

## 單元 49：最小平方迴歸分析 (課本 §7.7)

設有一些實際收集到的資料，描述  $y$  與  $x$  的關係，如

$$(2, 1), (5, 2), (7, 6), (9, 12), (11, 17)$$

問。如何根據這些資料，估計出  $y$  與  $x$  的真實關係，以致於可作推論之用？

答。根據資料建構出一個最配適的模型 (most fitted model)，如

(1) 線性模型 (linear model):

$$f(x) = 1.8566x - 5.024$$

(2) 二次式模型 (quadratic model):

$$g(x) = 0.1996x^2 - 0.7281x + 1.3749$$

如圖示，則可根據這些模型推論出  $y$  與  $x$  間的關係。

問 1. 如何建構？

問 2 何著較優？

答 2. 可採用如下定義的誤差平方和 (sum of the squared errors, SSE)，比較資料與模型間的差異，而決定出何者較優。

定義。設  $y = f(x)$  為資料點

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

的模型，則此模型的誤差平方和 (SSE)

$$\begin{aligned} S &\stackrel{\text{def}}{=} [f(x_1) - y_1]^2 + [f(x_2) - y_2]^2 \\ &\quad + \dots + [f(x_n) - y_n]^2 \end{aligned}$$

亦即，所有真實資料的  $y$  值與模型的  $y$  值間的差異的平方和。

接著，根據定義，線性模型的誤差平方和

$$\begin{aligned} S &= [f(2) - 1]^2 + [f(5) - 2]^2 + [f(7) - 6]^2 \\ &\quad + [f(9) - 12]^2 + [f(11) - 17]^2 \\ &\approx 16.9959 \end{aligned}$$

且二次模型的誤差平方和

$$\begin{aligned} S &= [g(2) - 1]^2 + [g(5) - 2]^2 + [g(7) - 6]^2 \\ &\quad + [g(9) - 12]^2 + [g(11) - 17]^2 \\ &\approx 1.8968 \end{aligned}$$

因為二次模型的誤差平方和較小，故二次模型較優。

答 1. 先以例說明。設資料為

$$(-3, 0), (-1, 1), (0, 2), (2, 3)$$

則建構一個最配適的線性模型乃相當於找一個線性函數

$$f(x) = ax + b$$

使得它的 SSE 最小，其中  $a$  與  $b$  為待定的未知常數，亦即，找  $a$  與  $b$  使得

$$\begin{aligned} S &= [f(-3) - 0]^2 + [f(-1) - 1]^2 \\ &\quad + [f(0) - 2]^2 + [f(2) - 3]^2 \end{aligned}$$

為最小。實際地計算  $f$  的函數值，並化簡，亦相當於使得

$$\begin{aligned} S &= (-3a + b - 0)^2 + (-a + b - 1)^2 \\ &\quad + (b - 2)^2 + (2a + b - 3)^2 \\ &= 9a^2 + b^2 - 6ab + a^2 + b^2 + 1 \\ &\quad - 2ab - 2b + 2a + b^2 + 4 - 4b \\ &\quad + 4a^2 + b^2 + 9 + 4ab - 12a - 6b \\ &= 14a^2 - 4ab + 4b^2 - 10a - 12b + 14 \end{aligned}$$

爲最小，此乃一雙變數函數的極值問題.

故，(1) 找臨界點，亦相當於解方程組

$$S_a = 28a - 4b - 10 = 0 \quad (1)$$

$$S_b = -4a + 8b - 12 = 0 \quad (2)$$

首先，將 (1) 式乘以 2，得

$$56a - 8b - 20 = 0$$

再將上式與 (2) 式相加，得

$$52a - 32 = 0$$

由此得

$$a = \frac{32}{52} = \frac{8}{13}$$

代入 (1) 式，得

$$b = \frac{1}{4} \left( 28 \cdot \frac{8}{13} - 10 \right) = \frac{112 - 65}{26} = \frac{47}{26}$$

因此，得第一類臨界點

$$\left( \frac{8}{13}, \frac{47}{26} \right)$$

無第二類臨界點，因爲一階偏導函數均有定義.

(2) 判斷. 因為只有一個第一類臨界點，故以二階偏導函數檢定法判斷. 首先，二階偏導函數為

$$S_{aa} = \frac{\partial}{\partial a}[28a - 4b - 10] = 28$$

且

$$S_{bb} = \frac{\partial}{\partial b}[-4a + 8b - 12] = 8$$

且

$$S_{ab} = \frac{\partial}{\partial b}[28a - 4b - 10] = -4$$

故，判別式

$$d(a, b) = (28)(8) - (-4)^2 > 0$$

為一常數，且恆大於 0，表示在臨界點有極值. 又

$$S_{aa}(a, b) = 28 > 0$$

亦為一常數，且恆大於 0，表示圖形上凹.

因此，在臨界點時，亦有上述判別式大於 0，且圖形上凹的現象，故當

$$a = \frac{8}{13}$$

且

$$b = \frac{47}{26}$$

時， $S$  有最小值，亦即，有最小的誤差平方和 (SSE).

所以，最配適的線性模型爲

$$f(x) = \frac{8}{13}x + \frac{47}{26}$$

同理，當有  $n$  個資料點

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

時，最配適的線性模型爲

$$f(x) = ax + b$$

其中

$$a = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

且

$$b = \frac{1}{n} \left( \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i \right)$$

並稱此直線爲最小平方迴歸線 (Least Squares Regression Line, LSRL).

**例 1.** 設由 1987 年至 1995 年的每小時平均工資  $y$ , 如下表:

年	1987	1988	1989	1990	1991
$y$	9.91	10.19	10.48	10.83	11.18
年	1992	1993	1994	1995	
$y$	11.46	11.74	12.06	12.35	

試求一最小平方迴歸線 (LSRL), 並估計 1998 年的平均工資.

<解> 令  $t = 0$  對應到 1990 年, 則資料點為

$$\begin{aligned} &(-3, 9.91), \quad (-2, 10.19), \quad (-1, 10.48) \\ &(0, 10.83), \quad (1, 11.18), \quad (2, 11.46) \\ &(3, 11.74), \quad (4, 12.06), \quad (5, 12.35) \end{aligned}$$

代入上述  $a$  與  $b$  的公式, 得 LSRL

$$y = 0.31t + 10.83$$

因此, 1998 年的工資, 乃相當於  $t = 8$ , 為

$$y(8) = 0.31(8) + 10.83 = 13.31$$

推廣. 最配適二次式模型

設有  $n$  個資料點

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

則最配適的二次式模型為

$$f(x) = ax^2 + bx + c$$

其中  $a, b, c$  滿足下列方程式

$$a \sum_{i=1}^n x_i^4 + b \sum_{i=1}^n x_i^3 + c \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i^2 y_i$$

且

$$a \sum_{i=1}^n x_i^3 + b \sum_{i=1}^n x_i^2 + c \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i$$

以及

$$a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i + cn = \sum_{i=1}^n y_i$$

並稱此拋物線為最小平方迴歸二次式 (Least Squares Regression Quadratic, LSRQ).