Vũ Thị Hương

MÔN HỌC: XÁC SUẤT THỐNG KÊ

1st April 2021

Mục lục

5.1 Hồi quy tuyến tính

Hàm hồi quy

Cho hai biến X, Y. Tương ứng kỳ vọng có điều kiện $f(X) = E[Y|X]$ là ước lượng tốt nhất của Y theo X theo nghĩa bình phương tối thiểu.

Hàm hồi quy của Y theo X

Quan hệ hàm số giữa X, Y xác định bởi $x \mapsto f(x) = E[Y|X = x]$, hay là $f(x) = E[Y|X = x]$ được hiểu là hàm hồi quy của Y theo X.

Các mô hình hồi quy

- Hàm hồi quy dạng tuyến tính $E[Y|X = x] = \beta_0 + \beta_1 x$.
- Hàm hồi quy dạng đa thức (bậc 2, bậc 3, . . .)
  $E[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2, v.v.$
- Hàm hồi quy dạng hyperbol: $E[Y|X = x] = \beta_0 + \dfrac{\beta_1}{x}$.
- Hàm hồi quy dạng mũ: $E[Y|X = x] = \beta_0 e^{\beta_1 x}$.

5.1 Hồi quy tuyến tính

Mô hình hồi quy tuyến tính đơn

Mô hình hồi quy tuyến tính đơn

$$Y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2). \tag{5.2}$$

Biến ε có trung bình là 0 nên hàm hồi quy của Y theo X là hàm tuyến tính

$$E[Y|x] = \beta_0 + \beta_1 x. \tag{5.3}$$

5.1 Hồi quy tuyến tính

Ước lượng hệ số hàm hồi quy từ mẫu thực nghiệm

Cho mẫu thực nghiệm:

$$(x_1, y_1), (x_2, y_3), \ldots, (x_n, y_n)$$

Xác định hệ số hồi quy sao cho đường thẳng $y = \beta_0 + \beta_1 x$ đi qua n điểm quan sát với sai số ít nhất có thể.
Ta biểu diễn

$$y_i = \beta_0 + \beta_1 x + e_i, \quad i = 1, 2, \ldots, n.$$

Gọi $\hat{y}_i = \beta_0 + \beta_1 x$ là giá trị dự báo theo mô hình tuyến tính.
Tương ứng

$$e_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x$$

là phần dư (sai số) giữa thực nghiệm và lý thuyết. Phần dư e_i cho thấy mức phù hợp với mô hình của quan sát thứ i.

5.1 Hồi quy tuyến tính

Ước lượng hệ số hàm hồi quy từ mẫu thực nghiệm

- Giá trị ước lượng của hệ số hồi quy β_0, β_1 được ký hiệu tương ứng là $\hat{\beta}_0, \hat{\beta}_1$.
- Phương pháp xác định là "phương pháp bình phương tối thiểu": Giá trị ước lượng ứng với tổng bình phương các sai số nhỏ nhất có thể, hay là

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- Các giá trị ước lượng $\hat{\beta}_0, \hat{\beta}_1$ là điểm dừng của hàm E:

$$\begin{cases} \frac{\partial E}{\partial \beta_0} & = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial E}{\partial \beta_1} & = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

5.1 Hồi quy tuyến tính

Ví dụ 5.1

Hệ phương trình chuẩn tắc

$$\begin{cases} n\beta_0 + \sum_{i=1}^{n} x_i\beta_1 & = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\beta_0 + \sum_{i=1}^{n} x_i^2\beta_1 & = \sum_{i=1}^{n} y_i x_i \end{cases} \tag{5.5}$$

Ví dụ 5.1

Kiểm tra mức độ phát thải khí nitro N_2O (mg/lít) do các xe ô-tô khi dừng phát ra ta có bảng số liệu sau:

x_i (giây)	2	2	4	6	10	10
y_i (mg/lít)	10	11	20	27	52	55

Hãy xác định hàm hồi quy tuyến tính biểu thị mối liên hệ mức độ ô nhiễm khí nitro đối với thời gian dừng của ô-tô.

5.1 Hồi quy tuyến tính

Ví dụ 5.1

Từ bảng số liệu trên ta tính các hệ số trong hệ phương trình chuẩn tắc

STT	x_i	y_i	x_iy_i	x_i^2
1	2	10	20	4
2	2	11	22	4
3	4	20	80	16
4	6	27	162	36
5	10	52	520	100
6	10	55	550	100
Tổng	34	175	1354	260

Ta có kích thước mẫu $n = 6$ và các số

$$\sum x_i = 34, \sum y_i = 175, \sum x_iy_i = 1354, \sum x_i^2 = 260.$$

5.1 Hồi quy tuyến tính

Ví dụ 5.1

Giả sử đường hồi quy tuyến tính là $y = \beta_0 + \beta_1 x$. Hệ phương trình chuẩn tắc là

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i & = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} y_i x_i \end{cases} \Leftrightarrow \begin{cases} 6\beta_0 + 34\beta_1 & = 175 \\ 34\beta_0 + 260\beta_1 & = 1354. \end{cases}$$

Giải hệ phương trình trên ta tính được hệ số dốc $\hat{\beta}_1 = 5,3812$ và hệ số tự do $\hat{\beta}_0 = -1,3267$. Vậy hàm hồi quy tuyến tính thực nghiệm là $y = -1,3267 + 5,3812x$.

5.1 Hồi quy tuyến tính

Ước lượng điểm của hệ số hồi quy

Ta coi các giá trị x_i là cố định và các giá trị quan sát Y_i là các biến ngẫu nhiên. Ta gọi các ước lượng điểm của các hệ số tự do β_0 và độ dốc β_1 đối với các dữ liệu $(x_i, Y_i), i = 1, 2, \ldots, n$ là B_0, B_1.

Nghiệm của hệ phương trình chuẩn tắc

$$\begin{cases} \hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i . \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ \hat{\beta}_0 = \dfrac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \dfrac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}. \end{cases}$$

hay là

$$\begin{cases} \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 = \bar{y} - \dfrac{S_{xy}}{S_{xx}} \bar{x} \end{cases}$$

5.1 Hồi quy tuyến tính

Ước lượng điểm của hệ số hồi quy

Ở đây S_{xx}, S_{xy}, S_{yy} được cho bởi

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}.\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

5.1 Hồi quy tuyến tính

Ước lượng điểm của hệ số hồi quy

Ước lượng điểm của hệ số β_1 và β_0 là

$$B_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i, \qquad B_0 = \bar{Y} - B_1\bar{x},$$

trong đó $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \beta_0 + \beta_1\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i = \beta_0 + \beta_1\bar{x} + \bar{\epsilon}.$

Định lý 5.1

Ước lượng điểm B_0, B_1 là các ước lượng không chệch của hệ số β_0, β_1. Hơn nữa, phương sai của các ước lượng này là

$$\mathbb{V}[B_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right), \mathbb{V}[B_1] = \frac{\sigma^2}{S_{xx}}.$$

5.1 Hồi quy tuyến tính

Ước lượng điểm cho phương sai của mô hình hồi quy

Đặt $T = \sum_{i=1}^{n} \left(Y_i - B_0 - B_1 x_i\right)^2$. Ta chỉ ra được

$$\mathbb{E}[T] = (n-2)\sigma^2.$$

Định lý 5.2

Ước lượng điểm $S^2 = \dfrac{T}{n-2}$ là ước lượng không chệch của giá trị phương sai $\sigma^2 = \mathbb{V}[\epsilon]$ trong mô hình hồi tuyến tính (5.1). Một điểm ước lượng tương ứng với mẫu thực nghiệm là

$$s^2 = \frac{E}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}(S_{yy} - \frac{S_{xy}^2}{S_{xx}}).$$

và gọi là sai số bình phương trung bình.

5.1 Hồi quy tuyến tính

Công thức khoảng tin cậy cho các hệ số hồi quy

Định lý 5.3

Khoảng tin cậy $\gamma = 1 - \alpha$ đối hệ số độ dốc β_1 là

$$\hat{\beta}_1 - t_{n-2,\alpha/2}\frac{s}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 - t_{n-2,\alpha/2}\frac{s}{\sqrt{S_{xx}}},$$

và khoảng tin cậy $\gamma = 1 - \alpha$ đối hệ số tự do β_0 là

$$\hat{\beta}_0 - t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} < \beta_0 < \hat{\beta}_0 + t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

5.1 Hồi quy tuyến tính

Kiểm định các giả thuyết về hệ số hồi quy

Định lý 5.4

a) Kiểm định giả thuyết thống kê $H_0 : \beta_1 = a; H_1 : \beta_1 \neq a$ với mức ý nghĩa α. Khi giả thuyết H_0 đúng, thống kê $T = \dfrac{B_1 - a}{\sqrt{S^2/S_{xx}}}$ có phân phối Student với $n-2$ bậc tự do. Bác bỏ giả thuyết H_0 nếu $|t| > t_{n-2,\alpha/2}$.

b) Kiểm định giả thuyết thống kê $H_0 : \beta_0 = b; H_1 : \beta_0 \neq b$ với mức ý nghĩa α. Khi giả thuyết H_0 đúng, thống kê $T = \dfrac{B_0 - b}{\sqrt{S^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}}$ có phân phối Student với $n-2$ bậc tự do. Bác bỏ giả thuyết H_0 nếu $|t| > t_{n-2,\alpha/2}$.

5.1 Hồi quy tuyến tính

Kiểm định các giả thuyết về hệ số hồi quy

Đối với sự kiểm định về sự tác động của hệ số dốc β_1 nghĩa là ta kiểm định giả thuyết $H_0 : \beta_1 = 0$ và đối thuyết $\beta_1 \neq 0$ ta có thể sử dụng phương pháp phân tích phương sai. Việc kiểm định này nhằm mục đích đánh giá về ý nghĩa của hàm hồi quy tuyến tính.

Nếu bác bỏ H_0 thì hàm hồi quy tuyến tính là có ý nghĩa. Trong trường hợp này thống kê để kiểm định khi H_0 đúng là thống kê $T = \dfrac{B_1}{\sqrt{S^2/S_{xx}}}$ có phân phối Student với $\nu = n - 2$ bậc tự do.

5.1 Hồi quy tuyến tính

Kiểm định các giả thuyết về hệ số hồi quy

Trong phân tích phương sai thì người ta hay sử dụng phân phối Fisher thay cho phân phối Student. Phân phối

$$F = T^2 = \frac{B_1^2 S_{xx}}{S^2}$$

có phân phối F với 1 và $\nu = n - 2$ bậc tự do. Bác bỏ giả thuyết H_0 nếu giá trị $f = \dfrac{S_{yy} - s^2}{s^2}$ lớn hơn giá trị mức phân vị $f_{1,n-2,\alpha}$.

Chú ý rằng với kiểm định Student ở trên thì ta có thể kiểm định một phía đối với hệ số dốc. Đối với kiểm định bằng phân phối F ta chỉ có thể kiểm định hai phía đối với hệ số này.

5.1 Hồi quy tuyến tính

Kiểm định các giả thuyết về hệ số hồi quy

Ví dụ 5.2

Thống kê về chỉ số BMI (kg/m^2) và độ tuổi của mẫu có kích thước $n = 9$ ta có bảng số liệu sau:

BMI	19,92	20,59	29,02	20,78	25,97	20,39	23,29	17,27	35,24
Tuổi	45	34	40	33	28	30	52	33	47

a) Tìm hàm hồi quy tuyến tính của BMI theo tuổi.

b) Tìm khoảng tin cậy của các hệ số hồi quy với độ tin cậy 95%.

Giải: Giả sử hàm hồi quy tuyến tính có dạng $y = \beta_0 + \beta_1 x$, với y là chỉ số BMI và x là số tuổi.

a) Hệ phương trình chuẩn tắc là

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i & = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} y_i x_i \end{cases}$$

5.1 Hồi quy tuyến tính

Kiểm định các giả thuyết về hệ số hồi quy

Ta lập bảng tính

STT	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	45	19,92	896,4	2025	396,8064
2	34	20,59	700,06	1156	423,9481
3	40	29,02	1160,8	1600	842,1604
4	33	20,78	685,74	1089	431,8084
5	28	25,97	727,16	784	674,4409
6	30	20,39	611,7	900	415,7521
7	52	23,29	1211,08	2704	542,4241
8	33	17,27	569,91	1089	298,2529
9	47	35,24	1656,28	2209	1241,8576
Tổng	342	212,47	8219,13	13556	5267,4509

Tương ứng ta thu được

$$\bar{x} = \frac{\sum x_i}{n} = 38,$$

5.1 Hồi quy tuyến tính

Kiểm định các giả thuyết về hệ số hồi quy

$$\bar{y} = \frac{\sum y_i}{n} = 23,6078,$$

$$S_{xy} = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i = 145,27,$$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\Big(\sum x_i\Big)^2 = 560,$$

$$S_{yy} = \sum y_i^2 - \frac{1}{n}\Big(\sum y_i\Big)^2 = 251,5064.$$

Hệ số hồi quy tuyến tính thực nghiệm là

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0,2594,$$

$$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} = 13,7502.$$

Vậy hàm hồi quy tuyến tính thực nghiệm là
$y = 13,7502 + 0,2594x.$

b) Với độ tin cậy $\gamma = 0,95 = 1 - \alpha$, nên $\alpha = 0,05$. Giá trị mức phân vị $t_{n-2;\alpha/2} = t_{7;0,025} = 2,365$. Ta có

$$s^2 = \frac{1}{n-2}(S_{yy} - \hat{\beta}_1 S_{xy}) = 30,5460.$$

Khoảng tin cậy cho hệ số độ dốc β_1 là

$$\begin{aligned}
&\left(\hat{\beta}_1 - t_{n-2,\alpha/2}\frac{s}{\sqrt{S_{xx}}}; \hat{\beta}_1 - t_{n-2,\alpha/2}\frac{s}{\sqrt{S_{xx}}}\right) \\
&= \left(0,2594 - 2,365\frac{5,5268}{23,6643}; 0,2594 - 2,365\frac{5,5268}{23,6643}\right) \\
&= (-0,2929; 0,8117).
\end{aligned}$$

5.1 Hồi quy tuyến tính

Kiểm định các giả thuyết về hệ số hồi quy

Khoảng tin cậy cho hệ số tự do β_0 là

$$\left(\hat{\beta}_0 - t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}; \hat{\beta}_0 + t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}\right)$$

$$= \left(13,7502 - 2,365\sqrt{30,5460(\frac{1}{9}+\frac{38^2}{560})};\right.$$

$$\left. 13,7502 + 2,365\sqrt{30,5460(\frac{1}{9}+\frac{38^2}{560})}\right)$$

$$= \Big(-7,68323; 35,1835\Big).$$

5.1 Hồi quy tuyến tính

Giá trị dự báo

Giả sử ta đã xây dựng được hàm hồi quy tuyến tính thực nghiệm là $y = \hat{\beta}_0 + \hat{\beta}_1 x$. Với giá trị x_0 của biến độc lập,thì ước lượng điểm của $y_0 = \mathbb{E}[Y|x_0] = \beta_0 + \beta_1 x_0$ tương ứng với x_0 là $\hat{Y}_0 = B_0 + B_1 x_0$.

Định lý 5.5

a) Ước lượng điểm $\hat{Y}_0 = B_0 + B_1 x_0$ tại giá trị x_0 là ước lượng không chệch của giá trị $y_0 = \beta_0 + \beta_1 x_0$. Hơn nữa, phương sai của ước lượng này là $\mathbb{V}[\hat{Y}_0] = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$.

b) Khoảng tin cậy với độ tin cậy $\gamma = 1 - \alpha$ cho giá trị y_0 là

$$\hat{y}_0 - t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} < y_0 < \hat{y}_0 + t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

5.1 Hồi quy tuyến tính

Giá trị dự báo

Ví dụ 5.3

Số liệu về điểm thi của 9 sinh viên gồm điểm thành phần (x) và điểm thi cuối kỳ (y) như sau:

x_i	$7,5$	5	7	7	8	$9,5$	$9,5$	10	$6,5$
y_i	8	$6,5$	8	5	$8,5$	10	10	10	$6,5$

a) Xác định hàm hồi quy tuyến tính biểu thị mối liên hệ giữa điểm thi cuối kỳ và điểm thành phần.

b) Nếu một sinh viên có điểm thành phần bằng 6 thì điểm thi cuối kỳ là bao nhiêu và tìm khoảng tin cậy 95% cho giá trị điểm thi tương ứng.

5.1 Hồi quy tuyến tính

Giá trị dự báo

Từ bảng số liệu trên ta lập bảng tính

STT	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	7,5	8	60	56,25	64
2	5	6,5	32,5	25	42,25
3	7	8	56	49	64
4	7	5	35	49	25
5	8	8,5	68	64	72,25
6	9,5	10	95	90,25	100
7	9,5	10	95	90,25	100
8	10	10	100	100	100
9	6,5	6,5	42,25	42,25	42,25
Tổng	70	72,5	583,75	566	609,75

Ta nhận được: $n = 9, \sum x_i = 70, \sum y_i = 72,5, \sum x_iy_i = 583,75,$ $\sum x_i^2 = 566, \sum y_i^2 = 609,75.$

5.1 Hồi quy tuyến tính

Giá trị dự báo

Từ đó ta tính được các giá trị
$\bar{x} = 7,7778, \bar{y} = 8,0556, S_{xy} = 19,8611, S_{xx} = 21,5556, S_{yy} = 25,722$

a) Giả sử hàm hồi quy tuyến tính cần tìm là $y = \beta_0 + \beta_1 x$ ta có hệ số hồi quy tuyến tính thực nghiệm là

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{19,8611}{21,5556} = 0,9214.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8,0556 - 0,9214.7,7778 = 0,8892.$$

Vậy hàm hồi quy tuyến tính thực nghiệm cần tìm là

$$y = 0,9214x + 0,8892.$$

b) Nếu sinh viên có điểm thành phần là $x_0 = 6$ thì điểm thi kết thúc học phần được ước lượng theo công thức

$$y_0 = \beta_0 + \beta_1 x = 0,9214.6 + 0,8892 = 6,4176.$$

5.1 Hồi quy tuyến tính

Giá trị dự báo

Với độ tin cậy $\gamma = 0,95 = 1 - \alpha$ với $\alpha = 0,05$. Giá trị điểm ước lượng cho phương sai là

$$s^2 = \frac{1}{n-2}(S_{yy} - \frac{S_{xy}^2}{S_{xx}}) = 1,0603.$$

Giá trị phân vị $t_{n-2,\alpha/2} = t_{7;0,025} = 2,365$. Khoảng tin cậy cho giá trị dự báo tại $x_0 = 6$ là

$$\left(\hat{y}_0 - t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}; \hat{y}_0 + t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right)$$

Ta tính phần sai số

$$2,365\sqrt{1,0603\left(\frac{1}{9} + \frac{(6 - 7,7778)^2}{21,5556}\right)} = 1,2361.$$

Thay số vào ta có khoảng tin cậy cho giá trị dự báo y_0 tại x_0 là

$$\Big(6,4176 - 1,2361; 6,4176 + 1,2361\Big) = (5,1815; 7,6537).$$

5.1 Hồi quy tuyến tính

Hệ số xác định

Sai số trong mô hình hồi quy tuyến tính

$$S_E = E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

biến thiên tổng thể là đại lượng

$$S_T = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Định nghĩa 5.1

Hệ số xác định trong mô hình hồi quy tuyến tính ký hiệu R^2 xác định bởi công thức

$$R^2 = \frac{S_T - S_E}{S_T} = 1 - \frac{S_E}{S_T} \tag{5.6}$$

5.1 Hồi quy tuyến tính

Hệ số xác định

Phương pháp tính hệ số xác định Ta biết công thức tính tổng bình phương sai số

$$S_E = S_{yy} - \hat{\beta}_1 S_{xy} = S_{yy} - \frac{S_{xy}}{S_{xx}} S_{xy}.$$

Do đó hệ số xác định được tính bằng công thức

$$R^2 = \frac{S_T - S_E}{S_T} = \frac{S_{xy}^2}{S_{xx} S_{xy}}.$$

Ví dụ 5.4

Trong Ví dụ 5.3 ta có
$S_{xy} = 19,8611; S_{xx} = 21,5556; S_{yy} = 25,7222$ do đó hệ số xác định bằng

$$R^2 = \frac{19,8611^2}{21,5556.25,7222} = 0,7114.$$

Nghĩa là có khoảng $71,14\%$ các giá trị được giải thích bởi mô hình hồi quy tuyến tính.

5.1 Hồi quy tuyến tính

Hệ số xác định

Trong thống kê hệ số xác định được dùng thường xuyên để đánh giá độ chính xác của mô hình hồi quy tuyến tính. Câu hỏi đặt ra là với mức nào của hệ số xác định thì dùng mô hình hồi quy tuyến tính. Trong thực hành người ta thường sự dụng mô hình hồi quy tuyến tính nếu giá trị của R^2 lớn hơn $0,7$. Đối với các trường hợp đặc biệt thì người ta xác định giá trị này phụ thuộc vào kinh nghiệm để quyết định có sử dụng mô hình hồi quy tuyến tính hay không.

Hàm xu thế tuyến tính Xét trường hợp một biến $x = t$ là biến thời gian thì hàm hồi quy tuyến tính dạng

$$y = \beta_0 + \beta_1 t$$

được gọi là hàm dự báo tuyến tính. Các hệ số hồi quy được tính tương tự như trường hợp hồi quy tuyến tính dựa vào dãy dữ liệu (t_i, y_i), $i = 1, 2, \ldots, n$.

5.1 Hồi quy tuyến tính

Đổi biến trong hồi quy tuyến tính

Chú ý 5.1

Đối với trường hợp tìm hàm dự báo tuyến tính, để giảm khối lượng tính toán ta sử dụng dịch chuyển mốc thời gian. Chẳng hạn chọn mốc thời gian là t_0 thì các mốc mới được tính là các mốc $t_i - t_0$. Khi đó ta tìm hàm hồi quy tuyến tính theo mốc mới và giá trị cần dự báo cũng phải được tính tương ứng với giá trị dịch chuyển tương ứng.

Ví dụ 5.5

Theo dõi về thu nhập GDP bình quân đầu người đơn vị (100 USD/năm) của Việt Nam quan các năm ta có bảng số liệu sau:

Năm	2005	2006	2007	2008	2009	2010	2011	2012
Thu nhập	7	7,96	9,19	11,45	11,60	12,73	15,17	17,49

a) Tìm hàm dự báo tuyến tính về thu nhập bình quân đầu người.
b) Dự báo về giá trị năm 2016.

5.1 Hồi quy tuyến tính

Đổi biến trong hồi quy tuyến tính

Giải: Từ bảng số liệu trên ta chuyển đổi mốc thời gian năm 2005 ứng với $t_1 = 0$ với bước nhảy bằng 1.

Năm	t_i	y_i	$t_i y_i$	t_i^2
2005	0	7	0	0
2006	1	7,96	7,96	1
2007	2	9,19	18,38	4
2008	3	11,45	34,35	9
2009	4	11,6	46,4	16
2010	5	12,73	63,65	25
2011	6	15,17	91,02	36
2012	7	17,49	122,43	49
Tổng	28	92,59	384,19	140

a) Giả sử hàm xu thế tuyến tính có dạng $y = a + bt$. Ta có $n = 8$ và phương trình chuẩn tắc là

$$\begin{cases} 8a + 28b & = 92,59 \\ 28a + 149b & = 384,19. \end{cases}$$

Nghiệm của phương trình chuẩn tắc là $a = 6,5633, b = 1,4315$.
Vậy hàm xu thế tuyến tính cần tìm là $y = 6,5633 + 1,4315t$.
b) Năm 2016 tương ứng với giá trị $t_0 = 11$ và dự báo thu nhập GDP bình quân đầu người đạt
$y_0 = 6,5633 + 1,4315.11 = 22,31$ hay 2231 USD/người.

5.1 Hồi quy tuyến tính

Đổi biến trong hồi quy tuyến tính

Hồi quy hyperbol. Nếu hàm hồi quy có dạng hyperbolic $y = \beta_0 + \frac{\beta_1}{x}$, ta đổi biến $z = \frac{1}{x}$ ta tìm hàm hồi quy tuyến tính $y = \beta_0 + \beta_1 z$. Các hệ số β_0, β_1 chính là các hệ số cần tìm. Giá trị dự báo y_0 đối với x_0 được tính theo công thức $y_0 = \hat{\beta}_0 + \frac{\hat{\beta}_1}{x_0}$.

Hồi quy dạng mũ. Nếu hàm hồi quy có dạng hàm số mũ $y = \beta_0 e^{\beta_1 x}$. Đây là hàm phi tuyến, ta sử dụng phép biến đổi logarit đưa về dạng tuyến tính như sau:

$$\ln y = \ln \beta_0 + \beta_1 x.$$

Khi đó ta đặt $z = \ln y$ và ta tìm hàm hồi quy tuyến tính của z đối với x. Giả sử hệ hàm hồi quy tìm được là $z = a + bx$ thì $\ln \beta_0 = a$ hay $\beta_0 = \exp(a)$ và $\beta_1 = b$. Giá trị dự báo y_0 đối với x_0 được tính theo công thức $z_0 = a + bx_0$ và $y_0 = \exp(z_0)$.

5.1 Hồi quy tuyến tính

Đổi biến trong hồi quy tuyến tính

Hồi quy phân thức. Nếu hàm hồi quy có dạng phân thức dưới dạng $y = \dfrac{x}{\beta_0 + \beta_1 x}$ thì ta có thể sử dụng mô hình hồi quy tuyến tính bằng cách biến đổi $\dfrac{1}{y} = \beta_1 + \beta_0 \dfrac{1}{x}$. Ta tìm hàm hồi quy tuyến tính đối với dữ liệu $(1/x_i, 1/y_i), i = 1, 2, \ldots, n$. Giả sử hàm hồi quy tuyến tính tìm được có dạng $\dfrac{1}{y} = a + b\dfrac{1}{x}$ thì các hệ số trong hàm phân thức xác định bởi $\beta_0 = b, \beta_1 = a$ và giá trị dự báo tại x_0 là $y_0 = \dfrac{x_0}{b + ax_0}$.

Ví dụ 5.6

Áp suất (P) của khí gas tương ứng với thể tích (V) trong bình. Đo áp suất của khí gas đối với các thể tích khác nhau ta được bảng dữ liệu sau:

Thể tích (cm^3)	50	60	70	90	100
Áp suất (kg/cm^2)	64,7	51,3	40,5	25,9	7,8

Với mô hình khí thì mối liên hệ giữa áp suất và thể tích cho bởi công thức $PV^\alpha = \beta$, với α, β là các hằng số.

a) Xác định các hệ số α, β.

b) Với thể tích khí gas là $80cm^3$ thì áp suất tương ứng bằng bao nhiêu?

5.1 Hồi quy tuyến tính

Đổi biến trong hồi quy tuyến tính

Giải:

STT	V_i	P_i	$x_i = \ln V_i$	$y_i = \ln P_i$	$x_i y_i$	x_i^2
1	50	64,7	3,9120	4,1698	16,3122	15,3039
2	60	51,3	4,0943	3,9377	16,1223	16,7637
3	70	40,5	4,2485	3,7013	15,7250	18,0497
4	90	25,9	4,4998	3,2542	14,6435	20,2483
5	100	7,8	4,6052	2,0541	9,4596	21,2076
Tổng			21,3598	17,1171	72,2625	91,5732

a) Với mô hình $PV^\alpha = \beta$ ta sử dụng phép biến đổi logarit hai vế ta được

$$\ln P = \ln \beta - \alpha \ln V.$$

Do đó ta sử dụng mô hình hồi quy tuyến tính đối với $(\ln V_i, \ln P_i)$, $i = 1, 2, \ldots, n$.

5.1 Hồi quy tuyến tính

Đổi biến trong hồi quy tuyến tính

Ta có $n = 5$ và
$\bar{x} = 4,2720, \bar{y} = 3,4234, S_{xy} = -0,8613, S_{xx} = 0,3246$. Gọi mô hình hồi quy tuyến tính là $y = a + bx$ ta có

$$b = \frac{S_{xy}}{S_{xx}} = -2,6535, a = \bar{y} - b\bar{x} = 14,7590.$$

Vậy ta có $\ln P = 14,7590 - 2,6535 \ln V$ do đó

$$\alpha = -b = 2,6535, \beta = \exp(a) = 2568862,888.$$

b) Với giá trị thể tích $V_0 = 80(cm^3)$ thì giá trị áp suất tướng ứng P_0 xác định bởi $\ln P_0 = 14,7590 - 2,6535 \ln V_0 =$
$14,7590 - 2,6535.\ln(80) = 3,1314$ Vậy áp suất khí gas tương ứng với thể tích $80cm^3$ là

$$P_0 = e^{3,1314} = 22,9(kg/cm^2).$$

5.2 Hệ số tương quan

Hệ số tương quan lý thuyết

Hệ số tương quan giữa hai biến ngẫu nhiên (X, Y) là

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{V}[X]}\sqrt{\mathbb{V}[Y]}}.$$

Hệ số tương quan lý thuyết ρ nói chung là không biết do ta không biết hàm phân phối đồng thời của (X, Y).

Định nghĩa 5.2

Hệ số tương quan ngẫu nhiên

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

là ước lượng điểm của hệ số tương quan ρ.

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Hệ số tương quan mẫu thực nghiệm, ký hiệu r, được xác định theo công thức

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Ước lượng điểm R là một ước lượng vững của hệ số tương quan ρ. Tuy nhiên R là ước lượng chệch của ρ với $\mathbb{E}[R] \approx \rho - \frac{\rho(1-\rho^2)}{2n}$ và phương sai của R là $\mathbb{V}[R] \approx \frac{(1-\rho^2)^2}{n}$. Hệ số tương quan thực nghiệm r là một điểm ước lượng của hệ số tương quan lý thuyết ρ. Để tính hệ số tương quan mẫu thực nghiệm r ta cần tính các đại lượng S_{xy}, S_{xx}, S_{yy} theo công thức ở phần trước.

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Ví dụ 5.7

Kiểm tra mức độ ô nhiễm (khí N_2O) do ô-tô khi dừng phát thải, ta có số liệu:

x_i (giây)	2	2	4	6	10	10
y_i (mg/lít)	10	11	20	27	52	55

Tính hệ số tương quan mẫu thực nghiệm.

Giải: Từ bảng số liệu trên ta lập bảng tính

STT	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	2	10	20	4	100
2	2	11	22	4	121
3	4	20	80	16	400
4	6	27	162	36	729
5	10	52	520	100	2704
6	10	55	550	100	3025
Tổng	34	175	1354	260	7079

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Ta có kích thước mẫu $n = 6$. Từ đó ta tính được

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 1354 - \frac{34.175}{6} = 362,3333$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 260 - \frac{34^2}{6} = 67,3333,$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 7079 - \frac{175^2}{6} = 1974,8333.$$

Hệ số tương quan mẫu

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{362,3333}{\sqrt{67,3333}\sqrt{1974,8333}} = 0,9936.$$

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Trong trường hợp biến ngẫu nhiên hai chiều ta gọi hàm hồi quy của Y đối với X là hàm $\varphi(x) = \mathbb{E}[Y|X = x]$.

Với trường hợp hồi quy tuyến tính ta có

$$\varphi(x) = \mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

Tương tự như trường hợp hồi quy tuyến tính ta đã xét ở phần trước, mô hình hàm hồi quy tuyến tính của Y theo X khi X là biến ngẫu nhiên được xác định bởi

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X. \tag{5.7}$$

Với mẫu quan sát $(x_i, y_i), i = 1, 2, \ldots, n$ của biến ngẫu nhiên hai chiều (X, Y), các hệ số hồi quy được xác định bằng phương pháp bình phương tối thiểu tương tự như trường hợp hồi quy tuyến tính và xác định bởi công thức

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Trong trường hợp này ta có sai số $E = S_{yy}(1 - r^2)$ và hệ số xác định $R^2 = r^2$. Ta nhận thấy nếu hệ số tương quan mẫu r gần giá trị 1 hoặc -1 thì sai số E càng bé, nghĩa là hồi quy tuyến tính càng chính xác và hệ số xác định R^2 cũng tăng.
Việc kiểm định và ước lượng khoảng tin cậy các hệ số hồi quy được xác định như phần hồi quy tuyến tính đối với trường hợp x không phải biến ngẫu nhiên.

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Ví dụ 5.8

Kiểm tra mức độ ô nhiễm do các xe ô-tô khi dừng phát ra ta có bảng số liệu sau:

x_i (giây)	2	2	4	6	10	10
y_i (mg/lít)	10	11	20	27	52	55

1. Tìm hàm hồi quy tuyến tính của mức độ ô nhiễm (y) theo thời gian dừng (x).
2. Hãy ước lượng các tham số của hàm hồi quy tuyến tính với độ tin cậy $\gamma = 95\%$.
3. Ước lượng giá trị y ứng với thời gian dừng 8 (giây) và tìm khoảng tin cậy cho giá trị ước lượng trên với độ tin cậy 95%.

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Giải: Từ bảng số liệu trên ta lập bảng tính (Ví dụ 5.8)

STT	x_i	y_i	x_iy_i	x_i^2	y_i^2
1	2	10	20	4	100
2	2	11	22	4	121
3	4	20	80	16	400
4	6	27	162	36	729
5	10	52	520	100	2704
6	10	55	550	100	3025
Tổng	34	175	1354	260	7079

Ta có kích thước mẫu $n = 6$. Từ đó ta tính được $\bar{x} = 5,6667; \bar{y} = 29,1667; S_{xy} = 362,3333; S_{xx} = 67,3333; S_{yy} = 1974,8333$.

Giả sử hàm hồi quy tuyến tính của y theo x là $y = a + bx$. Ta có hệ số độ dốc là $b = \frac{S_{xy}}{S_{xx}} = 5,3812$ và hệ số tự do bằng $a = \bar{y} - b\bar{x} = -1,3267$.

Vậy hàm hồi quy tuyến tính là $y = -1,3267 + 5,3812x$.

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Khoảng tin cậy $\gamma = 0,95 = 1 - \alpha$ với $\alpha = 0,05$ của hệ số độ dốc β_1 được xác định bởi công thức

$$\left(\hat{\beta}_1 - t_{n-2,\alpha/2}\frac{s}{\sqrt{S_{xx}}}; \hat{\beta}_1 - t_{n-2,\alpha/2}\frac{s}{\sqrt{S_{xx}}}\right).$$

Thay số vào ta tìm được khoảng tin cậy là $(4,5344; 6,2279)$.
Khoảng tin cậy $\gamma = 1 - \alpha$ với $\alpha = 0,05$ đối hệ số tự do β_0 là

$$\left(\hat{\beta}_0 - t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}; \hat{\beta}_0 + t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right).$$

Thay số vào ta tìm được khoảng tin cậy là $(-6,9006; 4,24711)$.

5.2 Hệ số tương quan

Ước lượng điểm cho hệ số tương quan

Khoảng tin cậy $\gamma = 1 - \alpha$ với $\alpha = 0,05$ của giá trị ước lượng $y_0 = -1,3267 + 5,3812.8 = 41,7229$ tại $x_0 = 8$ là

$$\left(\hat{y}_0 - t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}; \hat{y}_0 + t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right).$$

Thay số vào ta tìm được khoảng tin cậy là $(38,2660; 45,1795)$.

5.2 Hệ số tương quan

Ước lượng và kiểm định hệ số tương quan

Khoảng tin cậy của hệ số tương quan

Để xây dựng khoảng tin cậy cho hệ số tương quan, ta sử dụng phép biến đổi ngược của hàm $y = \tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$ là hàm $y = \operatorname{atanh}(x) = \dfrac{1}{2}\ln\dfrac{1+x}{1-x}$. Với kích thước mẫu đủ lớn $n \geq 25$, thống kê

$$T = \operatorname{atanh}(R) = \frac{1}{2}\ln\frac{1+R}{1-R}$$

có phân phối xấp xỉ phân phối chuẩn với giá trị trung bình

$$\mu_T = \mathbb{E}[T] = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

và phương sai

$$\sigma_T^2 = \mathbb{V}[T] = \frac{1}{n-3}.$$

5.2 Hệ số tương quan

Ước lượng và kiểm định hệ số tương quan

Trong thực hành chúng ta thường xấp xỉ giá trị trung bình của thống kê T bởi $\mu_T \approx \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ và bỏ qua lượng sai số $\frac{\rho}{2(n-1)}$.
Khoảng tin cậy $\gamma = 1-\alpha$ đối với giá trị trung bình của T là

$$\left(\mu_T - z_{\alpha/2}\sigma_T, \mu_T + z_{\alpha/2}\sigma_T\right).$$

Khi đó hệ số tương quan ρ thỏa mãn

$$\frac{1}{2}\ln\frac{1+r}{1-r} - z_{\alpha/2}\frac{1}{\sqrt{n-3}} < \frac{1}{2}\ln\frac{1+\rho}{1-\rho} < \frac{1}{2}\ln\frac{1+r}{1-r} + z_{\alpha/2}\frac{1}{\sqrt{n-3}}.$$

Sử dụng phép biến đổi trên ta được khoảng tin cậy đối với giá trị hệ số tương quan ρ là

$$\tanh\left(\operatorname{atanh} r - z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right) < \rho < \tanh\left(\operatorname{atanh} r + z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right).$$

5.2 Hệ số tương quan

Ước lượng và kiểm định hệ số tương quan

Định lý 5.6

Khoảng tin cậy $\gamma = 1 - \alpha$ đối với hệ số tương quan xác định bởi

$$\tanh\left(\operatorname{atanh} r - z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right) < \rho < \tanh\left(\operatorname{atanh} r + z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right).$$

5.2 Hệ số tương quan

Ước lượng và kiểm định hệ số tương quan

Kiểm định hệ số tương quan với giả thuyết $H_0 : \rho = 0$

Đối với bài toán kiểm định giả thuyết $H_0 : \rho = 0$ và đối thuyết: $H_1 : \rho \neq 0$. Ta sử dụng thống kê $T = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ có phân phối Student với $n-2$ bậc tự do khi giả thuyết H_0 đúng. Bác bỏ giả thuyết H_0 với mức ý nghĩa α nếu

$$|t| > t_{n-2,\alpha/2}.$$

Kiểm định hệ số tương quan với giả thuyết $H_0 : \rho = \rho_0$.

Đối với việc kiểm định giả thuyết giả thuyết $H_0 : \rho = \rho_0$, đối thuyết: $H_1 : \rho \neq \rho_0$, ta sử dụng thống kê

$$T = \Big(\operatorname{atanh}(R) - \operatorname{atanh}(\rho_0)\Big)\sqrt{n-3}$$

có phân phối xấp xỉ phân phối chuẩn tắc khi giả thuyết H_0 đúng. Bác bỏ giả thuyết H_0 với mức ý nghĩa α nếu $|t| > z_{\alpha/2}$.

5.2 Hệ số tương quan

Ước lượng và kiểm định hệ số tương quan

Ví dụ 5.9

TT	x_i	y_i	TT	x_i	y_i	TT	x_i	y_i	TT	x_i	y_i
1	4,36	16,9	11	2,795	21,6	21	2,91	21,9	31	2,556	33,5
2	4,054	15,5	12	3,41	16,2	22	2,91	21,9	32	2,2	34,2
3	3,605	19,2	13	3,07	20,8	23	1,975	34,1	33	2,02	31,8
4	3,94	18,5	14	3,62	18,6	24	1,915	35,1	34	2,13	37,3
5	2,155	30	15	3,41	18,1	25	2,67	27,4	35	2,19	30,5
6	2,56	27,5	16	3,84	17	26	1,99	31,5	36	2,815	22
7	2,3	27,2	17	3,725	17,6	27	2,135	29,5	37	2,6	21,5
8	2,23	30,9	18	3,955	16,5	28	2,57	28,4	38	1,925	31,9
9	2,83	20,3	19	3,83	18,2	29	2,595	28,8			
10	3,14	17	20	2,585	26,5	30	2,7	26,8			

- a) Tính hệ số tương quan mẫu và hàm hồi quy tuyến tính Y theo X.
- b) Tìm khoảng tin cậy 95% của hệ số tương quan.

5.2 Hệ số tương quan

Ước lượng và kiểm định hệ số tương quan

Giải: Ta có kích thước mẫu $n = 38$, từ bảng trên ta tính được

$$\sum x_i = 108,22 \qquad \sum y_i = 942,2 \qquad \sum x_i y_i = 2530,317$$
$$\sum x_i^2 = 326,461652 \qquad \sum y_i^2 = 24938,52$$

Từ đó ta tính được các giá trị

$$\bar{x} = 2,8479, \bar{y} = 24,7947, S_{xy} = -152,9694,$$
$$S_{xx} = 18,2625, S_{yy} = 1576,9189.$$

a) Hệ số tương quan mẫu

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = -0,9014.$$

Giả sử hàm hồi quy tuyến tính là $y = \beta_0 + \beta_1 x$ với hệ số hồi quy là

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -8,3762$$
$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = 48,6491.$$

Do đó hàm hồi quy tuyến tính cần tìm là

$$y = 48,6491 - 8,3762x.$$

b) Với mức ý nghĩa $\alpha = 0,05$, ta có $z_{\alpha/2} = 1,96$, kích thước mẫu $n = 38$ và

$$z_{\alpha/2}\frac{1}{\sqrt{n-3}} = \frac{1,96}{\sqrt{35}} = 0,3313.$$

Hệ số tương quan mẫu $r = -0,9014$ nên

$$\operatorname{atanh} r = \frac{1}{2}\ln\frac{1+r}{1-r} = -1,4797.$$

Khoảng tin cậy của hệ số tương quan ρ là

$$\Big(\tanh\Big(-1,4797 - 0,3313\Big); \tanh\Big(-1,4797 + 0,3313\Big)\Big)$$
$$= (-0,9479; -0,8172).$$

5.3 Hồi quy phi tuyến

Tỷ số tương quan

Tỷ số tương quan giữa biến ngẫu nhiên Y đối với biến ngẫu nhiên X ký hiệu là $\eta^2 = \eta^2(Y|X)$ được xác định bởi công thức

$$\eta^2 = \frac{\mathbb{E}[\mathbb{E}[Y|X]]^2 - [\mathbb{E}(Y)]^2}{\mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2},$$

trong đó $\mathbb{E}[Y|X]$ là kỳ vọng có điều kiện của Y đối với X.
Tỷ số tương quan nằm trong đoạn $[0, 1]$, nếu η^2 càng gần 1 thì sự phụ thuộc hàm giữa Y và X càng mạnh. Nếu hệ số tương quan gần ± 1 thì ta có η^2 gần 1, nghĩa là ta có sự phụ thuộc tuyến tính. Tuy nhiên nếu hệ số tương quan gần 0 và η^2 gần 1 thì ta có sự phụ thuộc phi tuyến.

Tỷ số tương quan mẫu Giả sử ta có n quan sát $(x_i, y_i), i = 1, 2, \ldots, n$ về biến ngẫu nhiên (X, Y). Gọi n_i là số lượng mẫu có thành phần thứ nhất bằng x_i, ta ký hiệu

$$\overline{y}|_{x_i} = \frac{1}{n_i} \sum_{(x_i, y_j)} y_j, \quad i = 1, 2, \ldots, k.$$

$$\overline{y^2}|_x = \frac{1}{n} \sum_{i}^{k} n_i (\overline{y}|_{x_i})^2$$

5.3 Hồi quy phi tuyến

Tỷ số tương quan

Tỷ số tương quan mẫu được tính theo công thức sau:

$$\eta^2 = \frac{\overline{y^2|x} - \bar{y}^2}{\overline{y^2} - \bar{y}^2}.$$

Đường hồi quy thực nghiệm Đường nối các các điểm $(x_i, y|_{x_i}), i = 1, 2, \ldots, k$ gọi là đường hồi quy thực nghiệm.

Ví dụ 5.10

Cho mẫu quan sát

$(-2;4), (-2;3.5), (-1;1), (0;0,5), (0;1), (0;1,5), (1;1), (1;2), (2;3,5), (2;$

Tính tỷ số tương quan mẫu.

Giải: Ta có bảng phân phối tần số của x_i là

x_i	-2	-1	0	1	2
n_i	2	1	3	2	2

5.3 Hồi quy phi tuyến

Tỷ số tương quan

Ta tính tổng các giá trị $y|x_i$ theo các giá trị x_i như sau

$$y|_{-2} = \frac{4 + 3,5}{2} = 3,75$$

$$y|_{-1} = \frac{1}{1} = 1$$

$$y|_0 = \frac{0,5 + 1 + 1,5}{3} = 1$$

$$y|_1 = \frac{1 + 2}{2} = 1,5$$

$$y|_2 = \frac{3,5 + 4}{3} = 3,75$$

Vậy ta có

$$\bar{y^2}|_x = \frac{2.3,75^2 + 1.1^2 + 3.1^2 + 2.1,5^2 + 2.3,75^2}{10} = 6,475.$$

5.3 Hồi quy phi tuyến

Tỷ số tương quan

Ta có bảng phân phối tần số của y_i là

y_i	$0,5$	1	$1,5$	2	$3,5$	4
n_i	1	3	1	1	2	2

Từ bảng phân phối trên ta tính được $\bar{y} = 2,2;\ \overline{y^2} = 6,6$. Tỷ số tương quan là

$$\eta^2_{} = \frac{\overline{y^2}|x - \bar{y}^2}{\overline{y^2} - \bar{y}^2} = \frac{6,475 - 2,2^2}{6,6 - 2,2^2} = 0,93.$$

Để so sánh với hệ số tương quan, ta có

$$\sum x_i = 1, \sum y_i = 22, \sum x_i y_i = 2, \sum x_i^2 = 19, \sum y_i^2 = 66.$$

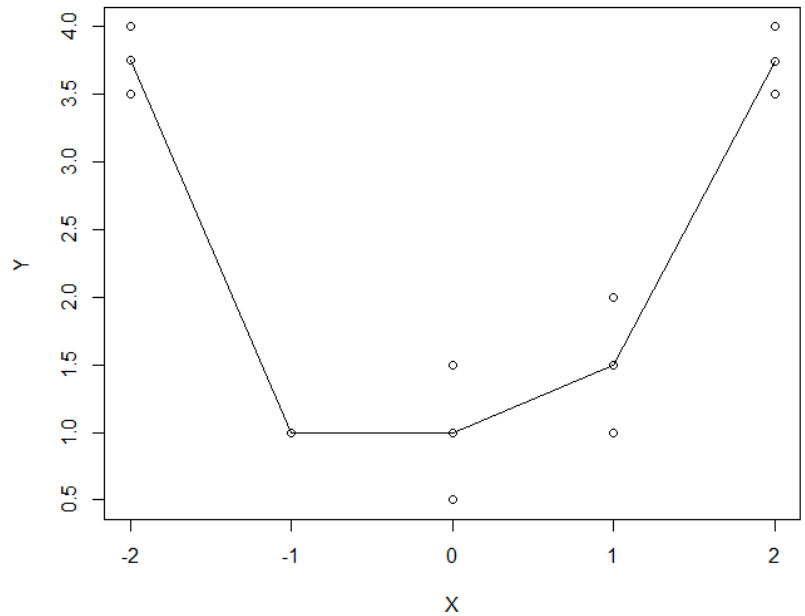
Vậy hệ số tương quan

$$r = -0,011.$$

Ta thấy hệ số tương quan gần bằng 0, vậy X và Y không có mối tương quan tuyến tính. Tuy nhiên ta có tỷ số tương quan gần bằng 1 nên X và Y có quan hệ phi tuyến.

5.3 Hồi quy phi tuyến

Tỷ số tương quan



Hình vẽ trên thể hiện các điểm dữ liệu và đường hồi quy thực nghiệm. Từ đồ thị của đường hồi quy thực nghiệm ta thấy không có mối quan hệ tuyến tính. Đường hồi quy phi tuyến này có dạng hàm bậc hai.

5.3 Hồi quy phi tuyến

Tỷ số tương quan

Khi X và Y có quan hệ phi tuyến thì ta cần tìm hàm phi tuyến dạng $y = \varphi(x)$ phù hợp với dữ liệu nhất. Vấn đề xác định hàm phi tuyến là một vấn đề khó, thường người ta sử dụng phương pháp trực quan vẽ các điểm (x_i, y_i) trên mặt phẳng hoặc vẽ đường hồi quy thực nghiệm rồi từ đó đưa ra các giả thuyết về dạng hàm phi tuyến hoặc dựa vào kinh nghiệm để đưa ra dạng của hàm phi tuyến. Trong phần tiếp theo đề cập một số dạng hàm phi tuyến thường sử dụng đối với hồi quy phi tuyến.

Nếu hàm hồi quy có dạng bậc 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

thì ta tìm các hệ số hồi quy để sai số là bé nhất. Ta sử dụng phương pháp bình phương tối thiểu để tổng bình phương các sai số

$$E = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right)^2$$

đạt giá trị nhỏ nhất.

5.3 Hồi quy phi tuyến

Hồi quy đa thức

Hồi quy bậc hai. Tương tự như phương pháp hồi quy tuyến tính, ta lấy đạo hàm của hàm E đối với các hệ số hồi quy.

$$\frac{\partial E}{\partial \beta_0} = 2\sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\Big)(-1) = 0$$

$$\frac{\partial E}{\partial \beta_1} = 2\sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\Big)(-x_i) = 0$$

$$\frac{\partial E}{\partial \beta_2} = 2\sum_{i=1}^{n}\Big(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\Big)(-x_i^2) = 0$$

5.3 Hồi quy phi tuyến

Hồi quy đa thức

Ta có hệ phương trình tuyến tính sau

$$\begin{cases} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 & = \sum_{i=1}^{n} x_i y_i \\ \beta_0 \sum_{i=1}^{n} x_i^2 + \beta_1 \sum_{i=1}^{n} x_i^3 + \beta_2 \sum_{i=1}^{n} x_i^4 & = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

Giải hệ phương trình trên ta được các hệ số hồi quy của hàm hồi quy bậc hai.

Ví dụ 5.11

Cho mẫu quan sát

$(-2;4), (-2;3,5), (-1;1), (0;0,5), (0;1), (0;1,5), (1;1), (1;2), (2;3,5), (2;$

Tìm hàm hồi quy bậc hai của y theo x.

5.3 Hồi quy phi tuyến

Hồi quy đa thức

Giải: Giả sử hàm hồi quy bậc hai có dạng $y = \beta_0 + \beta_1 x + \beta_2 x^2$.
Khi đó ta có hệ phương trình

$$\begin{cases} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 & = \sum_{i=1}^{n} x_i y_i \\ \beta_0 \sum_{i=1}^{n} x_i^2 + \beta_1 \sum_{i=1}^{n} x_i^3 + \beta_2 \sum_{i=1}^{n} x_i^4 & = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

Tính các hệ số từ số liệu được cho ta được

$$\begin{cases} \beta_0 10 + \beta_1 1 + \beta_2 19 & = 22 \\ \beta_0 1 + \beta_1 19 + \beta_2 1 & = 2 \\ \beta_0 19 + \beta_1 1 + \beta_2 67 & = 64 \end{cases}$$

Giải hệ trên ta được $\beta_0 = 0,8313, \beta_1 = 0,0237, \beta_2 = 0,7191$.
Vậy hàm hồi quy bậc hai cần tìm là

$$y = 0,8313 + 0,0237x + 0,7191x^2.$$

5.3 Hồi quy phi tuyến

Hồi quy đa thức

Hồi quy đa thức Giả sử rằng hàm hồi quy có dạng đa thức dạng

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

với n cặp giá trị quan sát $(x_i, y_i), i = 1, 2, \ldots, n$. Ta xác định các hệ số hồi quy bởi phương pháp bình phương tối thiểu, nghĩa là xác định các hệ số $\beta_j, j = 0, 1, \ldots, k$ để hàm

$$E = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k) \right]^2$$

đạt giá trị nhỏ nhất. Các hệ số hồi quy được tìm tương tự như trường hợp hồi quy bậc hai bằng cách cho các đạo hàm riêng của E đối với các hệ số hồi quy bằng 0. Khi đó ta có hệ phương trình tuyến tính gồm $k + 1$ ẩn của các hệ số hồi quy. Giải hệ phương trình này ta được hàm hồi quy đa thức thực nghiệm.
Biến phụ thuộc Y liên hệ với k biến độc lập (biến hồi quy) theo mô hình

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (5.8)$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

được gọi là mô hình hồi quy tuyến tính nhiều chiều (hay còn gọi là hồi quy đa biến) với k biến hồi quy $x_1, x_2, \ldots, x_k$ và $\epsilon \sim N(0, \sigma^2)$. Các tham số $\beta_j, j = 1, \ldots, k$ gọi là các hệ số hồi quy hay hệ số góc và hệ số β_0 gọi là hệ số tự do. Các tham số β_j biểu diễn sự thay đổi giá trị trung bình của biến phụ thuộc Y khi thay đổi biến hồi quy x_j khi các biến hồi quy $x_i, i \neq j$ không thay đổi. Chẳng hạn như trong một công ty thì mức lương của nhân viên có thể được xác định theo các yếu tố như trình độ, kinh nghiệm, giới tính, độ tuổi hoặc giá nhà có thể xác định theo các yếu tố như vị trí, diện tích, số phòng ngủ, số phòng tắm, v.v.. Tương tự như phần hồi quy tuyến tính đơn biến ta ký hiệu $\hat{\beta}_j$ là điểm ước lượng cho các hệ số β_j. Giá trị trung bình của biến phụ thuộc Y tại điểm $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ xác định bởi

$$y = \mathbb{E}[Y|\boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

được gọi là hàm hồi quy tuyến tính nhiều chiều.

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Phương pháp bình phương tối thiểu Phương pháp bình phương tối thiểu được sử dụng để tính các hệ số hồi quy dựa trên n mẫu quan sát $(x_{i1}, x_{i2}, \ldots, x_{ik}), i = 1, 2, \ldots, n$ và n quan sát về giá trị của biến phụ thuộc Y là $(y_1, y_2, \ldots, y_n)$ tương ứng với mẫu trên.

y	x_1	x_2	$\ldots$	x_k
y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

Table: Bảng dữ liệu hồi quy

Với mỗi quan sát $(x_{i1}, x_{i2}, \ldots, x_{ik})$ và y_i thỏa mãn mô hình hồi quy trong phương trình (5.8), nghĩa là ta có

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Ta tìm các hệ số hồi quy sao cho tổng bình phương các sai số e_i là nhỏ nhất có thể. Ta xét hàm

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2$$

Các hệ số hồi quy được tính bằng cách giải hệ sau:

$$\begin{cases} \frac{\partial E}{\partial \beta_0} & = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) = 0 \\ \frac{\partial E}{\partial \beta_j} & = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right) x_{ij} = 0, j = 1, 2, \ldots, k \end{cases}$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Viết tường minh hệ trên ta có hệ phương trình sau:

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_2 \sum_{i=1}^{n} x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik} & = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + \beta_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} x_{i1}x_{ik} & = \sum_{i=1}^{n} x_{i1}y_i \\ \vdots & \vdots \\ \beta_0 \sum_{i=1}^{n} x_{ik} + \beta_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \beta_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik}^2 & = \sum_{i=1}^{n} x_{ik}y_i \end{cases}$$

Giải hệ trên ta ước lượng được các hệ số hồi quy là $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Hệ phương trình trên được gọi là hệ phương trình chuẩn tắc cho hàm hồi quy tuyến tính nhiều chiều. Hàm hồi quy tuyến tính nhiều chiều thực nghiệm là

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k.$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Ví dụ 5.12

Giá của máy tính xách tay (y) (triệu đồng) được xác định bởi các yếu tố như hiệu suất hoạt động (x_1) và khả năng kết nối (x_2). Thống kê về 10 loại máy tính trên thị trường ta có bảng số liệu sau:

Loại	x_1	x_2	y
Dell XPS	10	9,48	35,7
HP Envy TS	9,38	9,45	27,7
HP Envy	9,38	9,23	26,3
Samsung ATIV	8,75	9,2	25,5
Sony Vaio S	8,13	8,9	17
Dell Ins	8,13	8,82	10,6
Samsung 5	7,5	8,75	18,5
Sony Vaio F	8,13	8,73	14,4
ThinkPad T	7,5	8,6	15,9
Samsung 3	7,5	8,58	12,6

Tìm hàm hồi quy tuyến tính dạng $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Giải: Giả sử hàm hồi quy tuyến tính cần tìm là $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Ta có phương trình chuẩn tắc xác định các hệ số hồi quy là

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{i1} + \beta_2 \sum_{i=1}^{n} x_{i2} & = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_{i1} + \beta_1 \sum_{i=1}^{n} x_{i1}^2 + \beta_2 \sum_{i=1}^{n} x_{i1}x_{i2} & = \sum_{i=1}^{n} x_{i1}y_i \\ \beta_0 \sum_{i=1}^{n} x_{i2} + \beta_1 \sum_{i=1}^{n} x_{i2}x_{i1} + \beta_2 \sum_{i=1}^{n} x_{i2}^2 & = \sum_{i=1}^{n} x_{i2}y_i \end{cases}$$

Từ bảng số liệu trên ta tính được các hệ số trong phương trình chuẩn tắc trên là

$$\begin{cases} 10\beta_0 + 84,4\beta_1 + 89,74\beta_2 & = 204,2 \\ 84,4\beta_0 + 719,572\beta_1 + 760,0319\beta_2 & = 1780,605 \\ 89,74\beta_0 + 760,0319\beta_1 + 806,36\beta_2 & = 1854,777 \end{cases}$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Giải hệ phương trình trên ta có nghiệm là các hệ số hồi quy

$$\hat{\beta}_0 = -160,073, \hat{\beta}_1 = 0,909, \hat{\beta}_2 = 19,258.$$

Vậy phương trình hồi quy tuyến tính là

$$y = -160,073 + 0,909x_1 + 19,258x_2.$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Tính chất của hệ số hồi quy Ta có thể xây dựng hồi quy tuyến tính nhiều chiều bằng cách sử dụng ký hiệu ma trận. Phương pháp này cho ta cách diễn đạt ngắn gọn hơn. Giả sử ta có bộ n giá trị quan sát của k biến độc lập và một biến phụ thuộc dạng

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), i = 1, 2, \ldots, n.$$

Gọi mô hình hồi quy tuyến tính là

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

với $\varepsilon \sim N(0, \sigma^2)$. Khi đó, các giá trị tại các điểm quan sát là

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, i = 1, 2, \ldots, n.$$

Trong đó các biến ngẫu nhiên $\varepsilon_i, i = 1, 2, \ldots, n$ độc lập, cùng phân phối chuẩn $N(0, \sigma^2)$.

Các phương trình này được viết dưới dạng ma trận như sau

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{5.9}$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

trong đó

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Khi đó ta cần tìm véc-tơ β sao cho

$$E = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

đạt giá trị nhỏ nhất. Giải phương trình của đạo hàm hàm sai số E bằng 0 ta được phương trình

$$\frac{\partial E}{\partial \beta} = \frac{1}{2}(\mathbf{X}^T \mathbf{X}\beta - \mathbf{X}^T \mathbf{y}) = 0,$$

như vậy ta được hệ phương trình tuyến tính

$$\mathbf{X}^T \mathbf{X}\beta = \mathbf{X}^T \mathbf{y}.$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Giải hệ trên và ta ký hiệu nghiệm là véc-tơ

$$B = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \tag{5.10}$$

Véc-tơ B là ước lượng điểm của véc-tơ β.
Theo cách biểu diễn bởi ma trận thì các hệ số hồi quy được xác định bởi công thức

$$B = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \varepsilon). \tag{5.11}$$

Ta có

$$\mathbb{E}[B] = \mathbb{E}\left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\beta + \varepsilon)\right] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\beta = \beta.$$

Do đó $B = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ là ước lượng không chệch của β. Ma trận hiệp phương sai của ước lượng B là $\mathrm{cov}(B) = \sigma^2 C$ với ma trận $C = (\mathbf{X}^T \mathbf{X})^{-1}$, nghĩa là $\mathbb{V}[\hat{\beta}_j] = \sigma^2 C_{jj}, \mathrm{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$.

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Định lý 5.7

Ước lượng điểm $B = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ là ước lượng không chệch của véc-tơ hệ số hồi quy β với ma trận hiệp phương sai là $\text{cov}(B) = \sigma^2 C$ trong đó $C = (\mathbf{X}^T\mathbf{X})^{-1}$.

Định lý 5.8

Xét mô hình hồi quy tuyến tính nhiều chiều $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, trong đó các ε_i độc lập, cùng phân phối chuẩn $N(0, \sigma^2)$. Khi đó, một điểm ước lượng từ ước lượng không chệch của σ^2 là sai số trung bình bình phương của các phần dư xác định bởi

$$s^2 = \frac{S_E}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}.$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Ta có phương sai tổng thể là

$$S_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

và ta gọi

$$S_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

là phương sai hồi quy. Vậy ta có

$$S_T = S_R + S_E.$$

Biểu thức phương sai tổng thể S_T có $n-1$ bậc tự do và số hạng S_R có chứa k bậc tự do nên S_E có $n-k-1$ bậc tự do, nên khi tính giá trị sai số trung bình ta chia cho số bậc tự do $n-k-1$. Tương tự như trường hợp hồi quy tuyến tính, hệ số xác định R^2 dùng để đo sự chính xác của mô hình hồi quy tuyến tính. Hệ số xác định R^2 được tính bởi công thức

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Định nghĩa 5.3

Hệ số xác định được tính bởi công thức

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T}.$$

Hệ số xác định trên chưa tính đến số biến được xác định trong mô hình hồi quy tuyến tính. Để tính đến số biến trong hàm hồi quy tuyến tính ta sử dụng hệ số xác định hiệu chỉnh R'^2 cho bởi công thức sau

Định nghĩa 5.4

Hệ số xác định hiệu chỉnh được cho bởi công thức

$$R'^2 = 1 - \frac{S_E/(n-k-1)}{S_T/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2). \qquad (1)$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Ví dụ 5.14

Trong Ví dụ 5.12 ta có được phương trình hồi quy tuyến tính là $y = -160,073 + 0,909x_1 + 19,258x_2$ khi đó ta tính được các giá trị từ mô hình lý thuyết ứng với mỗi mẫu quan sát và ta có

$$S_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 95,3361.$$

Ta tính phương sai của tổng thể là

$$S_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 576,496.$$

Khi đó hệ số xác định bởi công thức $R^2 = 1 - \dfrac{S_E}{S_T} = 0,8346.$

Nghĩa là có khoảng $83,46\%$ giá trị được giải thích bởi mô hình hồi quy tuyến tính trên. Hệ số xác định hiệu chỉnh là

$$R'^2 = 1 - \frac{9}{7}(1 - 0,8346) = 0,7874.$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Khoảng tin cậy cho giá trị ước lượng Đối với một giá trị quan sát mới $\boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0k})$ thì giá trị dự báo tương ứng đối với mô hình hồi quy tuyến tính là

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}.$$

Giá trị $\hat{y}_0$ là một điểm ước lượng của giá trị $y_0 = \mathbb{E}[Y|\boldsymbol{x}_0]$. Ước lượng điểm của giá trị này là $B_0 + B_1 x_{01} + B_2 x_{02} + \cdots + B_k x_{0k}$ với $B = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. Khoảng tin cậy cho giá trị quan sát y_0 có trong định lý sau.

Định lý 5.9

Khoảng tin cậy $\gamma = 1 - \alpha$ đối với giá trị quan sát mới y_0 của biến phụ thuộc Y tại $x_0 = (x_{01}, x_{02}, \ldots, x_{0k})$ được cho bởi công thức

$$\hat{y}_0 - t_{\nu,\alpha/2}\sqrt{s^2(1 + x_0'^T C x_0')} < y_0 < \hat{y}_0 + t_{\nu,\alpha/2}\sqrt{s^2(1 + x_0'^T C x_0')}$$

với $\nu = n - k - 1$, $C = (\mathbf{X}^T \mathbf{X})^{-1}$, và $x_0' = (1, x_{01}, x_{02}, \ldots, x_{0k})$.

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Ví dụ 5.14

Trong Ví dụ 5.12 ta có được phương trình hồi quy tuyến tính là

$$y = -160,073 + 0,909x_1 + 19,258x_2.$$

Giả sử một máy tính có hiệu suất hoạt động bằng $x_{01} = 9,5$ và khả năng kết nối là $x_{02} = 9,0$ thì giá tiền tương ứng là $y_0 = -160,073 + 0,909.9,5 + 19,258.9,0 = 21,885$ (triệu đồng).

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Khoảng tin cậy và kiểm định hệ số hồi quy. Trong hồi quy tuyến tính nhiều chiều việc xác định sự ảnh hưởng của một hoặc một vài biến nào đó rất quan trọng. Sự ảnh hưởng của một biến nào đó được đánh giá thông qua hệ số hồi quy tương ứng với biến đó. Ở phần trên ta biết hệ số ước lượng B_j có kỳ vọng là β_j và phương sai là $\sigma^2 C_{jj}$, và ta biết $s^2 = \dfrac{S_E}{n-k-1}$ là một điểm ước lượng không chệch của σ^2 và ta gọi ước lượng điểm tương ứng là S^2. Do đó ta sử dụng thống kê

$$T = \frac{B_j - \beta_j}{\sqrt{S^2 C_{jj}}}$$

có phân phối Student với $\nu = n - k - 1$ bậc tự do để kiểm định và tìm khoảng tin cậy của các hệ số hồi quy.

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Định lý 5.10

a) Khoảng tin cậy $\gamma = 1 - \alpha$ đối với hệ số hồi quy β_j là

$$\hat{\beta}_j - t_{\nu,\alpha/2}\sqrt{s^2 C_{jj}} < \beta_j < \hat{\beta}_j + t_{\nu,\alpha/2}\sqrt{s^2 C_{jj}}$$

b) Kiểm định giả thuyết $H_0 : \beta_j = b_j$ với đối thuyết $H_1 : \beta_j \neq b_j$ với mức ý nghĩa α, bác bỏ giả thuyết H_0 nếu $|t| > t_{\nu,\alpha/2}$ với

$$t = \frac{\hat{\beta}_j - b_j}{\sqrt{s^2 C_{jj}}}.$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Trong phân tích thống kê nhiều chiều người ta quan tâm đến hệ số dốc riêng $\beta_j, j = 1, 2, \ldots, k$ có thực sự ảnh hưởng đến mô hình hồi quy tuyến tính nhiều chiều hay không. Sử dụng phương pháp phân tích phương sai để kiểm định bài toán

$$H_0 : \beta_j = 0, \forall j = 1, 2, \ldots, k$$

$$H_1 : \text{Có ít nhất một hệ số } \beta_j \neq 0, j = 1, 2, \ldots, k.$$

Khi giả thuyết H_0 đúng, ta sử dụng thống kê

$$F = \frac{S_R/k}{S_E/(n-k-1)} = \frac{S_R/k}{S^2}$$

có phân phối Fisher với số bậc tự do k và $n-k-1$. Bác bỏ giả thuyết H_0 nếu

$$f = \frac{S_R/k}{s^2} > f_{k,n-k-1,\alpha}.$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Ví dụ 5.15 Trong Ví dụ 5.12 ta có ma trận

$$\mathbf{X} = \begin{pmatrix} 1 & 10 & 9,48 \\ 1 & 9,38 & 9,45 \\ 1 & 9,38 & 9,23 \\ 1 & 8,75 & 9,2 \\ 1 & 8,13 & 8,9 \\ 1 & 8,13 & 8,82 \\ 1 & 7,5 & 8,75 \\ 1 & 8,13 & 8,73 \\ 1 & 7,5 & 8,6 \\ 1 & 7,5 & 8,58 \end{pmatrix}$$

Tương ứng

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 10 & 84,4 & 89,74 \\ 84,4 & 719,572 & 760,0319 \\ 89,74 & 760,0319 & 806,36 \end{pmatrix}$$

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Tính nghịch đảo của ma trận trên ta được ma trận

$$C = (\mathbf{X}^T \mathbf{X})^{-1} = \begin{pmatrix} 446.505 & 25.641 & -73.859 \\ 25.641 & 1.784 & -4.535 \\ -73.859 & -4.535 & 12.496 \end{pmatrix}$$

Phương trình hồi quy tuyến tính là

$$y = -160,073 + 0,909x_1 + 19,258x_2.$$

Tìm khoảng tin cậy 95% đối với các hệ số β_1 và β_2.

5.4 Hồi quy nhiều chiều

Mô hình hồi quy tuyến tính nhiều chiều

Giải: a) Ta có $C_{11} = 1,784, s^2 = \dfrac{S_E}{n-k-1} = 13,619$. Tra bảng Student ta có $t_{\nu,\alpha/2} = t_{7;0,025} = 2,841$. Vậy khoảng tin cậy 95% đối với giá trị β_1 là

$$\left(\hat{\beta}_1 - t_{7;0,025}\sqrt{s^2 C_{11}}; \hat{\beta}_1 + t_{7;0,025}\sqrt{s^2 C_{11}}\right)$$

$$= \left(0,909 - 2,841\sqrt{13,619.1,784}; 0,909 + 2,841\sqrt{13,619.1,784}\right)$$

$$= \Big(-13,0947; 14,9126\Big).$$

b) Tương tự như trên ta có khoảng tin cậy đối với hệ số β_2 là

$$\left(\hat{\beta}_2 - t_{7;0,025}\sqrt{s^2 C_{22}}; \hat{\beta}_2 + t_{7;0,025}\sqrt{s^2 C_{22}}\right)$$

$$= \left(19,258 - 2,841\sqrt{13,619.12,496}; 19,258 + 2,841\sqrt{13,619.12,4}\right.$$

$$= \Big(-17,804; 56,32\Big).$$