Chương 12

HỒI QUY TUYẾN TÍNH ĐƠN GIẢN VÀ TƯƠNG QUAN

Trong bài toán hai mẫu đã nghiên cứu trong chương 9 ta đi so sánh giá trị các tham số của hai phân phối x và y. Mặc dù mẫu đã cho ghép cặp trong chương 9 ta đã không sử dụng thông tin của một biến để nghiên cứu về biến còn lại. Dây chính xác là vấn đề mà hồi quy nghiên cứu: đó là khám phá ra mối quan hệ giữa hai hay nhiều biến, từ đó ta có thể thu được thông tin của một biến khi biết các biến còn lại.

Khi nói về x và y có mối quan hệ với nhau ta thường nghĩ đến mối quan hệ hàm $y = f(x)$. Ví dụ x là số sản phẩm A của cửa hàng bán lẻ B bán được trong 1 ngày. Mỗi sản phẩm A bán được cửa hàng lời 25 ngàn đồng, y số tiền lời của cửa hàng B thu được khi bán sản phẩm A trong 1 ngày. Khi đó $y = 25x$. Nếu một ngày cửa hàng B bán được $x = 8$ sản phẩm A thì cửa hàng thu được số tiền lời tương ứng là $y = 8.25 = 200$ ngàn đồng.

Tuy nhiên, trong thực tế có nhiều biến x, y có mối quan hệ với nhau nhưng x và y không có quan hệ hàm số $y = f(x)$. Ví dụ như điểm thi đại học môn Toán x và điểm môn Toán trung bình khi học đại học y của sinh viên trường đại học M có mối quan hệ với nhau. Tuy nhiên, trường hợp hai sinh viên có điểm Toán thi đại học giống nhau có điểm trung bình môn Toán khi học tại trường M là khác nhau thì có thể xảy ra.

Phân tích hồi quy là một phần của thống kê nghiên cứu mối quan hệ giữa hai hay nhiều biến không theo kiểu quan hệ hàm số. Trong chương 12, ta khái quát quan hệ hàm tuyến tính $y = \beta_0 + \beta_1 x$ cho một quan hệ xác suất tuyến tính và rút ra các kết luận cho mô hình. Trong chương 13, ta sẽ xét về kỹ thuật cho các mô hình cụ thể và nghiên cứu các mối quan hệ phi tuyến tính cũng như mối quan hệ giữa nhiều hơn hai biến.

12.1 Mô hình hồi quy tuyến tính đơn giản

Quan hệ tuyến tính $y = \beta_0 + \beta_1 x$ giữa hai biến x và y là mối quan hệ toán học đơn giản giữa hai biến x và y. Tập hợp các điểm x và y xác định bởi $y = \beta_0 + \beta_1 x$ thuộc đường thẳng có hệ số dốc β_1 và hệ số tự do β_0 (tức là đường thẳng $y = \beta_0 + \beta_1 x$ cắt trục tung Oy tại điểm có tung độ bằng β_0).

Nếu hai biến x, y không có mối quan hệ hàm với nhau thì với giá trị x cố định ta không xác định được giá trị chắc chắn tương ứng của y. Ví dụ như quan sát mối quan hệ về số lượng từ vựng của trẻ em và độ tuổi của chúng thì tương ứng với một trường hợp cụ thể của

tuổi trẻ em như $x = 5$ tuổi thì số lượng từ vựng của trẻ 5 tuổi là biến ngẫu nhiên y. Quan sát cụ thể đếm số lượng từ của trẻ A có tuổi bằng 5 thấy số lượng từ của em A khoảng 2000 từ. Khi đó ta nói một giá trị quan sát được của Y tương ứng $x = 5$ là $y = 2000$.

Tổng quát, các biến có giá trị được cố định bởi người quan sát sẽ ký hiệu là x và gọi là **biến độc lập**, **biến để dự đoán** hay **biến giải thích** với x cố định, biến thứ hai sẽ là ngẫu nhiên. Ta thường ký hiệu biến ngẫu nhiên là Y và giá trị của nó là y và gọi biến này là **biến phụ thuộc** hay **biến đáp ứng, biến mong đợi**.

Thông thường các quan sát biến độc lập sẽ được thực hiện. Kí hiệu $x_1, x_2, ..., x_n$ là các giá trị quan sát được của biến độc lập; Y_i và y_i tương ứng là biến ngẫu nhiên và giá trị biến ngẫu nhiên tương ứng với giá trị quan sát x_i. Dữ liệu tồn tại dưới dạng ghép cặp $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Biểu diễn của dữ liệu này được gọi là một **biểu đồ chấm** sẽ cho những cảm nhận sơ bộ về mối quan hệ tự nhiên nào đó. Trong biểu đồ này, mỗi (x_i, y_i) được biểu diễn bởi một chấm trong hệ tọa độ hai chiều.

Mô hình tuyến tính xác suất

Đối với mô hình $y = \beta_0 + \beta_1 x$, giá trị quan sát của y là một hàm tuyến tính đối với x. Tổng quát hóa xấp xỉ này ta giả sử rằng giá trị mong đợi của Y là hàm tuyến tính của x cố định giá trị của Y khác với giá trị mong đợi một lượng ngẫu nhiên.

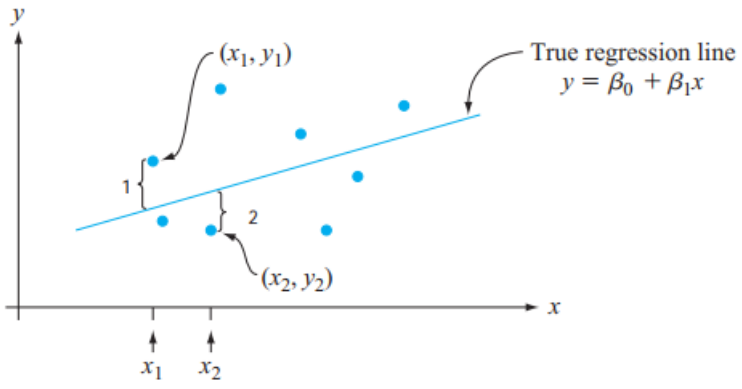
Định nghĩa 12.1: Mô hình tuyến tính đơn giản

Các định tham số β_0, β_1 và σ^2 thỏa mãn với giá trị cố định x bất kỳ của biến độc lập, biến phụ thuộc là biến ngẫu nhiên quan hệ với x qua **phương trình mô hình**

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{12.1}$$

Đại lượng ϵ trong mô hình là một biến ngẫu nhiên, được giả sử là có phân phối chuẩn với $E(\epsilon) = 0$ và $V(\epsilon) = \sigma^2$.

Biến ϵ thường được gọi **độ lệch ngẫu nhiên** hay **sai số ngẫu nhiên** trong mô hình. Nếu không có ϵ, cặp quan sát (x, y) bất kỳ sẽ tương ứng với một điểm chính thuộc đường $y = \beta_0 + \beta_1 x$, gọi là **đường hồi quy đúng (hay tổng thể)**. Suy luận về sai số ngẫu nhiên cho phép (x, y) nằm phía trên đường hồi quy đúng (khi $\epsilon > 0$) hay nằm dưới đường hồi quy đúng (khi $\epsilon < 0$). Các điểm $(x_1, y_1), ..., (x_n, y_n)$ thu được từ n quan sát được biểu diễn thành các chấm về đường hồi quy đúng. Trong trường hợp các chấm xấp xỉ tuyến tính mô hình hồi quy tuyến tính đơn giản có thể được xem xét có dùng để biểu diễn cho số liệu này.



Hình 12.3: Các điểm tương ứng với các quan sát từ mô hình hồi quy tuyến tính đơn giản.

Ký hiệu x^* là một giá trị cụ thể của biến độc lập x và

$\mu_{Y.x^*}$ là giá trị trung bình của Y khi x có giá trị x^*

$\sigma^2_{Y.x^*}$ là phương sai của Y khi x có giá trị x^*

Thay cho các ký hiệu $E(Y|x^*)$ và $V(Y|x^*)$. Ví dụ như, x là tuổi của một đứa trẻ và y là số lượng từ của trẻ, thì $\mu_{Y.5}$ là lượng từ trung bình của tất cả các đứa trẻ 5 tuổi trong tổng thể biết và $\sigma^2_{Y.5}$ là đại lượng mô tả sự phân tán các giá trị của y (lượng từ vựng của những đứa trẻ 5 tuổi) trong tổng thể so với giá trị trung bình $\sigma^2_{Y.5}$.

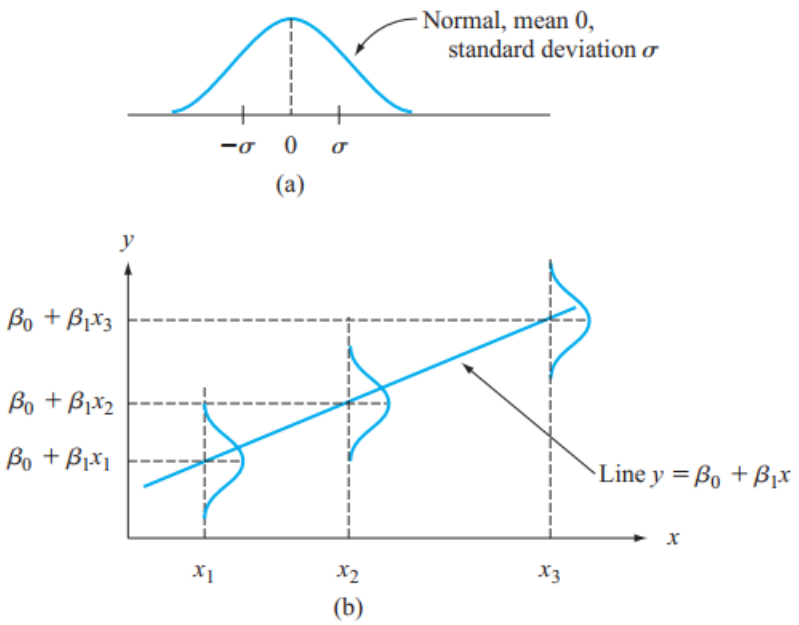
Với x cố định chỉ có duy nhất đại lượng ϵ trong vế phải của (12.1) là biến ngẫu nhiên.

$$\mu_{Y.x^*} = E(\beta_0 + \beta_1 x^* + \epsilon) = \beta_0 + \beta_1 x^* + E(\epsilon) = \beta_0 + \beta_1 x^*$$
$$\sigma^2_{Y.x^*} = V(\beta_0 + \beta_1 x^* + \epsilon) = V(\beta_0 + \beta_1 x^*) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

Thay x^* bởi x ta có $\mu_{Y.x} = \beta_0 + \beta_1 x$ tức là trung bình của y là hàm tuyến tính của x. Đường hồi quy đúng $y = \beta_0 + \beta_1 x$ suy ra là đường của các giá trị trung bình của Y trên một giá trị cụ thể của x. Hệ số dốc β_1 được diễn tả như sự thay đổi mong đợi của Y tương ứng với sự tăng lên một đơn vị của x. $\sigma^2_{Y.x^*} = \sigma^2$ đẳng thức cho thấy phương sai của Y tại mỗi giá trị khác nhau của x là như nhau (phương sai không đổi).

Trong ví dụ tuổi của trẻ và lượng từ vựng, mô hình cho thấy lượng từ vựng trung bình của trẻ thay đổi tuyến tính với tuổi của trẻ và có phương sai không đổi tại mọi độ tuổi của trẻ.

Với mỗi x cố định Y là tổng của một hằng số $\beta_0 + \beta_1 x$ và sai số ngẫu nhiên có ϵ có phân phối chuẩn nên bản thân Y cũng có phân phối chuẩn.



Hình 12.4: (a) phân phối của ϵ, (b) phân phối của Y với các giá trị khác nhau của x.

Tham số σ^2 xác định khu vực phân tán của đường cong chuẩn so với giá trị trung bình (cũng là cao độ của đường thẳng. Khi σ^2 nhỏ một điểm quan sát (x, y) sẽ hầu hết rơi vào vị trí gần đường hồi quy đúng và ngược lại khi σ^2 lớn.

Ví dụ 12.1: Giả sử x và y có quan hệ hồi quy tuyến tính đơn giản với đường hồi quy

đúng là $y = 65 - 1.2x$ và $\sigma = 8$. Khi đó với giá trị cố định x^* bất kỳ y có giá trị trung bình là $65 - 1.2x^*$ và độ lệch chuẩn là 8.

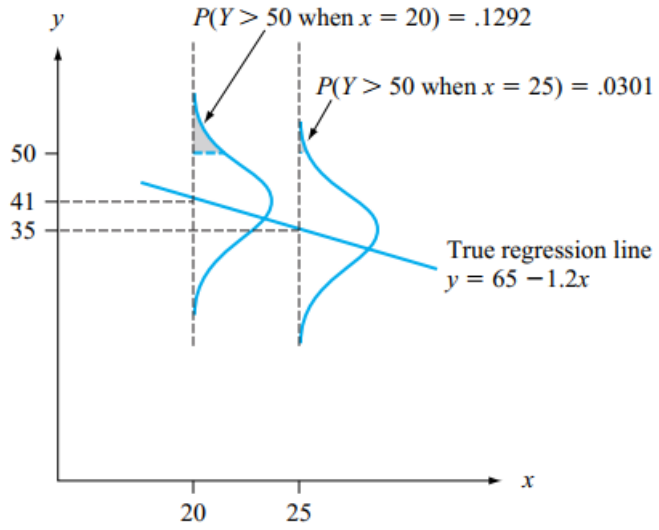
Với $x = 20$ biến ngẫu nhiên Y có giá trị trung bình $\mu_{Y.20} = 65 - 1.2(20) = 41$

$$P(Y > 50|x = 20) = p\left(Z = \frac{Y - 41}{8} > \frac{50 - 41}{8}\right) = 1 - \varnothing(1.13)$$
$$= 0.1292$$

Với $x = 25$ thì $\mu_{Y.25} = 65 - 1.2(25) = 35$

$$P(Y > 50|x = 25) = p\left(Z > \frac{50 - 35}{8}\right) = 1 - \varnothing(1.88)$$
$$= 0.0301$$

Các giá trị xác suất này được minh họa trong hình 12.5.



Hình 12.5: xác suất dựa trên mô hình hồi quy tuyến tính đơn giản.

Giả sử Y_1 là quan sát được $x = 25$ và Y_2 là quan sát được thực hiện khi $x = 24$

Khi đó $Y_1 - Y_2$ là biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình $E(Y_1 - Y_2) = E(Y_1) - E(Y_2) = \beta_1 = -1.2$, giá trị phương sai $V(Y_1 - Y_2) = V(Y_1) + V(Y_2) = \sigma^2 + \sigma^2 = 128$, giá trị độ lệch chuẩn $\sigma = \sqrt{128} = 11.314$. Xác suất Y_1 lớn hơn Y_2 là

$$P(Y_1 - Y_2 > 0) = p(Z > \frac{0 - (-1.2)}{11.314}) = p(Z > 0.11) = 0.4562$$

Tức là mặc dù ta kỳ vọng Y giảm khi x tăng 1 đơn vị, quan sát của Y tại $x + 1$ chưa chắc đã lớn hơn quan sát của Y tại x.

Bài tập 12.1

1. Tỷ số năng suất cho mẫu phép ngàm trong bể phốt phát được tính bằng cách lấy trọng lượng lớp phốt phát phủ ngoài chia cho lượng kim loại bị mất (cả hai cùng lấy đơn vị: mg/ft^2). Bài báo "Statistical Process Control of a Phosphate Coating Line" (Wire J. Intl., May 1997: 78–81) đưa ra bộ số liệu ghép cặp giữa nhiệt độ trong thùng (x) và tỷ số năng suất (y)

Temp.	170	172	173	174	174	175	176
Ratio	.84	1.31	1.42	1.03	1.07	1.08	1.04
Temp.	177	180	180	180	180	180	181
Ratio	1.80	1.45	1.60	1.61	2.13	2.15	.84
Temp.	181	182	182	182	182	184	184
Ratio	1.43	.90	1.81	1.94	2.68	1.49	2.52
Temp.	185	186	188				
Ratio	3.00	1.87	3.08				

(a) Biểu diễn biểu đồ gốc lá của nhiệt độ (x) và tỷ số năng suất (y).

(b) Giá trị của tỷ số hiệu suất được xác định đầy đủ và duy nhất ở nhiệt độ trong thùng hay không?

(c) Vẽ biểu đồ chấm của dữ liệu. Biểu đồ có cho thấy ta có khả năng dự đoán tỷ số năng suất y bởi nhiệt độ trong thùng ngàm x hay không? Giải thích.

2. Cho bộ số liệu ghép cặp

x	17	32	35	40	40	48	65	70	84	88	94	97
y	38	62	54	68	85	80	93	105	116	117	127	114
x	99	100	110	111	120	123	134	168	172	178	182	191
y	132	136	134	139	142	170	149	164	188	195	200	215

Vẽ biểu đồ chấm của dữ liệu trên. Biểu đồ có cho thấy mối quan hệ x và y?

3. Giả sử rằng biến ngẫu nhiên Y có quan hệ hồi quy tuyến tính đơn giản với tính độc lập X với phương trình hồi quy đúng có dạng $y = 1800 + 1.3x$

(a) Giá trị trung bình của Y là bao nhiêu khi $x = 2500$.

(b) Giá trị Y thay đổi trung bình là bao nhiêu khi x tăng thêm 1 đơn vị.

(c) Trả lời câu b khi x tăng thêm 100 đơn vị.

(d) Trả lời câu b khi x giảm đi 100 đơn vị.

4. Tiếp tục với giả thuyết trong bài 3, giả sử thêm độ lệch chuẩn của sai số ngẫu nhiên ε là 350.

(a) Tính xác suất giá trị của Y vượt quá 5000 khi giá trị của x là 2000.

(b) Trả lời câu a khi giá trị của x là 2500.

(c) Thực hiện hai quan sát độc lập tương ứng với $x = 2000$ và $x = 2500$. Tính xác suất quan sát thứ hai (tương ứng với $x = 2500$) lớn hơn quan sát thứ nhất (tương ứng với $x = 2000$) là 1000. Ký hiệu Y_1, Y_2 là các quan sát tương ứng với $x = x_1$ và $x = x_2$. Hỏi x_2 cần lớn hơn x_1 bao nhiêu để $P(Y_2 > Y_1) = 0.95$?

5. Tốc độ dòng chảy $y(m^3/\text{phút})$ trong một thiết bị sử dụng cho đo lường chất lượng khí phụ thuộc vào áp suất rơi x (đơn vị: inch) qua máy lọc thiết bị. Giả sử rằng giá trị của x trong khoảng 5 đến 20. Hai biến quan hệ qua mô hình hồi quy tuyến tính đơn giản với hàm hồi quy đúng $y = -0.12 + 0.95x$.

(a) Tính kỳ vọng tốc độ dòng chảy thay đổi tương ứng khi áp suất rơi x tăng.

(b) Tốc độ dòng chảy thay đổi trung bình bao nhiêu khi áp suất rơi giảm 5 inch.

(c) Tốc độ dòng chảy trung bình là bao nhiêu khi áp suất rơi là 10 inch? 15 inch?

(d) Giả sử $\sigma = 0.025$ và áp suất rơi là 10 inch. Tính xác suất giá trị tốc độ dòng chảy vượt quá 0.84?

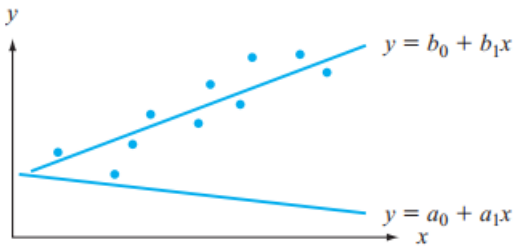
(e) Tính xác suất tốc độ dòng chảy khi áp suất rơi và là 10 inch vượt quá tốc độ dòng chảy khi áp suất rơi là 11 inch.

6. Biến phụ thuộc Y quan hệ với biến độc lập x qua mô hình hồi quy tuyến tính đơn giản có hàm hồi quy đúng $y = 4000 + 10x$. Biết rằng $P(Y > 5500|x = 100) = 0.05$ và $P(Y > 6500|x = 200) = 0.10$. Tính độ lệch chuẩn σ của sai số ngẫu nhiên ϵ.

12.2 Ước lượng hệ số hồi quy

Ta sẽ giả sử trong mục này và các mục khác là các biến x và y có quan hệ hồi quy tuyến tính đơn giản. Các giá trị β_0, β_1 và σ^2 sẽ hầu hết không biết được trong một nghiên cứu. Thay vào đó có mẫu dữ liệu gồm n cặp giá trị $(x_1, y_1), ..., (x_n, y_n)$ từ dữ liệu này sẽ ước lượng được mô hình tham số và đường hồi quy đúng. Các quan sát được giả thuyết thực hiện một cách độc lập tức là y_i là giá trị quan sát của Y_i với $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ và n độ lệch $\epsilon_1, \epsilon_2, ..., \epsilon_n$ là độc lập. Dẫn đến $Y_1, Y_2, ..., Y_n$ độc lập.

Qua mô hình các điểm quan sát sẽ bị phân phối bởi đường hồi quy đúng theo cách ngẫu nhiên. Hình 12.6 chỉ ra một kiểu các chấm của các quan sát của hai biến. Theo trực giác đường $y = a_0 + a_1 x$ không hợp lý để ước lượng đường hồi quy đúng $y = \beta_0 + \beta_1 x$ bởi vì nếu $y = a_0 + a_1 x$ là đường hồi quy đúng thì các điểm quan sát hầu hết phải nằm gần đường này. Dường thẳng $y = b_0 + b_1 x$ là ước lượng hợp lý bởi hầu hết các điểm nằm gần đường này hơn.



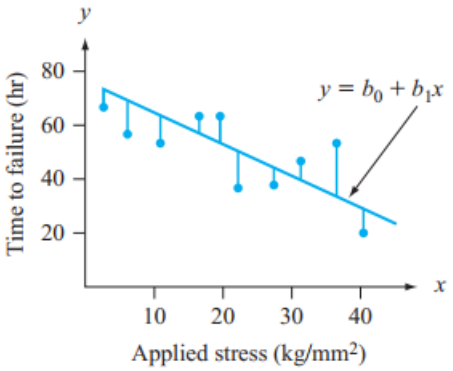
Hình 12.6: hai ước lượng khác nhau của đường hồi quy đúng.

Theo hình 12.6 và thảo luận đã nói dẫn đến ước lượng của $y = \beta_0 + \beta_1 x$ nên là đường thẳng phù hợp nhất với các điểm quan sát. Theo nguyên lý bình phương tối thiểu được đưa ra bởi nhà toán học người Đức - Gauss (1777 - 1855) đường thẳng phù hợp với dữ liệu là đường có khoảng cách thẳng đứng giữa các điểm quan sát và đường này là nhỏ nhất (xem hình 12.7). Để đo sự phù hợp ta lấy tổng các bình phương các độ lệch. Đường phù hợp nhất là đường có tổng các bình phương các độ lệch là nhỏ nhất.

Nguyên lý bình phương tối thiểu

Độ lệch thẳng đứng của điểm (x_i, y_i) với đường $y = b_0 + b_1 x$ là *(cao độ của điểm)* - *(chiều cao của đường)* $= y_i - (b_0 + b_1 x_i)$. Tổng các bình phương độ lệch thẳng đứng của các điểm $(x_1, y_1), ..., (x_n, y_n)$ tới đường thẳng là

$$f(b_0, b_1) = \sum_{x=1}^{n} [y_i - (b_0 + b_1 x_i)]^2.$$



Hình 12.7: độ lệch của các dữ liệu quan sát với đường $y = b_0 + b_1 x$

.

Ước lượng điểm của β_0 và β_1 ký hiệu là $\widehat{\beta}_0$ và $\widehat{\beta}_1$ và gọi là **ước lượng bình phương tối thiểu** là các giá trị làm cực tiểu $f(b_0, b_1)$. Tức là $\widehat{\beta}_0$ và $\widehat{\beta}_1$ thỏa mãn $f(\widehat{\beta}_1, \widehat{\beta}_1) \leq f(b_0, b_1)$ với mọi giá trị b_0 và b_1. **Đường hồi quy ước lượng** hay **đường bình phương tối thiểu** là đường có phương trình là $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

Đạo hàm riêng $f(b_0, b_1)$ theo b_0, b_1 và giải các phương trình các đạo hàm riêng bằng 0

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$
$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

Rút gọn nhân tử -2 và biến đổi tương đương ta có các phương trình sau gọi là **các phương trình chuẩn**

$$nb_0 + (\sum x_i)b_1 = \sum y_i$$
$$(\sum x_i)b_0 + (\sum x_i^2)b_1 = \sum x_i y_i$$

Các phương trình này là các phương trình tuyến tính theo hai ẩn b_0 và b_1. Cho rằng không phải tất cả các x_i là như nhau, ước lượng bình phương tối thiểu là nghiệm duy nhất của hệ phương trình này.

Ước lượng bình phương tối thiểu của hệ số dốc β_1 của đường hồi quy đúng là

$$b_1 = \widehat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{12.2}$$

Ước lượng bình phương tối thiểu của một hệ số tự do β_0 của đường hội quy đúng là

$$b_0 = \widehat{\beta}_0 = \frac{\sum y_i - \widehat{\beta}_1 \sum x_i}{n} = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{12.3}$$

Trong đó

$$S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)$$
$$S_{xx} = \sum(x_i - \overline{x})^2 = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2.$$

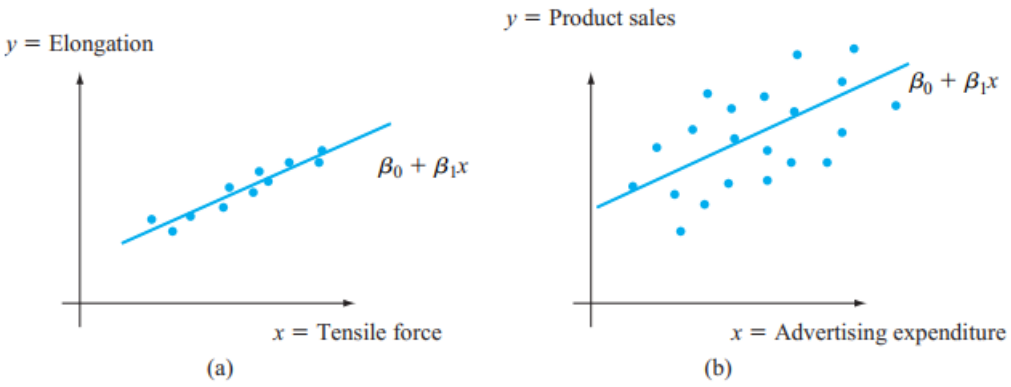
Ví dụ 12.2: Cho bộ số liệu ghép cặp

x	132	128	120	115	114	102	105	98	95	98
y	46	48	51	52	54	55	53	57	56	58
x	92	91	89	78	75	73	68	65	71	59
y	61	60	61	63	65	66	67	72	65	70

$$\sum x_i = 1868 \qquad \sum x_i^2 = 183166$$
$$\sum y_i = 1180 \qquad \sum x_i y_i = 107326$$
$$n = 20 \qquad \sum y_i^2 = 70614$$
$$S_{xy} = 8694.8 \qquad b_1 = -0.3319225284$$
$$n = -2886 \qquad b_0 = 90.00156415$$

Ước lượng σ^2 và σ

Hệ số σ^2 xác định lượng thay đổi vốn có trong mô hình hồi quy. Khi σ^2 lớn dẫn tới các quan sát (x_i, y_i) khá phân tán so với đường hồi quy thực, còn khi σ^2 nhỏ các điểm quan sát (x_i, y_i) tiến dần về đường hồi quy thực (xem hình 12.9). Ước lượng của σ^2 và quá trình kiểm định giả thuyết thống kê sẽ trình bày trong hai mục tiếp theo. Bởi vì phương trình đường thẳng là không biết, ước lượng được dựa trên phạm vi mẫu quan sát chệch so với đường ước lượng. Độ lệch lớn gợi ý giá trị σ^2 lớn còn các độ chệch có độ lớn là nhỏ thì gợi ý tới giá trị σ^2 là nhỏ.



Hình 12.9: Kiểu mẫu cho a) phương sai nhỏ, b) phương sai lớn

.

Định nghĩa 12.2:

Các giá trị thích hợp (hay dự đoán) $\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n$ tính được bằng cách thế lần lượt các giá trị $x_1, x_2, ..., x_n$ vào phương trình hồi quy ước lượng: $\widehat{y}_1 = \widehat{\beta}_0 + \widehat{\beta}_1 x_1, \widehat{y}_2 = \widehat{\beta}_0 + \widehat{\beta}_2 x_2, ..., \widehat{y}_n = \widehat{\beta}_0 + \widehat{\beta}_n x_n$ các phần dư là hiệu giữa các giá trị quan sát và giá trị dự đoán $y_1 - \widehat{y}_1, y_2 - \widehat{y}_2, ..., y_n - \widehat{y}_n$.

Thông thường các độ lệch với trung bình trong một mẫu được sử dụng tính giá trị ước lượng $s^2 = \sum \dfrac{(x_i - \overline{x})^2}{n-1}$, ước lượng của σ^2 trong phân tích hồi quy được dựa trên tổng các bình phương độ lệch và tiếp tục sử dụng kí hiệu s^2 cho ước lượng của phương sai nên đường bố đối với ký hiệu S^2 trước đó.

Định nghĩa 12.3: Tổng bình phương các sai số (hay tổng bình phương các độ lệch) kí hiệu là SSE là

$$SSE = \sum (y_i - \widehat{y}_i)^2 = \sum \left[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right]^2$$

và ước lượng của σ^2 là

$$\widehat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \widehat{y}_i)^2}{n-2}$$

Hiệu $n-2$ trong s^2 là bậc tự do (df) tương ứng của SSE và ướng lượng s^2,

Công thức tương đương để tính giá trị SSE là

$$SSE = \sum y_i^2 - \widehat{\beta}_0 \sum y_i - \widehat{\beta}_1 \sum x_i y_i.$$

Ví dụ 12.3: Cho bộ số liệu ghép cặp

x	15	23	35	40	48	55	61	69	75	83
y	4.9	4.7	4.5	4.6	4.4	4.1	4.2	4.0	3.8	3.7
x	90	98	102	108	112	119	125	132	140	
y	3.6	3.4	3.5	3.3	3.0	2.8	2.6	2.3	2.0	

Cỡ mẫu $n = 19$

$\sum x_i = 1530;$ $\sum x_i^2 = 149030$

$\sum y_i = 69.4;$ $\sum y_i^2 = 266$

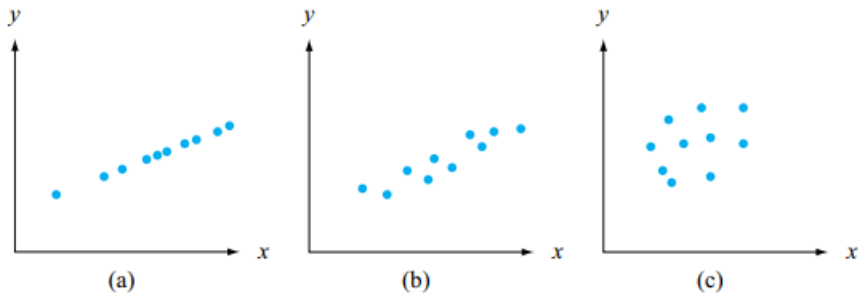
$\sum x_i y_i = 5032.5$

Từ đó ta tính được

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 25824.73684$$
$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = -556.0263158$$
$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -0.02153076406$$
$$\widehat{\beta}_0 = \frac{\sum y_i - \widehat{\beta}_1 \sum x_i}{n} = \overline{y} - \widehat{\beta}_1 \overline{x} = -5.386424685$$
$$SSE = 0.535696993.$$

Hệ số xác định (hệ số đo độ phù hợp của mô hình)

Hình 12.10 biểu diễn 3 biểu đồ chấm của mẫu ghép cặp. Trong cả 3 biểu đồ cao độ của các điểm khác nhau là biến thiên cho có sự thay đổi trong các giá trị quan sát y. Các điểm trong biểu đồ 1 chính xác thuộc một đường thẳng. Trong trường hợp này 100% mẫu quan sát được cho là x và y quan hệ tuyến tính. Biểu đồ 12.10(b) không rơi chính xác trên một đường, nhưng so sánh tất cả các độ chệch của y so với đường bình phương tối thiểu là nhỏ. Trong trường hợp này có lý do các giá trị khác nhau của y có thể tính được bằng cách xấp xỉ tuyến



Hình 12.10: Sử dụng mô hình để giải thích sự biến thiên của y: a) dữ liệu với mọi sự biến thiên giải thích được, b) dữ liệu với hầu hết sự biến thiên giải được, c) dữ liệu với rất ít sự biến thiên giải thích được.

tính với biến được yêu cầu bởi mô hình hồi quy tuyến tính đơn giản. Khi biểu đồ giống hình 12.10(c) có mức độ biến đổi giữa đường bình phương tối thiểu với các quan sát y nên mô hình hồi quy tuyến tính đơn giản không sử dụng để giải thích y bởi x.

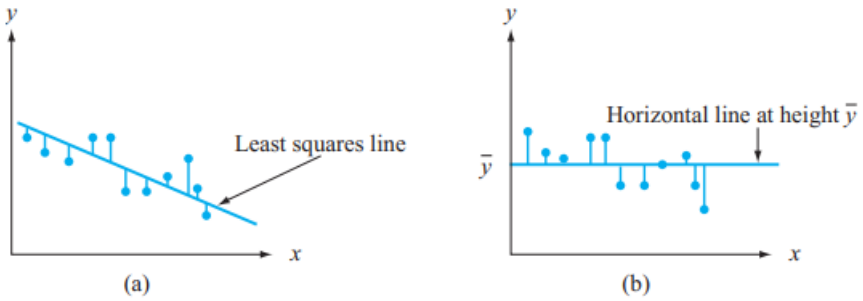
Tổng bình phương các sai số có thể được hiểu như một độ đo. Đo xem có bao nhiêu sự biến thiên của y không giải thích được bởi mô hình. Trong hình 12.10(a) $SSE = 0$ tức là không có sự biến thiên nào không được giải thích, SSE là nhỏ trong hình 12.10(b) và SSE lớn trong hình 12.10(c).

Một giá trị đo tổng lượng thay đổi các giá trị quan sát y cho bởi **tổng các bình phương**

$$SST = S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2.$$

Tổng các bình phương là tổng của các bình phương độ lệch với trung bình mẫu và các giá trị quan sát y. Giống như SSE là tổng của bình phương các độ lệch với đường bình phương tối thiểu $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$, SST là tổng các bình phương tối thiểu so với đường nằm ngang tại $\overline{y}$ (do đó các độ lệch thẳng đứng là $y_i - \overline{y}$) như hình 12.11. Hơn nữa vì tổng các bình phương độ lệch so với đường bình phương tối thiểu là nhỏ hơn tổng các bình phương tối thiểu so với các đường khác, SSE $<$ SST trừ khi đường nằm ngang là đường bình phương tối thiểu.

Tỷ số $\dfrac{SSE}{SST}$ là tỷ lệ của tổng các biến thiên mà không thể giải thích bởi mô hình hồi quy tuyến tính đơn giản và $1 - SSE/SST$ (là một số nằm giữa 0 và 1) là tỷ lệ của các biến thiên của y được giải thích bởi mô hình.



Hình 12.11: Minh họa tổng các bình phương: a) SSE: tổng các bình phương độ lệch so với đường bình phương tối thiểu b) SST: tổng các bình phương độ lệch so với đường nằm ngang.

Định nghĩa 12.4: Hệ số xác định ký hiệu là r^2, cho bởi công thức

$$r^2 = 1 - \frac{SSE}{SST}$$

Hệ số xác định như tỷ lệ sự thay đổi (biến thiên) của quan sát y có thể giải thích bởi mô hình hồi quy tuyến tính đơn giản.

Hệ số r^2 càng cao mô hình hồi quy đơn giản giải thích cho sự thay đổi của y càng tốt (thích hợp). Khi sử dụng phân tích hồi quy trong các phần mềm thống kê, r^2 hay $100r^2$ (số phần trăm sự biến thiên có thể giải thích bởi mô hình) là một kết quả sẽ được đưa ra. Nếu r^2 nhỏ nhà nghiên cứu sẽ mong muốn sử dụng mô hình khác (như hồi quy phi tuyến hay mô hình hồi bội có nhiều hơn một biến độc lập) để giải thích cho sự thay đổi của y.

Ví dụ 12.4: Tiếp tục ví dụ trên ta có:

$$\begin{aligned} SST &= \sum y_i^2 - \frac{1}{n}(\sum y_i)^2 \\ &= 266 - \frac{1}{19}69.4^2 = \frac{5941}{475} \\ &= 12.507368421052631578 44 \end{aligned}$$

$$\begin{aligned} SSE &= \sum y_i^2 - \widehat{\beta}_0 \sum y_i - \widehat{\beta}_1 \sum x_i y_i \\ &= 266 - (5.386424685).(69.4) + 0.0215307640650345 \\ &= 0.535696993 \end{aligned}$$

$$r^2 = 1 - \frac{SSE}{SST} = 0.957169488$$

Vẽ biểu đồ chấm

Bài tập 12.2

7. (a) Xác định phương trình đường bình phương tối thiểu cho số liệu ghép cặp trong bài 2.
   (b) Dự đoán giá trị của y khi biết $x = 35$ và tính giá trị phần dư tương ứng.
   (c) Tính SSE và một giá trị ước lượng điểm của σ.
   (d) Tỷ lệ sự biến thiên thay đổi của y có thể giải thích được quan hệ xấp xỉ tuyến tính giữa hai biến là bao nhiêu?

8. Bài báo "Characterization of Highway Runoff in Austin, Texas, Area" (J. of Envir. Engr., 1998: 131–137). Khảo sát về lượng mưa $x(m^3)$ và lượng nước thoát $y(m^3)$ tại một điểm cụ thể được bộ số liệu tương ứng

x	5	12	14	17	23	30	40	47
y	4	10	13	15	15	25	27	46
x	55	67	72	81	96	112	127	
y	38	46	53	70	82	99	100	

(a) Vẽ biểu đồ chấm cho dữ liệu này. Biểu đồ chấm có đưa tới việc dùng mô hình hồi quy tuyến tính đơn giản?

(b) Hãy tính một giá trị ước lượng điểm cho hệ số dốc và hệ số tự do của đường hồi quy tổng thể.

(c) Tính một giá trị ước lượng điểm cho lượng nước thoát trung bình khi lượng nước mưa là $50m^3$.

(d) Tính một giá trị ước lượng điểm cho độ lệch chuẩn.

(e) Tỷ lệ sự thay đổi của lượng nước thoát có thể giải thích bởi mối quan hệ hồi quy tuyến tính giữa lượng nước mưa x và lượng nước thoát y?

9. Cho bộ số liệu ghép cặp

x	102.3	87	82	76	92	89.2	85.5	93.5	79	76.7
y	85	81	67.7	58.7	84.3	83.3	78	69	67.5	58.3

(a) Xác định đường bình phương tối thiểu của dữ liệu và giải thích các hệ số.

(b) Tính hệ số xác định r^2 và giải thích.

(c) Tính và giải thích một giá trị ước lượng điểm cho độ lệch chuẩn σ trong mô hình hồi quy tuyến tính đơn giản.

10. Theo dữ liệu công bố trong bài báo "An Experimental Correlation of Oxides of Nitrogen Emissions from Power Boilers Based on Field Data" (J. of Engr. for Power, July 1973: 165–170) với x là tốc độ lan của đám cháy ($MBtu/hr - ft^2$) và y là tốc độ giải phóng ra khí $NO_x(ppm)$

x	100	125	125	150	150	200	200
y	150	140	180	210	190	320	280
x	250	250	300	300	350	400	400
y	400	430	440	390	600	610	670

(a) Giả sử mô hình hồi quy tuyến tính đơn giản có hiệu lực hãy xác định một ước lượng của đường hồi quy đúng.

(b) Ước lượng tốc độ giải phóng khí NO_x trung bình khi tốc độ lan tỏa của đám cháy là 225.

(c) Khi tốc độ lan tỏa của đám cháy giảm 50 thì tốc độ giải phóng khí NO_x thay đổi trung bình là bao nhiêu?

(d) Có thể dùng đường hồi quy ước lượng để dự đoán tốc độ giải phóng khí NO_x. Khi tốc độ lan của đám cháy là 500? Tại sao?

11. Cho bộ số liệu ghép cặp

x	0.05	0.1	0.15	0.21	0.28	0.32
y	0.45	0.53	0.52	0.59	0.67	0.89
x	0.37	0.41	0.46	0.51	0.57	0.63
y	0.88	0.93	1.05	1.36	1.48	1.72
x	0.7	0.76	0.82	0.89	0.92	0.98
y	1.69	1.85	1.81	2.01	2.38	2.39

(a) Tính giá trị ước lượng bình phương tối thiểu cho β_0, β_1 trong mô hình hồi quy tuyến tính đơn giản cho bộ số liệu ghép cặp này.

(b) Dự đoán giá trị của y khi $x = 0.5$.

(c) Tính một giá trị ước lượng cho σ.

(d) Tính giá trị của tổng các bình phương thay đổi SST và giá trị hệ số xác định r^2 và đưa ra bình luận về các giá trị này.

(e) Giả sử rằng thay vì tìm đường bình phương tối thiểu qua các điểm $(x_1, y_1), ..., (x_n, y_n)$ ta tìm đường bình phương tối thiểu qua các điểm $(x_1 - \overline{x}, y_1), ..., (x_n - \overline{x}, y_n)$. Vẽ biểu đồ chấm cho các điểm $(x_1, y_1), ..., (x_n, y_n)$ rồi vẽ biểu đồ chấm cho các điểm $(x_1 - \overline{x}, y_1), ..., (x_n - \overline{x}, y_n)$. Dùng các biểu đồ này để giải thích bằng trực quan mối quan hệ giữa các biểu đồ chấm và đường phương tối thiểu tương ứng.

(f) Giả sử thay vì tìm mô hình $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i (i = 1, 2, ..., n)$ ta tìm mô hình $y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \varepsilon_i (i = 1, 2, ..., n)$. Tìm ước lượng bình phương tối thiểu cho β_0^* và mối quan hệ với $\widehat{\beta}_0$ và $\widehat{\beta}_1$.

12.3 Kết luận về hệ số β_1

12.4 Kết luận liên quan đến $\mu_{y.x^*}$ và các giá trị dự đoán cho y

12.5 Hệ số tương quan

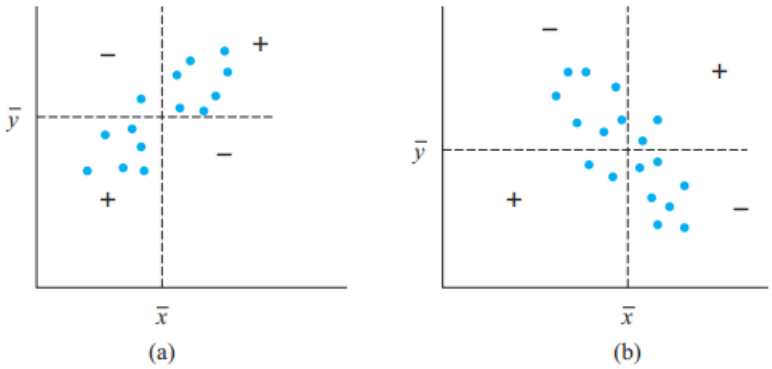
Có nhiều nghiên cứu cần chỉ ra có hay không mối quan hệ giữa hai biến hay đúng hơn là có thể sử dụng biến này để dự đoán biến kia hay không. Trong mục này trước tiên ta sẽ xét hệ số tương quan mẫu r như là một độ đo về mối quan hệ giữa hai biến x và y trong mẫu, sau đó xét về mối quan hệ giữa r và hệ số tương quan ρ đã định nghĩa trong chương 5. **Hệ số tương quan mẫu**

Cho n cặp số $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Rất tự nhiên khi ta nói x và y có mối quan hệ dương (đồng biến) nếu giá trị lớn của x ghép cặp với giá trị lớn của y và giá trị nhỏ của x ghép cặp với giá trị nhỏ của y. Còn x và y có mối quan hệ âm (nghịch biến) nếu giá trị lớn của x ghép cặp với giá trị nhỏ của y và giá trị nhỏ của x ghép cặp với giá trị lớn của y. Xét đại lượng:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

Nếu mối quan hệ dương là mạnh thì x_i lớn hơn $\overline{x}$ giá trị ghép cặp y_i tương ứng lớn hơn $\overline{y}$, do đó $(x_i - \overline{x})(y_i - \overline{y}) > 0$ là tích này cũng sẽ âm khi cả hai giá trị x_i, y_i nhỏ hơn các trung bình $\overline{x}, \overline{y}$ tương ứng. Suy ra x, y có mối quan hệ dương (đồng biến) thì S_{xy} sẽ dương. Lập luận tương tự chỉ ra rằng nếu x, y có mối quan hệ âm thì tích $(x_i - \overline{x})(y_i - \overline{y})$ sẽ âm. Minh họa trong hình 12.19

Mặc dù S_{xy} có vẻ như là một độ đo tin cậy đo mối quan hệ giữa x, y tuy nhiên khi thay đổi đơn vị của x hay y thì độ lớn sẽ thay đổi hoặc rất lớn hoặc rất gần 0. Ví dụ như bằng $S_{xy} = 25$ khi x có độ đo là m và $S_{xy} = 25000$ khi x có đơn vị là mm và $S_{xy} = 0.205$ khi x có đơn vị là km. Vì vậy ta cần một độ đo mối quan hệ giữa x và y mà không phụ thuộc vào đơn



Hình 12.19: (a) biểu đồ chấm với S_{xy} dương, (b) biểu đồ chấm với S_{xy} âm.
[+ nghĩa là trung bình $(x_i - \overline{x})(y_i - \overline{y}) > 0$ và - nghĩa là trung bình $(x_i - \overline{x})(y_i - \overline{y}) < 0$]

vị độ đo được sử dụng để đo các biến này. Từ đó hệ số tương quan mẫu được đưa ra từ việc điều chỉnh S_{xy}.

Định nghĩa 12.5:

Hệ số tương quan mẫu cho n cặp giá trị $(x_1, y_1), ..., (x_n, y_n)$ là

$$r = \frac{S_{xy}}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Tính chất của r

Ta có một số tính chất quan trọng của r như sau:

1. Giá trị của r không phụ thuộc vào biến nào đánh nhãn là x hay y.
2. Giá trị của r độc lập với các đơn vị dùng để đo biến x, y.
3. $-1 \leq r \leq 1$
4. $r = 1$ nếu và chỉ nếu mọi cặp (x_i, y_i) nằm trên một đường thẳng với hệ số góc dương và $r = -1$ nếu mọi cặp (x_i, y_i) cùng nằm trên một đường thẳng với hệ số góc âm.
5. Bình phương của hệ số tương quan mẫu là giá trị của hệ số xác định sự thích hợp của mô hình hồi quy tuyến tính đơn giản, theo ký hiệu $(r)^2 = r^2$.

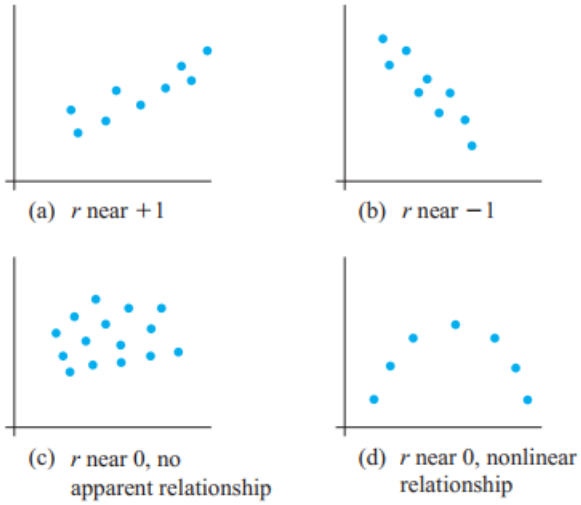
Tính chất 1 chỉ ra có sự tương phản trong phân tích hồi quy khi mọi đại lượng ta quan tâm (hệ số dốc, hệ số tự do, s^2, ...) phụ thuộc vào việc trong hai biến ta xem biến nào là biến phụ thuộc. Tuy nhiên tính chất 5 chỉ ra rằng tỷ lệ sự thay đổi (biến thiên) của biến phụ thuộc có thể giải thích được bởi mô hình hồi quy tuyến tính đơn giản không phụ thuộc vào việc biến nào có vai trò gì.

Cách phát biểu khác của tính chất 2 là giá trị của r không thay đổi nếu mỗi x_i được thay bởi cx_i và nếu mỗi y_i được thay bởi dy_i cũng như nếu mỗi x_i được thay bởi $x_i - a$ và mỗi y_i được thay bởi $y_i - b$.

Tính chất 3 cho rằng giá trị lớn nhất của r, tương ứng với quan hệ cực dương là 1 còn tương ứng với quan hệ âm nhất là $r = -1$. Theo tính chất 4, tương quan dương và âm nhất chỉ khi mọi điểm (x_i, y_i) đều nằm trên một đường thẳng. Các hình dạng khác của đám mây điểm

sẽ gợi ý về mối quan hệ số giữa các biến nếu trị tuyệt đối của r gần 1, suy ra đo mức độ phụ thuộc tuyến tính giữa các biến.

Một giá trị r gần bằng 0 ta có thể nói giữa các biến không có quan hệ tuyến tính nhưng không thể khẳng định không có mối quan hệ nào giữa các biến này. Hình 12.20 minh họa các hình dạng khác nhau của đám mây điểm tương ứng với giá trị khác nhau của r.



Hình 12.20 Biểu đồ chấm của các dữ liệu tương ứng với các giá trị r khác nhau.

Quy tắc kết luận về mức độ tương quan giữa hai biến mạnh hay yếu dựa trên giá trị của hệ số tương quan mẫu r.

Yếu	**Trung bình**	**Mạnh**
$-0.5 \le r \le 0.5$	$-0.8 < r < -0.5$ hoặc $0.5 < r < 0.8$	$r \ge 0.8$ hay $r \le -0.8$

Kết luận về hệ số tương quan của tổng thể

Hệ số tương quan mẫu r là một độ đo mức độ của quan hệ giữa x và y trong mẫu quan sát. Ta có thể xem các cặp giá trị (x_i, y_i) được rút ra từ tổng thể ghép cặp với (x_i, y_i) có hàm sát xuất hay mật độ đồng thời. Trong chương 5 ta định nghĩa hệ số tương quan

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$$

Trong đó

$$Cov(X,Y) = \begin{cases} \sum\limits_x \sum\limits_y (x-\mu_x)(y-\mu_y)p(x,y) & \text{với } (X,Y) \text{ rời rạc} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-\mu_x)(y-\mu_y)f(x,y) & \text{với } (X,Y) \text{ liên tục} \end{cases}$$

Nếu $p(x,y)$ hay $f(x,y)$ mô tả phân phối của cả tổng thể ghép cặp thì ρ là độ do mức độ mạnh của quan hệ giữa x và y trong tổng thể.

Hệ số tương quan tổng thể ρ là một tham số hay một đặc tính của tổng thể như μ_x, μ_y, σ_x và σ_y, do đó ta có thể sử dụng hệ số tương quan mẫu để đưa ra kết luận về ρ. cụ thể r là 1 giá trị ước lượng điểm cho ρ và ước lượng tương ứng là

$$\widehat{\rho} = R = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

Ví dụ 12.5: Cho mẫu ghép cặp

x	562	552	562	537	526	515	505	480	460	450	445	435	421	418	400
y	2.1	2.3	2.2	2.5	2.5	2.7	2.8	2.8	3.2	3.3	3.3	3.5	3.6	3.6	3.7

$$\begin{aligned} n &= 15 \\ \sum x_i &= 7268 \\ \sum x_i^2 &= 3565402 \\ \sum y_i &= 44.1 \\ \sum y_i^2 &= 133.89 \\ \sum x_i y_i &= 20940.6 \end{aligned}$$

Hệ số tương qua mẫu

$$\begin{aligned} r &= \frac{20940.6 - (7268)(44.1)/15}{\sqrt{3565402 - (7268)^2/15}\sqrt{133.89 - (44.1)^2/15}} \\ &= -0.9919050782 \end{aligned}$$

Một giá trị ước lượng điểm cho hệ số tươgn quan tổng thể ρ là $\hat{\rho} = r = -0.9919050782$.

Khoảng ước lượng với mẫu nhỏ về quá trình kiểm định trình bày trong chương 7 đến 9 dựa trên giả sử tổng thể có phân phối chuẩn. Để kiểm định giả thuyết về ρ cần giả thiết tương tự và phân phối của cặp giá trị (x, y) trong tổng thể. Bây giờ ta giả sử X và Y là ngẫu nhiên trong khi các hồi quy ta thực hiện trên các giá trị x cố định thu được từ các thí nghiệm.

Giả thiết

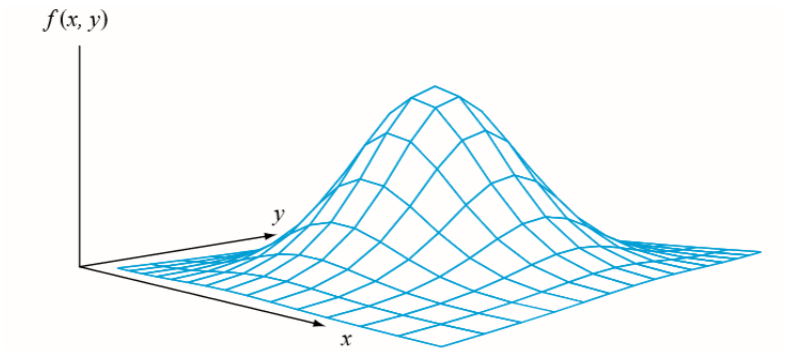
Phân phối xác suất đồng thời của (X, Y) là xác định bởi

$$f(x,y) = \frac{1}{2\pi.\sigma_1\sigma_2\sqrt{1-p^2}} e^{-[((x-\mu_1)/\sigma_1)^2 - 2p(x-\mu_1)(y-\mu_2)/\sigma_1\sigma_2 + ((y-\mu_2)/\sigma_2)^2]/[2(1-p^2)]}$$

$$-\infty < x < +\infty$$
$$-\infty < y < +\infty$$

trong đó μ_1 và σ_1 là trung bình và độ lệch chuẩn của X và μ_2 và σ_2 là trung bình và độ lệch chuẩn của Y. Hàm $f(x, y)$ gọi là **hàm mật độ của phân phối chuẩn 2 chiều**.

Phân phối chuẩn hai chiều hiển nhiên là phức tạp tuy nhiên với mục đích của chúng ta chỉ cần những hiểu biết sơ bộ về các tích chất của phân phối chuẩn hai chiều. Mặt cong xác định bởi hàm $f(x, y)$ nằm phía trên mặt phẳng xy (vì $f(x, y)$ không âm) và có hình dạng quả chuông trong không gian ba chiều, minh họa trong hình 12.21. Cắt mặt cong này bởi mặt phẳng vuông góc với mặt phẳng xy bất kì ta có giao tuyến là đường cong của hàm mât độ của phân phối chuẩn nào đó. Cụ thể hơn nếu $Y = x$ thì phân phối điều kiện của Y là chuẩn với trung bình $\mu_{Y.x} = \mu_2 - \rho\mu_1\frac{\sigma_2}{\sigma_1} + \rho\sigma_2\frac{x}{\sigma_1}$ và phương sai $(1-\rho^2)\sigma_2^2$. Dây chính xác là mô hình được sử dụng trong hồi quy tuyến tính đơn giản với $\beta_0 = \mu_2 - \rho\mu_1\frac{\sigma_2}{\sigma_1}, \beta_1 = \rho\frac{\sigma_2}{\sigma_1}$ và $\sigma^2 = (1-\rho^2)\sigma_2^2$ với x là biến độc lập. Dẫn tới nếu các cặp quan sát (x_i, y_i) thực sự được rút ra từ phân phối



Hình 12.21 Đồ thị hàm mật độ cua phân phối chuẩn hai chiều.

chuẩn hai chiều thì mô hình hồi quy tuyến tính đơn giản là xấp xỉ các nghiên cứu về Y với x cố định. Nếu $\rho = 0$ thì $\mu_{Y.x} = \mu_2$ tức là khi $\rho = 0$ hàm mật độ các suất đồng thời $f(x, y)$ có thể phân tích thành $f_1(x).f_2(y)$, dẫn tới X, Y là các biến độc lập . Giả sử rằng các cặp giá trị được rút ra từ một phân phối chuẩn hai chiều, ta đi kiểm định giả thuyết và ρ và xây dựng khoảng ước lượng cho ρ. Không có cách hoàn chỉnh thỏa mãn việc kiểm tra giả thuyết phân phối chuẩn hai chiều. Một cách không không hoàn chỉnh là xây dựng biểu đồ mẫu cho các x_i; biểu đồ mẫu cho các y_i. Phân phối chuẩn hai chiều được suy ra khi phân phối của X, Y đều chuẩn. Nếu có biểu đồ có hình dáng khác đáng kể so với biểu đồ phân phối chuẩn thì các bước kiểm định sau không nên sử dụng cho trường hợp cỡ mẫu n nhỏ.

Kiểm định giả thuyết hệ sô tương quan bằng 0.

Khi kiểm định giả thuyết $H_0 : \rho = 0$ là đúng thì tiêu chuẩn kiểm định

$$T = \frac{R\sqrt{n-2}}{sqrt1 - R^2}$$

có phân phối t với $n - 2$ bậc tự do.

Đối thuyết	Miền bác bỏ giả thuyết với mức ý nghĩa α
$H_a : \rho > 0$	$t \geq t_{\alpha,n-2}$
$H_a : \rho < 0$	$t \leq -t_{\alpha,n-2}$
$H_a : \rho \neq 0$	hoặc $z \geq t_{\alpha/2,n-2}$ hoặc $z \leq -t_{\alpha/2,n-2}$

Giá trị P xác suất dựa trên $n - 2$ bậc tự do có thể tính dựa trên các mô tả trên.

Lập luận khác liên quan đến ρ

Mệnh đề

Khi $(X_1, X_1), ..., (X_n, Y_n)$ là một mẫu lấy từ tổng thể có phân phối chuẩn hai chiều, biến ngẫu nhiên

$$V = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

có phân phối xấp xỉ chuẩn với trung bình và phương sai

$$\mu_V = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \qquad \sigma_V^2 = \frac{1}{n-3}$$

Khi n khá nhỏ thì xấp xỉ này không có hiệu lực.

Tiêu chuẩn kiểm định cho giả thuyết $H_0 : \rho = \rho_0$ là

$$Z = \frac{V - \frac{1}{2}\ln[(1+\rho_0)/(1-\rho_0)]}{1/\sqrt{n-3}}$$

Đối thuyết	Miền bác bỏ giả thuyết với mức ý nghĩa α
$H_a : \rho > \rho_0$	$z \geq z_\alpha$
$H_a : \rho < \rho_0$	$z \leq -z_\alpha$
$H_a : \rho \neq \rho_0$	hoặc $z \geq z_{\alpha/2}$ hoặc $z \leq -z_{\alpha/2}$

P xác suất có thể tính tương tự theo cách kiểm định z ở trước.

Ví dụ 12.6: Cho bộ số liệu ghép cặp

x	55.10	44.83	46.32	51.10	49.89	45.20	48.18	46.70	54.31	41.50
y	49.10	31.20	32.80	42.60	42.50	32.70	36.21	40.40	37.42	30.80
x	47.50	52.00	52.25	50.86	51.66	54.77	57.06	57.84	55.22	
y	36.34	44.80	41.75	39.35	44.07	43.40	45.30	39.08	41.89	

$S_{xx} = 367.74, S_{yy} = 488.54, S_{xy} = 322.37$ từ đó tính được $r = 0.761$. Ta có thể kết luận mức độ tương quan giữa x, y ít nhất là trung bình?

Trong phần trước ta kết luận mức độ tương quan tuyến tính là trung bình khi $0.5 < \rho < 0.8$, vì vậy ta kiểm định giả thuyết $H_0 : \rho = 0.5$ với đối thuyết $H_a : \rho > 0.5$.

Ta tính được các giá trị $v = 0.5ln(\frac{1+0.761}{1-0.761}) = 0.999$ và $0.5ln(\frac{1+0.5}{1-0.5}) = 0.549$
Suy ra $z = (0.999 - 0.549)\sqrt{19-3} = 1.80$. P giá trị cho kiểm định một phía lớn hơn là 0.0359. Giả thuyết không bị bác bỏ với mức ý nghĩa 5% nhưng được chấp nhận với mức ý nghĩa 1%.

Để thu được ước lượng cho ρ trước tiên ta xuất phát từ khoảng cho $\frac{1}{2}\ln[(1+\rho_0)/(1-\rho_0)]$. Chuẩn hóa V, viết biểu thức xác suất và biến đổi các bất đẳng thức ta thu được

$$\left(v - \frac{z_{\alpha/2}}{\sqrt{n-3}}; v + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)$$

là một khoảng ước lượng cho μ_V với độ tin cậy $100(1-\alpha)\%$ với $v = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$. Dẫn tới ta có khoảng ước lượng cho ρ.

Khoảng ước lượng với độ tin cậy $100(1-\alpha)\%$ cho hệ số tương quan ρ là

$$\left(\frac{e^{2C_1}-1}{e^{2C_1}+1}; \frac{e^{2C_2}-1}{e^{2C_2}+1}\right)$$

với $C_1 = V - \frac{z_{\alpha/2}}{\sqrt{n-3}}; C_2 = V + \frac{z_{\alpha/2}}{\sqrt{n-3}}$

Ví dụ 12.7: Giả sử số liệu ghép cặp có thống kê mẫu $\sum x_i = 285.90, \sum x_i^2 = 4409.55, \sum y_i = 690.30, \sum y_i^2 = 29040.29$ và $\sum x_i y_i = 10818.56$. Hệ số tương quan mẫu $r = 0.733$, từ đó tính được $v = 0.935$. Với cỡ mẫu $n = 20$, khoagr tin cậy 95% cho μ_v là

$$(0.935 - 1.96/\sqrt{17}, 0.935 + 1.96/\sqrt{17}) = (0.460, 1.410) = (c_1, c_2).$$

Khoảng tin cậy 95% cho ρ là

$$[\frac{e^{2.(0.46)}-1}{e^{2.(0.46)}+1}, \frac{e^{2.(1.41)}-1}{e^{2.(1.41)}+1}] = (0.43, 0.89).$$