Khi x thay đổi kỳ vọng M(Y/x) thay đổi theo và các điểm P(x,M(Y/ x)) chạy trên một đường thẳng gọi là đường hồi quy tuyến tính Y theo X.

Nếu theo dõi biến Y trước thì ứng với một giá trị y của Y có vô số giá trị của biến X có trung bình là kỳ vọng M(X/ y). Điểm Q(y, M(X/ y) chạy trên một đường thẳng gọi là đường hồi quy tuyến tính X theo Y.

Như vậy, về mặt lý thuyết, khi có phân phối chuẩn hai chiều các đường hồi quy tuyến tính Y theo X và hồi quy tuyến tính X theo Y chính là các đường kỳ vọng có điều kiện M(Y/x) và M(X/y).

Trong trường hợp tổng quát của phân phối hai chiều các đường kỳ vọng có điều kiện có thể là đường thẳng hoặc đường cong và được gọi là hồi quy Y theo X (hay X theo Y). Trong thực nghiệm chúng ta khảo sát 2 biến định lượng bằng cách lấy mẫu với dung lượng n khá lớn.

Thay cho đường hồi quy tuyến tính lý thuyết có đường hồi quy thực nghiệm. Gọi (x, y) là toạ độ của một điểm chạy trên đường thẳng hồi quy, $\bar{x}$ và $\bar{y}$ là trung bình cộng của X và Y, s_x và s_y là độ lệch chuẩn của X và Y, phương trình hồi quy tuyến tính thực nghiệm có dạng:

$$y - \bar{y} = r_{XY}\,\frac{s_Y}{s_X}(x - \bar{x}) \qquad (6.4)$$

Nếu viết phương trình đường thẳng dưới dạng y = a + bx thì:

hệ số góc $b = r_{XY}\frac{s_Y}{s_X}$ tung độ gốc $a = \bar{y} - b\bar{x}$ (6.5)

Nếu dùng công thức (6.2) để tính hệ số tương quan thì:

hệ số góc $b = \dfrac{\sum x_i y_i - \dfrac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}}$ tung độ gốc $a = \dfrac{\sum y_i - b\sum x_i}{n}$ (6.6)

Nếu dùng công thức (8.1) để tính hệ số tương quan thì:

hệ số góc $b = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ tung độ gốc $a = \bar{y} - b\bar{x}$ (6.7)

Đường hồi quy tuyến tính thực nghiệm X theo Y có phương trình:

$x - \bar{x} = d(y - \bar{y})$ với hệ số góc $d = r_{XY}\frac{s_X}{s_Y}$

Nếu viết dưới dạng x = c + dy thì hoành độ gốc $c = \bar{x} - c\bar{y}$

Nếu nhân hệ số góc b của hồi quy tuyến tính Y theo X với hệ số góc d của hồi quy tuyến tính X theo Y thì được r^2_{xy}:

$$b \times d = r^2_{XY}$$

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (gram) của một loại trứng gà. Tiến hành đo đường kính lớn và cân khối lượng của 10 quả trứng. Số liệu thu thập được như sau:

Ta đã có: $\bar{x} = 56$; $\bar{y} = 58$; $s^2_x = 4{,}8889$; $s^2_y = 3{,}1111$; $r_{XY} = 0{,}4558$

Hồi quy tuyến tính Y theo X

$$y - 58 = 0{,}4558\frac{\sqrt{3{,}1111}}{\sqrt{4{,}8889}}(x - 56)$$

Viết dưới dạng y = a + bx thì

Nếu tính theo (5.5) ta có:

hệ số góc $b = 0{,}4558\frac{\sqrt{3{,}1111}}{\sqrt{4{,}8889}} = 0{,}3636$ và tung độ gốc a = 58 - 0,3636. 56 = 37,6384

Nếu tính theo (5.6) ta có:

hệ số góc $b = \frac{16}{44} = 0{,}3636$ và tung độ gốc $a = \frac{580 - 0{,}3636 \times 560}{10} = 37{,}6384$

6.3.2. Đường thẳng gần đúng của Y theo X

Xét bài toán thường gặp trong các thí nghiệm nông nghiệp và sinh học sau:

Một biến X định lượng có các giá trị x_i(i = 1, n), biến này hoặc do chúng ta chủ động điều khiển ví dụ thời gian cai sữa, mức protein trong khẩu phần, mật độ nuôi trong chuồng, liều lượng thuốc, . . . , hoặc quan sát trong tự nhiên như tuổi của vật nuôi, thời gian tiết sữa, số con đẻ ra trên lứa, số con cai sữa, tiêu tốn thức ăn . . .

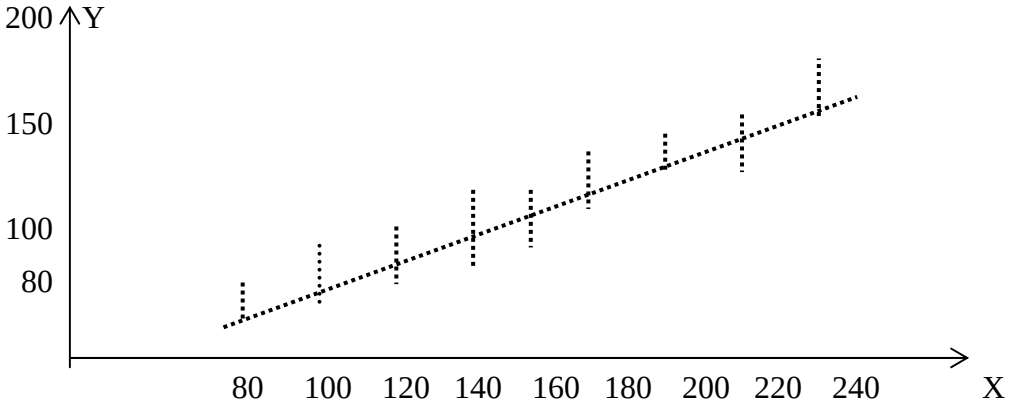
Biến thứ hai là một biến Y mà qua quan sát thấy thay đổi theo X, ví dụ khối lượng vật nuôi thay đổi theo tuổi, năng suất sữa trong một chu kỳ thay đổi theo thời gian tiết sữa, chỉ tiêu Y về phản xạ của chuột thay đổi theo lượng thuốc X đã tiêm ...

Vấn đề đặt ra là tìm một hàm của X để tính gần đúng các giá trị của Y.

Hàm này thường chọn trong các lớp hàm: bậc nhất (tuyến tính), bậc hai, lôgarít, mũ . . . hàm phải đơn giản và dễ lý giải về mặt chuyên môn.

Nếu dùng x_i làm hoành độ, y_i làm tung độ thì có n điểm quan sát $M_i(x_i,y_i)$ và bài toán ở đây là dùng một đường thẳng, đường parabôn, đường lôgarít, đường mũ, . . . để lý giải sự thay đổi của Y theo X, đường này không buộc phải đi qua tất cả các điểm mà chỉ cần đi "sát", đi "gần" các điểm quan sát M_i.

Trong phần hàm nhiều biến của toán học cao cấp sau khi tính đạo hàm riêng có đề cập đến đường thẳng "tốt" nhất theo nguyên tắc (hay phương pháp) bình phương bé nhất.



Hồi quy tuyến tính Y theo X

Giả sử chọn đường gần đúng là đường thẳng $z = a + bx$ ta có mô hình tuyến tính sau:

$$y_i = z_i + e_i = a + bx_i + e_i \qquad (6.8)$$

e_i là độ chênh lệch giữa giá trị thực y_i và giá trị tương ứng z_i trên đường thẳng (thường gọi e_i là sai số hay phần dư).

Theo nguyên tắc bình phương bé nhất thì đường "tốt" nhất trong các đường thẳng dùng làm đường gần đúng là đường có tổng bình phương các phần dư Σe^2_i nhỏ nhất.

Dùng cách tính cực trị của hàm hai biến để tìm min Σe^2_i thu được hệ hai phương trình (gọi là hệ phương trình chuẩn) để tìm a và b.

$$\begin{cases} an + b\,\Sigma x_i = \Sigma y_i \\ a\Sigma x_i + b\,\Sigma x^2_i = \Sigma x_i y_i \end{cases}$$

Có nhiều cách giải hệ hai phương trình bậc nhất với hai ẩn số. Nếu dùng định thức để giải thì có ngay kết quả sau:

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \qquad a = \frac{\sum y_i - b\sum x_i}{n} \qquad (6.9)$$

trùng với công thức (5.6) đã dùng để tính các hệ số hồi quy a và b ở phần a/

Nếu các **biến ngẫu nhiên e_i** trong mô hình tuyến tính (5.8) phân phối chuẩn thoả mãn 3 điều kiện:

a/ Kỳ vọng bằng 0

b/ Phương sai bằng nhau (6.10)

c/ Độc lập với nhau.

thì sau khi tính các hệ số theo (5.9) có thể tính được sai số của các hệ số, phân tích và đánh giá các nguồn biến động, phân tích sai số dự báo.

Đường thẳng gần đúng tốt nhất vừa tìm được theo (8.9) trong trường hợp này cũng được gọi là đường hồi quy tuyến tính Y theo X.

(Để phân biệt có khi người ta gọi đường này là đường hồi quy tuyến tính dạng I, còn đường trung bình trong mô hình phân phối chuẩn hai chiều ở a/ là đường hồi quy tuyến tính dạng II).

Trong mô hình hồi quy tuyến tính dạng I biến X (không ngẫu nhiên) được gọi là biến độc lập, biến giải thích hay biến điều khiển còn biếnY (ngẫu nhiên) thay đổi theo X được gọi là biến phụ thuộc, biến kết quả hay biến đáp.

Trở lại đường hồi quy tuyến tính ở phần a/, nếu chọn trước biến ngẫu nhiên X và coi như biến độc lập thì biến thay đổi theo Y trong phân phối chuẩn hai chiều thoả mãn các điều kiện vừa nêu ở (5.10). Như vậy đường hồi quy tuyến tính dạng II, theo nghĩa đường trung bình của biến Y theo biến X, cũng chính là đường hồi quy tuyến tính theo nghĩa vừa trình bầy: "đường thẳng gần đúng tốt nhất đối với biến Y", tức là đường hồi quy tuyến tính dạng I.

Tóm lại khi cần tính hồi quy tuyến tính theo nghĩa "Đường thẳng gần đúng tốt nhất đối với biến Y thì dù X là biến không ngẫu nhiên với các sai số e_i của mô hình thoả mãn điều kiện (5.10), hay X là biến ngẫu nhiên trong mô hình phân phối chuẩn hai chiều ta đều có thể tính các hệ số a và b bằng cách dùng các công thức (5.5), (5.6), (5.7) hoặc giải hệ 2 phương trình chuẩn.

Việc tính sai số của a và b, việc phân tích biến động chung thành biến động do hồi quy và biến động do sai số, việc tính và đánh giá dự báo hoàn toàn giống nhau.

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (gram) của một loại trứng gà. Tiến hành đo đường kính lớn và cân khối lượng của 10 quả trứng. Số liệu thu thập được như sau:

Ta đã có: n = 10; n = 10; $\Sigma x_i = 560$; $\Sigma y_i = 580$; $\Sigma x_i^2 = 31404$; $\Sigma x_i y_i = 32496$

$$10a + 560b = 580$$

$$560a + 31404b = 32496$$

Giải hệ phương trình ta được a = 37,6 ; b = 0,364. Như vậy hồi quy tuyến tính khối lượng theo đường kính lớn của trứng là:

$$y = 37,6 + 0,364x$$

6.4. Kiểm định đối với hệ số tương quan và các hệ số hồi quy

Trong mô hình phân phối chuẩn hai chiều thì hệ số tương quan mẫu là một thống kê có kỳ vọng là hệ số tương quan lý thuyết ρ. Để kiểm định giả thiết H_0: ρ = 0 với đối thiết H_1: ρ ≠ 0 phải tính giá trị T_{TN} theo công thức:

$$T_{TNR} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad \text{rồi so với giá trị tới hạn } t(\alpha/2,n-2) \quad (6.11)$$

Kết luận:

Nếu $|T_{TN}| \leq t(\alpha/2,n-2)$ thì chấp nhận H_0, ngược lại thì bác bỏ H_0

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (gram) của một loại trứng gà.

Ta đã có: n = 10; r = 0,4558

$$T_{TN} = \frac{0,4558}{\sqrt{\frac{1-0,4558^2}{10-2}}} = 1,448 ; \qquad t(0,025;8) = 2,306$$

Kết luận: chấp nhận H_0: ρ=0

Để kiểm định giả thiết H_0 : ρ = ρ_0 với đối thiết H_1: ρ ≠ ρ_0 thường thực hiện phép biến đổi

$$z = \frac{1}{2}\ln(\frac{1+r}{1-r})$$

Biến này phân phối chuẩn với kỳ vọng $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ và phương sai 1/(n-3)

Từ đó có quy tắc kiểm định:

$$Z_{TN} = \frac{\sqrt{n-3}}{2}\left(\ln(\frac{1+r}{1-r}) - \ln(\frac{1+\rho_0}{1-\rho_0})\right) = \frac{\sqrt{n-3}}{2}\ln\left(\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right)$$

so với giá trị tới hạn z(α/2) của phân phối chuẩn tắc

Kết luận: Nếu $|Z_{TN}| \leq z(\alpha/2)$ thì chấp nhận H_0, ngược lại thì bác bỏ H_0

Trong mô hình hồi quy tuyến tính y = a + bx các sai số được giả thiết phân phối chuẩn $N(0,\sigma^2)$.

Sau khi tính các hệ số a và b của đường hồi quy có thể tính được chênh lệch giữa giá trị quan sát (y_i) và giá trị tương ứng trên đường hồi quy (y^H_i)

$$y^H_i = a + bx_i \rightarrow e_i = y_i - y^H_i = y_i - (a + bx_i)$$

Phương sai σ^2 được ước lượng bởi se^2

$$SE^2 = \frac{\sum_{i=1}^{n}(y_i - (a + bx_i))^2}{(n-2)} \qquad (6.12)$$

SE được gọi là sai số của một quan sát trong mô hình hồi quy tuyến tính.

Tung độ gốc a có sai số:

$$SE(a) = SE\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (6.13)$$

Hệ số góc b có sai số:

$$SE(b) = \frac{se}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (6.14)$$

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (gram) của một loại trứng gà.

x	y	$y^H_i = 37,6+0,364x_i$	$e_i = y_i - y^H_i$	e^2_i
57	61	58,36	2,64	6,95
54	59	57,27	1,73	2,98
55	58	57,64	0,36	0,13
52	56	56,55	-0,55	0,30
55	57	57,64	-0,64	0,40
60	59	59,45	-0,45	0,21
56	56	58,00	-2,00	4,00
56	58	58,00	0,00	0,00
57	56	58,36	-2,36	5,59
58	60	58,73	1,27	1,62
560	580	580	0,00	22,18

Ta có: $\Sigma e^2_i = 22,182$; $SE^2 = 22,182 / (10-2) = 2,773$; $se = 1,664$;

$\Sigma x^2_i = 31404$; $(x_i - \bar{x})^2 = 44$

$$SE(a) = 1,664\sqrt{\frac{31404}{10\times 44}} = 14,07 \quad \text{và} \quad SE(b) = \frac{1,664}{\sqrt{44}} = 0,251$$

Từ đó có quy tắc kiểm định đối với các hệ số a và b

Giả thiết H_{0A}: a = 0 đối thiết H_{1A} : a ≠ 0

Tính $T_{TNA} = \frac{a}{s(a)}$ so với giá trị tới hạn t(α/2, n-2)

Kết luận:

Nếu $|T_{TNA}| \leq t(\alpha/2, n-2)$ thì chấp nhận H_{0A}, nếu ngược lại thì bác bỏ H_{0A}

Giả thiết H_{0B}: b = 0 đối thiết H_{1B} : b ≠ 0

Tính $T_{TNB} = \frac{b}{s(b)}$ và so với giá trị tới hạn t(α/2, n-2)

Kết luận:

Nếu $|T_{TNB}| \leq t(\alpha/2, n-2)$ thì chấp nhận H_{0B}, nếu ngược lại thì bác bỏ H_{0B}

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (gram) của một loại trứng gà.

T_{TNA} = 37,6 / 14,07 = 2,672	t(0,025 ;8) = 2,306	Kết luận: a ≠ 0
T_{TNB} = 0,364 / 0,251 = 1,450	t(0,025,5) = 2,306	Kết luận: b = 0

6.5. Dự báo theo hồi quy tuyến tính

Khi có đường hồi quy tuyến tính thì có thể dùng đường đó để dự báo giá trị Y_M ứng với giá trị x_M ngoài các giá trị x_i đã có của mẫu quan sát:

$$y_M = a + b\, x_M \qquad (6.15)$$

Trong ví dụ 6.1 hồi quy khối lượng theo đường kính lớn của trứng là

$$y = 37,6 + 0,364x$$

Dùng đường hồi quy để dự báo khối lượng một quả trứng có đường kính lớn là 59mm

$$y_{59} = 37,6 + 0,364 \times 59 = 59,076 \text{gram}$$

Các dự báo này cho ta một giá trị dự báo y_M và có thể tính được sai số dự báo, sai số này lớn dần nếu điểm dự báo x_M ở xa giá trị $\bar{x}$, như vậy dự báo xa $\bar{x}$ không tốt vì sai số quá lớn.

Sai số dự báo $$SE_M = SE\sqrt{1+\frac{1}{n}+\frac{(x_M-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \qquad (6.16)$$

Với ví dụ 1 ta có sai số dự báo là $SE_{59} = 1,664\sqrt{1+\frac{1}{10}+\frac{(59-56)^2}{44}} = 1,834$

6.6. Phân tích phương sai và hồi quy

Dựa theo ý tưởng của phương pháp phân tích phương sai có thể khảo sát tổng bình phương toàn bộ (biến động toàn bộ của y)

$$SS_{TO} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Có thể tách SS_{TO} thành hai tổng bình phương: 1) tổng bình phương do hồi quy SS_R và 2)tổng bình phương do sai số SS_E

$$SS_R = \sum_{i=1}^{n} (y_i^H - \bar{y})^2 \quad \text{với } y^H_i = a + bx_i \text{ (giá trị trên đường hồi quy)}$$

$$SSE = \sum_{i=1}^{n} (y_i - y_i^H)^2 = \sum_{i=1}^{n} e_i^2$$

Từ đó có bảng phân tích phương sai sau:

Nguồn biến động	df	SS	MS	F_{TN}	F tới hạn
Hồi quy	1	SS_R	$MS_R = SS_R/df_R$	MS_R / MS_E	$F(\alpha,df_R,df_E)$
Sai số	n-2	SS_E	$MS_E = SS_E/df_E = se^2$		
Toàn bộ	n-1	SS_{TO}			

Giả thiết H_0 : Không có hồi quy (hệ số hồi quy b = 0) với đối thiết H_1 : hệ số b ≠ 0

Nếu $F_{TN} \leq F(\alpha,df_R,df_E)$ thì chấp nhận H_0 ngược lại thì chấp nhận H_1

Chia SS_R cho SS_{TO} được $\frac{SS_R}{SS_{TO}} = r^2$ và SS_E cho SS_{TO} được $\frac{SSE}{SSTO} = 1 - r^2$

r^2 được gọi là hệ số xác định (6.16)

Ta còn có $$F_{TN} = \frac{msR}{msE} = \frac{r^2}{\frac{1-r^2}{n-2}} = T^2_{tnR} \qquad (6.17)$$

Như vậy kiểm định F tương đương với kiểm định T đối với hệ số tương quan r và tương đương với kiểm định T đối với hệ số góc b.

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (gram) của một loại trứng gà.

Từ đó có bảng phân tích phương sai sau:

Nguồn biến động	df	SS	MS	F_{TN}	F tới hạn
Hồi quy	1	5,818	5,818	2,10	0,185
Sai số	8	22,182	2,773		
Toàn bộ	9	28,000			

Kết luận : Vì F_{TN} > F tới hạn cho nên giả thiết H_0 bị bác bỏ

$F_{TN} = 5,818 / 2,773 = 2,10 = (1,449)^2 = (T_{TNB})^2 = (T_{TNR})^2$

6.7. Bài tập

6.7.1

Xác định mối liên hệ giữa khối lượng của gà mái (kg) và thu nhận thức ăn trong một năm (kg). Tiến hành quan sát trên 10 gà mái và thu được kết quả như sau :

Khối lượng gà mái	2,3	2,6	2,4	2,2	2,8	2,3	2,6	2,6	2,4	2,5
Khối lượng thức ăn	43	46	45	46	50	46	48	49	46	47

Xây dựng phương trình hồi quy tuyến tính và tính hệ số tương quan.

6.7.2

Một thí nghiệm được tiến hành để xác định mối liên hệ giữa khối lượng thân thịt lợn (kg) và độ dày mỡ lưng (mm). Tiến hành xác định các chỉ tiêu vừa nêu trên 8 thân thịt lợn, kết quả thu được như sau :

Khối lượng thân thịt	100	130	140	110	105	95	130	120
Độ dày mỡ lưng	42	38	53	34	35	31	45	43

Xây dựng phương trình hồi quy tuyến tính và tính hệ số tương quan.

6.7.3

Để xác định khối lượng của cừu (kg) thông qua chu vi lồng ngực, tiến hành cân đo trên 66 cứu. Số liệu thu được như sau :

Khối lượng (Y) và chu vi lồng ngực (X) của cừu

Y	X	Y	X	Y	X	Y	X	Y	X	Y	X
30	76	20	63	28	77	29	73	18	62	19	67
24	71	28	70	25	71	30	74	28	70	27	69
20	63	22	65	27	72	21	64	27	71	31	74
25	69	28	72	28	74	28	74	30	73	23	67
25	67	25	67	25	65	48	89	28	72	22	63
19	62	20	62	20	64	17	60	22	69	35	75
35	77	35	78	35	78	46	86	48	90	44	84
37	84	43	81	32	73	43	84	31	73	31	73
39	78	36	81	33	80	44	82	39	80	45	86
43	88	41	87	36	82	43	80	33	79	35	78
38	78	36	76	35	74	39	81	34	74	39	76

Xây dựng phương trình hồi quy tuyến tính.