數學 2

進階講義

最小平方法與迴歸直線方程式

景美女中‧莊嘉銘 老師

信望愛文教基金會

7-2-3 最小平方法與迴歸直線方程式

定理證明或說明

1. 最小平方法

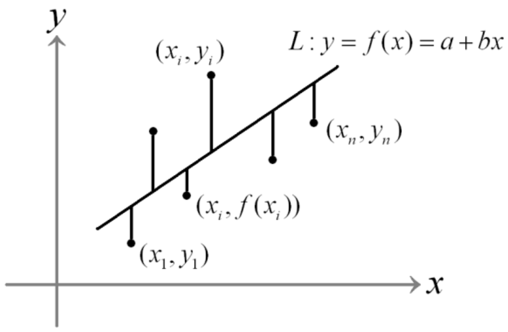
給定 n 個二維數據 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

欲求線性函數 $y = f(x) = a + bx$ 使得誤差平方和最小

即誤差 $E = (f(x_1) - y_1)^2 + (f(x_2) - y_2)^2 + \cdots + (f(x_n) - y_n)^2$ 為最小

2. 迴歸直線

將 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 視為坐標平面上的點，然後求一條直線 $y = a + bx$，使得下圖中 n 段鉛直線段長度平方和為最小，此直線即稱為迴歸直線



3. 迴歸直線的求法

設 $E = (a + bx_1 - y_1)^2 + (a + bx_2 - y_2)^2 + \cdots + (a + bx_n - y_n)^2$

$= na^2 + 2a[b(\sum x_i) - \sum y_i] + \sum (bx_i - y_i)^2$ ………(A)

$= b^2(\sum x_i^2) + 2b(a(\sum x_i) - \sum x_i y_i) + \sum (a - y_i)^2$ ………(B)

在(A)式中可視為 a 之一元二次式　　∴欲有最小值

故取 $a = \dfrac{-[b(\sum x_i) - \sum y_i]}{n}$　　$\therefore na + b(\sum x_i) = \sum y_i$ ……… ①

在(B)式中可視為 b 之一元二次式　　∴欲有最小值

故取 $b = \dfrac{-(a(\sum x_i) - \sum x_i y_i)}{(\sum x_i^2)}$　　$\therefore a(\sum x_i) + b(\sum x_i^2) = \sum x_i y_i$ ……… ②

即a，b需滿足方程組$\begin{cases} na+b(\sum x_i)=\sum y_i \cdots\cdots\cdots\cdots\cdots ① \\ a(\sum x_i)+b(\sum x_i^2)=\sum x_i y_i \cdots\cdots ② \end{cases}$

4. 補充說明

由①：$na+b(\sum x_i)=\sum y_i$　　$\therefore na+nb\overline{X}=n\overline{Y}$　　$\therefore \boxed{a=\overline{Y}-b\overline{X}}$

又由克拉瑪公式可得

$$b=\frac{\begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}}=\frac{n\sum x_i y_i-(\sum x_i)(\sum y_i)}{n\sum x_i^2-(\sum x_i)^2}=\frac{\sum x_i y_i-n\overline{X}\,\overline{Y}}{\sum x_i^2-n\overline{X}^2}$$

$$=\frac{\sum(x_i-\overline{X})(y_i-\overline{Y})}{\sum(x_i-\overline{X})^2}=r\cdot\frac{S_y}{S_x}$$（r為相關係數）

$$\therefore \boxed{b=r\cdot\frac{S_y}{S_x}=\frac{\sum(x_i-\overline{X})(y_i-\overline{Y})}{\sum(x_i-\overline{X})^2}=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}}$$

$\therefore$最小平方直線$y=a+bx$　$\Rightarrow y=(\overline{Y}-b\overline{X})+bx$

$\therefore \boxed{y-\overline{Y}=b(x-\overline{X})}$

即此直線必過($\overline{X}$，$\overline{Y}$)且斜率$=r\cdot\dfrac{S_y}{S_x}=\dfrac{\sum(x_i-\overline{X})(y_i-\overline{Y})}{\sum(x_i-\overline{X})^2}$

關鍵字

最適合直線、迴歸直線、預估值、標準差

例題 1

已知某班學生的數學成績(X)與地理成績(Y)的算術平均數分別為 $\mu_x = 65$，$\mu_y = 70$，且其相關係數為 $r = 0.8$。若 Y 對 X 的迴歸直線過點 $(5, 46)$，選出下列正確的選項為何？

(1) Y 對 X 的迴歸直線斜率為 0.8

(2) Y 對 X 的迴歸直線必過點 $(65, 70)$

(3) Y 的標準差大於 X 的標準差

(4) Y 的標準差小於 X 的標準差

(5) 若已知該班某位同學數學成績 70 分，則他的地理成績必為 72 分

Ans：

(1)(2)：Y 對 X 的迴歸直線為 $y = mx + k$

迴歸直線過點 $(5, 46)$，也會通過 $(\mu_x, \mu_y) = (65, 70)$

故 $\begin{cases} 46 = 5m + k \\ 70 = 65m + k \end{cases}$

$\Rightarrow m = \dfrac{2}{5}$、$k = 44$

得迴歸直線為 $y = \dfrac{2}{5}x + 44$，故斜率為 $\dfrac{2}{5} = 0.4$

(3)(4)：由於 $0.4 = m = r \cdot \dfrac{\sigma_y}{\sigma_x} \Rightarrow \dfrac{\sigma_y}{\sigma_x} = \dfrac{0.4}{0.8} = \dfrac{1}{2}$

$\Rightarrow \sigma_y = \dfrac{1}{2}\sigma_x < \sigma_x$

(5)相關性並無因果關係，故地理成績「必」為 72 分的說法不正確

故選(2)(4)

例題 2

已知變數 x 的算術平均數 $\mu_x=3$，標準差 $\sigma_x=1$，變數 y 的算術平均數 $\mu_y=4$，標準差 $\sigma_y=5$，變數 x 與變數 y 的相關係數 $r_{xy}=0.4$，若 $p=-2x+1$，$q=y-3$，則：

(1) 變數 p 與變數 q 的相關係數為何？

(2) q 對 p 的迴歸直線方程式為何？

Ans：

(1) 因為 $p=-2x+1$，$q=y-3$ $\quad\therefore r_{pq}=-r_{xy}=-0.4$

(2) 因為 $\mu_p=-2\mu_x+1=-5$，$\mu_q=\mu_y-3=1$

且 $\sigma_p=|-2|\sigma_x=2\sigma_x=2$，$\sigma_q=\sigma_y=5$

所以斜率 $m=r\cdot\dfrac{\sigma_q}{\sigma_p}=(-0.4)\times\dfrac{5}{2}=-1$

$\therefore q$ 對 p 的迴歸直線為 $y-1=(-1)(x+5)\Rightarrow y=-x-4$

例題 3

全班數學段考第一次成績的平均分數是 60 分，標準差為 12 分，第二次成績平均分數是 69 分，標準差為 10 分，兩次成績的相關係數是 0.6，小龍第一次成績是 66 分，請問小龍第二次成績預測是多少分？

Ans：

設第一次成績為 x 分，第二次成績為 y 分

則 Y 對 X 的迴歸直線方程式 $y-69=0.6\times\dfrac{10}{12}(x-60)$

$\Rightarrow y=0.5x+39$

所以當 $x=66$ 時，$y=0.5\times66+39=72$

例題 4

一組10個二維數據(x,y)，滿足$\sum_{i=1}^{10}x_i=20$，$\sum_{i=1}^{10}y_i=100$，$\sum_{i=1}^{10}x_i^{\ 2}=85$，$\sum_{i=1}^{10}y_i^{\ 2}=1500$，$\sum_{i=1}^{10}x_iy_i=326$，求：

(1) 這組數據的相關係數

(2) 這組數據Y對X的迴歸直線

(3) 利用Y對X的迴歸直線，預測$x=12$時，y的值

Ans：

(1) $$r=\frac{\sum_{i=1}^{10}x_iy_i-n\mu_x\mu_y}{\sqrt{\sum_{i=1}^{10}x_i^{\ 2}-n\mu_x^{\ 2}}\times\sqrt{\sum_{i=1}^{10}y_i^{\ 2}-n\mu_y^{\ 2}}}$$

$$=\frac{326-10\times2\times10}{\sqrt{85-10\times2^2}\times\sqrt{1500-10\times10^2}}=0.84$$

(2) 設$y=mx+k$

則$$m=\frac{\sum_{i=1}^{10}x_iy_i-n\mu_x\mu_y}{\sum_{i=1}^{10}(x_i-\mu_x)^2}=\frac{326-10\times2\times10}{85-10\times4}=\frac{14}{5}$$

將$(\mu_x,\mu_y)=(2,10)$代入$y=\frac{14}{5}x+k$，得$k=\frac{22}{5}$

故Y對X的迴歸直線為$y=\frac{14}{5}x+\frac{22}{5}$

(3) $x=12$代入$y=\frac{14}{5}x+\frac{22}{5}$，得$y=38$

例題 5

某雜誌想發行一本新書，在上市以前依不同的單價 X 元，調查市場的購買力為 Y 萬本，調查結果如下：

X	80	90	110	120
Y	11	12	8	9

試求 Y 對 X 的最適合直線為何？

Ans：

先求 $\mu_x = 100$ 及 $\mu_y = 10$

$x_i - \mu_x$	−20	−10	10	20
$y_i - \mu_y$	1	2	−2	−1

$$\therefore \sum_{i=1}^{4}(x_i - \mu_x)(y_i - \mu_y) = -20 - 20 - 20 - 20 = -80$$

$$\sum_{i=1}^{4}(x_i - \mu_x)^2 = 400 + 100 + 100 + 400 = 1000$$

$\therefore Y$ 對 X 的最適合直線為 $y - 10 = \dfrac{-80}{1000}(x - 100)$

$$\Rightarrow y = -0.08x + 18$$

溫故知新

習題 1

一組數據(x_i , y_i)，$i=1,2,\cdots,10$，μ_x，μ_y分別x_i為與y_i($i=1,2,\cdots,10$)的算術平均數，下列各敘述何者是不正確的？

(1) 相關係數r必小於1

(2) 相關係數不受單位的影響

(3) y對x的迴歸直線必過原點

(4) y對x的迴歸直線必過(μ_x , μ_y)

(5) 相關係數與迴歸直線之斜率同號

習題 2

已知變數x的算術平均數$\mu_x=3$，標準差$\sigma_x=\dfrac{3}{2}$，變數y的算術平均數$\mu_y=4$，標準差$\sigma_y=5$，變數x與變數y的相關係數$r_{xy}=0.6$，若$p=2x+3$，$q=-y-1$，則：

(1) 變數p與變數q的相關係數為何？

(2) q對p的迴歸直線方程式為何？

習題 3

身高y(公分)對體重x(公斤)的迴歸直線為$y=\dfrac{5}{4}x+100$，則體重為52公斤的人其身高的預測值為何？

習題 4

高一某次英文與國文競試後，全班50位同學的成績(x_i, y_i)，$i=1,2,\cdots,50$，(x_i，y_i分別表第i位同學的英文成績及國文成績)，整理得下面的數值：

$$\sum_{i=1}^{50} x_i = 3500，\sum_{i=1}^{50} y_i = 4000，\sum_{i=1}^{50} {x_i}^2 = 249900，\sum_{i=1}^{50} {y_i}^2 = 350625，$$

$\sum_{i=1}^{50} x_i y_i = 289800$，則：

(1) 這50位同學英文成績與國文成績的相關係數

(2) 依最小平方法，求國文成績(y)對英文成績(x)的最佳直線方程式

習題 5

給定5組(X, Y)數據如下：

X	2	1	4	5	3
Y	1	3	7	6	3

(1) 求Y對X的迴歸直線方程式

(2) 利用迴歸直線，預測$x=8$時，y值應為多少？

解答與解析

習題 1：(1)(3)

【詳解】(1) ╳：可能等於1

(3) ╳：不一定過原點

(5) ○：斜率$m = r \cdot \dfrac{\sigma_y}{\sigma_x}$　∴斜率與相關係數同號

答案為(1)(3)

習題 2：(1) -0.6 (2) $y=-x+4$

【詳解】(1) 因為 $p=2x+3$，$q=-y-1$ $\therefore r_{pq}=-r_{xy}=-0.6$

(2) 因為 $\mu_p=2\mu_x+3=9$，$\mu_q=-\mu_y-1=-5$

且 $\sigma_p=2\sigma_x=3$，$\sigma_q=|-1|\sigma_y=\sigma_y=5$

所以斜率 $m=r\cdot\dfrac{\sigma_q}{\sigma_p}=(-0.6)\times\dfrac{5}{3}=-1$

$\therefore q$ 對 p 的迴歸直線為 $y+5=(-1)(x-9)\Rightarrow y=-x+4$

習題 3：165 公分

【詳解】$x=52$ 代入得 $y=\dfrac{5}{4}\times52+100=65+100=165$

習題 4：(1) 0.8 (2) $y=2x-60$

【詳解】先求出 $\mu_x=70$ 及 $\mu_y=80$

(1) $$r=\frac{\sum_{i=1}^{50}x_iy_i-50\cdot\mu_x\cdot\mu_y}{\sqrt{(\sum_{i=1}^{50}x_i^2-50\mu_x^2)(\sum_{i=1}^{50}y_i^2-50\mu_y^2)}}$$

$$=\frac{289800-50\times70\times80}{\sqrt{249900-50\times(70)^2}\sqrt{350625-50\times(80)^2}}$$

$$=\frac{9800}{\sqrt{4900}\sqrt{30625}}=\frac{9800}{70\times175}=0.8$$

(2) $$m=\frac{\sum_{i=1}^{50}(x_i-\mu_x)(y_i-\mu_y)}{\sum_{i=1}^{50}(x_i-\mu_x)^2}=\frac{\sum_{i=1}^{50}x_iy_i-50\mu_x\mu_y}{\sum_{i=1}^{50}x_i^2-50\mu_x^2}=\frac{9800}{4900}=2$$

故國文成績(y)對英文成績(x)的最佳直線方程式為 $y-80=2(x-70)$

$\Rightarrow y=2x-60$

習題 5：(1) $y=\frac{6}{5}x+\frac{2}{5}$ (2)10

【詳解】(1) X 的平均數 $\mu_x=3$，Y 的平均數 $\mu_y=4$

$x-\mu_x$	$y-\mu_y$	$(x-\mu_x)^2$	$(x-\mu_x)(y-\mu_y)$
-1	-3	1	3
-2	-1	4	2
1	3	1	3
2	2	4	4
0	-1	0	0
總	和	10	12

因為 $m=\frac{\sum_{i=1}^{5}(x_i-\mu_x)(y_i-\mu_y)}{\sum_{i=1}^{5}(x_i-\mu_x)^2}=\frac{12}{10}=\frac{6}{5}$

故 Y 對 X 的迴歸直線為 $y-4=\frac{6}{5}(x-3)$，即 $y=\frac{6}{5}x+\frac{2}{5}$

(2) 將 $x=8$ 代入 $y=\frac{6}{5}x+\frac{2}{5}$，得 $y=10$