數學傳播　41卷3期, pp. 26-38

最小平方法與迴歸分析

蔡聰明

19 世紀的統計學, 主題是最小平方法以及迴歸分析 (Least Square Method and Regression Analysis), 即要找一條直線或曲線來適配 (fit) 一組觀測數據。根據統計學史家 Stigler, S. M. [1] 的說法, 最小平方法與微積分形成了類推的對比:

最小平方法、統計學	.....	微積分
觀測數據的演算		函數的演算
觀測數據的組合		數的組合

最小平方法在 1805 年由 Legendre (1752∼1833) 發其端, 接著高斯對測量數據發展出誤差論 (theory of errors), 由天文的觀測數據探尋星球運行的軌道。到了 19 世紀後半, 高爾頓 (Galton) 把它應用到生物的遺傳現象, 發展出迴歸分析, 大大地深化了統計學。

所謂迴歸問題就是探討: 以最小平方法求得迴歸係數, 迴歸直線, 以及相關係數, 使得我們對兩統計變量的關係有相當清晰的理解。

二次多項函數求極值有微分法與配方法。我們採用初等的配方法, 順便求出極值, 得到迴歸直線, 並且也自然地看出柯西−施瓦茲不等式與相關係數, 顯示配方法是豐收的。

高中數學教科書對於相關係數的定義與性質多半是語焉不詳, 本文可以補其不足, 適合於未學過微積分的高中生研讀。

一、迴歸問題

對於母群體(population) Ω, 我們同時觀測兩個統計變量:

$$X, Y \ : \ \Omega \to \mathbb{R}$$

例如某班同學的身高 X 與體重 Y, 數學成績 X 與物理成績 Y。觀測 Ω 的 n 個樣本點, 得到

如下的統計數據:

X	x_1	x_2	x_3	$\cdots$	x_n
Y	y_1	y_2	y_3	$\cdots$	y_n

我們要來處理這些統計數據, 對它們做分析, 找函數關係, 最後做出各種結論。

把統計數據 $\{(x_k, y_k) : k = 1, 2, \ldots, n\}$ 在坐標平面上描點出來, 見圖 1, 叫做散布圖 (scatter diagram), 這讓我們粗略看出兩個統計變量的變化大勢, 例如 x 變大時, y 也差不多變大。

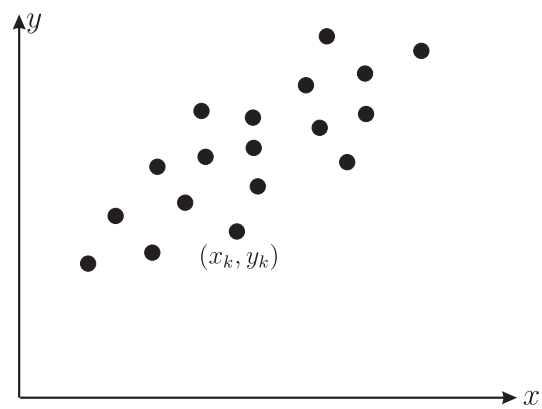


圖 1

要找 y 與 x 的函數關係, 本文我們只選用最簡單的一次函數 $y = ax + b$ (直線) 來對這一堆數據 $\{(x_k, y_k) \mid k = 1, 2, \ldots, n\}$ 作最佳適配 (best fit) 的工作。

「最佳」的意味就是讓「偏差」的總和爲最小。最自然的偏差是考慮爲點線距, 但這不好處理; 若改爲垂縱坐標差之和 $\sum_{k=1}^{n}[y_k-(ax_k+b)]$, 則會內力抵消, 也不好; 改爲 $\sum_{k=1}^{n}|y_k-(ax_k+b)|$ 仍然不好, 因爲絕對值不方便處理。最後改爲相差的平方和 $\sum_{k=1}^{n}[y_k-(ax_k+b)]^2$, 這在數學上最方便處理, 也不失其意義。

於是就有直線迴歸問題:

問題 1: (Y 對 X 的迴歸直線 $y = ax + b$) 求 a 與 b 使得縱軸方向的總平方差的平均

$$f(a, b) = \frac{1}{n}\sum_{k=1}^{n}[y_k - (ax_k + b)]^2 \tag{1}$$

取最小值。見圖 2。

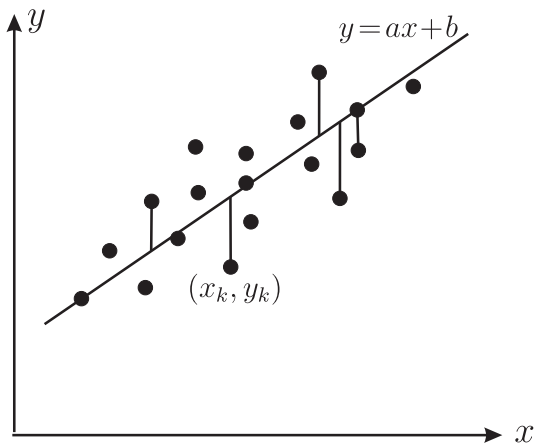


圖 2

問題2: (X 對 Y 的迴歸直線 $x = cy + d$) 求 c 與 d 使得橫軸方向的總平方差的平均

$$g(c,d) = \frac{1}{n}\sum_{k=1}^{n}[x_k - (cy_k + d)]^2 \tag{2}$$

取最小值。見圖 3。

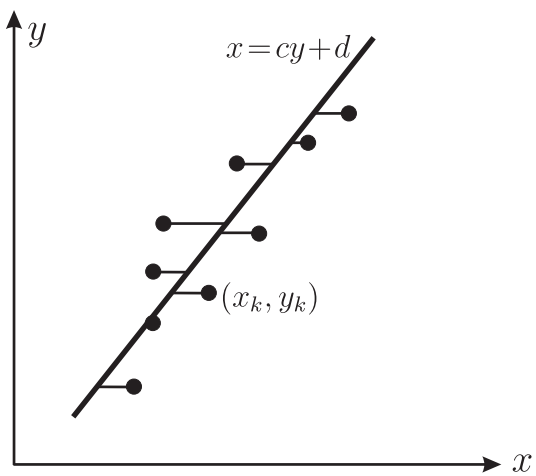


圖 3

注意: $f(a,b)$ 與 $g(c,d)$ 皆大於等於 0, 並且兩者都等於 0 $\Leftrightarrow$ 數據全落在一直線上。這個觀察對於以後要介紹的相關係數有重要意義。

二、配方法求極值

首先注意到, 配方法的優點是初等, 並且一舉求出極值點與極值, 而又能分辨出是最大值

或最小值。

爲了對 $f(a,b)$ 施展配方法求最小值, 我們一概採用統計的標準記號, 這很方便於計算與理解:

算術平均 (arithmetic mean): $\bar{x} = \dfrac{1}{n}\sum_{k=1}^{n} x_k$, $\bar{y} = \dfrac{1}{n}\sum_{k=1}^{n} y_k$

變異數 (variance): $\sigma_x^2 = \dfrac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$, $\sigma_y^2 = \dfrac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2$

標準差 (standard deviation): $\sigma_x = \sqrt{\sigma_x^2}$, $\sigma_y = \sqrt{\sigma_y^2}$ (開平方)

共變異數 (covariance): $\sigma_{xy} = \dfrac{1}{n}\sum_{k=1}^{n}[(x_k - \bar{x})(y_k - \bar{y})]$.

注意, 變異數與共變異數具有密切關係: $\sigma_{xx} = \sigma_x^2$ 且 $\sigma_{yy} = \sigma_y^2$。因此 σ_x^2 與 σ_y^2 是 σ_{xy} 的單元化, 而 σ_{xy} 是 σ_x^2 與 σ_y^2 的兩元化。

例 1: 用配方法求兩變數多項函數 $f(x,y) = x^2 - xy + y^2 - 2x + y - 3$ 的極值。

$$\begin{aligned} f(x,y) &= x^2 - xy + y^2 - 2x + y - 3 \\ &= x^2 - 2x\Big(\frac{y}{2}+1\Big) + \Big(\frac{y}{2}+1\Big)^2 - \Big(\frac{y}{2}+1\Big)^2 + y^2 + y - 3 \\ &= \Big(x - \frac{y}{2} - 1\Big)^2 + \frac{3}{4}y^2 - 4 \end{aligned}$$

因此 $y = 0$, $x = 1$ 是最小點, 而最小值爲 -4。 □

頭腦的體操: 假設 $a \neq 0$ 且 $b^2 - 4ac \neq 0$。對一般兩變數二次式

$$p(x,y) = ax^2 + bxy + cy^2 + dx + ey + f$$

作配方, 以求極值點與極值。

回到直線迴歸問題。因爲 $y = ax + b$ 是數據 $\{(x_k, y_k) \mid k = 1, 2, \ldots, n\}$ 的最佳適配直線, 所以應該會通過形心 $(\bar{x}, \bar{y})$, 因此我們將式子 $y_k - ax_k - b$ 改爲用記號 $\bar{x}$ 與 $\bar{y}$ 來表達

$$y_k - ax_k - b = (y_k - \bar{y}) - a(x_k - \bar{x}) + (\bar{y} - a\bar{x} - b)$$

接著將 $\dfrac{1}{n}\sum_{k=1}^{n}(y_k - ax_k - b)^2$作展開並且用變異數與共變異數來改寫

$$f(a,b) = \frac{1}{n}\sum_{k=1}^{n}(y_k - ax_k - b)^2 = \frac{1}{n}\sum_{k=1}^{n}[(y_k - \bar{y}) - a(x_k - \bar{x}) + (\bar{y} - a\bar{x} - b)]^2$$

$$= \frac{1}{n}\sum_{k=1}^{n}(y_k-\bar{y})^2 + a^2\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2 + (\bar{y}-a\bar{x}-b)^2 - 2a\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$
$$= \sigma_y^2 + a^2\sigma_x^2 + (\bar{y}-a\bar{x}-b)^2 - 2a\sigma_{xy} \tag{3}$$

注意到, 展開式中還有兩個交叉項, 但皆爲 0。將 $a^2\sigma_x^2 - 2a\sigma_{xy}$ 對 a 作配方

$$a^2\sigma_x^2 - 2a\sigma_{xy} = \sigma_x^2\Big[a^2 - 2a\frac{\sigma_{xy}}{\sigma_x^2} + \Big(\frac{\sigma_{xy}}{\sigma_x^2}\Big)^2\Big] - \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_x^2\Big(a^2 - \frac{\sigma_{xy}}{\sigma_x^2}\Big)^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}$$

代回到 (3) 式

$$f(a,b) = \sigma_x^2\Big(a-\frac{\sigma_{xy}}{\sigma_x^2}\Big)^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} + \sigma_y^2 + (\bar{y}-a\bar{x}-b)^2$$
$$= \sigma_x^2\Big(a-\frac{\sigma_{xy}}{\sigma_x^2}\Big)^2 + \sigma_y^2\Big(1-\frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\Big) + (\bar{y}-a\bar{x}-b)^2 \tag{4}$$

至此配方完成。我們可以合理地假設 $\sigma_x^2 > 0$ 且 $\sigma_y^2 > 0$, 因爲若 $\sigma_x^2 = 0$ 且 $\sigma_y^2 = 0$, 則所有數據崩塌爲一點 , 變成無聊。

我們觀察到, 第一項與第三項皆爲非負數, 第二項爲常數, 而第三項含有兩個未知數 a 與 b。我們先選取 $a = \dfrac{\sigma_{xy}}{\sigma_x^2}$, 讓 (4) 式的第一項變成 0; 以此 a 代入第三項, 使得第三項只含有一個未知數 b, 再取 $b = \bar{y} - a\bar{x}$, 讓第三項也變成 0。那麼由 (4) 式我們直接就可以讀出下面的諸多結果:

定理 1:

(i) 當 $a = \dfrac{\sigma_{xy}}{\sigma_x^2}$ 且 $b = \bar{y} - a\bar{x}$ 時, $f(a,b)$ 有最小值 $\min(f) = \sigma_y^2\Big(1 - \dfrac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\Big)$.

(ii) $0 \le \dfrac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} \le 1$.

(iii) $0 \le \min(f) \le \sigma_y^2$.

(iv) y 對 x 的迴歸直線 $y = ax + b$ 通過形心 $(\bar{x}, \bar{y})$, 其方程式爲

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x}). \tag{5}$$

同理, x 對 y 的迴歸直線 $x = cy + d$ 也通過形心 $(\bar{x}, \bar{y})$, 並且方程式爲

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y}). \tag{6}$$

頭腦的體操: 給統計數據:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

(i) 求 Y 對 X 的迴歸直線。(ii) 當 $x = 10$ 時, 估算 y 的值。

答: $y = \dfrac{65}{120}x + 1$, $y = \dfrac{65}{120} \times 10 + 1 = \dfrac{65}{12} + 1 \fallingdotseq 6.42$。

三、柯西–施瓦茲不等式

因爲對任何 a 與 b, $f(a,b) \geq 0$, 由 f 的連續性知 $f(a,b)$ 的最小值亦爲非負數, 於是就有下列一連串的等價式, 包括柯西–施瓦茲不等式與相關係數的性質:

$$\begin{aligned}
&\min(f) = \sigma_y^2\Big(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\Big) \geq 0 \\
\Leftrightarrow\ & 1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} \geq 0 \quad \text{或} \quad \sigma_{xy}^2 \leq \sigma_x^2\sigma_y^2 \quad \text{(柯西–施瓦茲不等式)} \\
\Leftrightarrow\ & \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} \leq 1 \\
\Leftrightarrow\ & \left|\frac{\sigma_{xy}}{\sigma_x\sigma_y}\right| \leq 1 \quad \text{(相關係數的性質)}
\end{aligned}$$

其次我們探討在上面不等式中等號成立的意涵。

$$\begin{aligned}
&\sigma_y^2\Big(1 - \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\Big) = 0 \\
\Leftrightarrow\ & f(a,b) = \frac{1}{n}\sum_{k=1}^{n}(y_k - ax_k - b)^2 \quad \text{的最小值爲} \quad 0 \\
\Leftrightarrow\ & \text{數據 } \{(x_k, y_k) : k = 1, 2, \ldots, n\} \text{ 完全落在迴歸直線 } y = ax + b \text{ 上} \\
\Leftrightarrow\ & \text{向量 } (x_1, x_2, \ldots, x_n) \text{ 與 } (y_1, y_2, \ldots, y_n) \text{ 爲線性相依。}
\end{aligned}$$

將 $\sigma_{xy}^2 \leq \sigma_x^2\sigma_y^2$ 代回原來的數據立得通常的柯西–施瓦茲不等式:

$$\left[\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})\right]^2 \leq \sum_{k=1}^{n}(x_k - \bar{x})^2 \sum_{k=1}^{n}(y_k - \bar{y})^2$$

定理2: (柯西–施瓦茲不等式)

假設 u_k 與 v_k, $k = 1, 2, \ldots, n$, 爲任意實數, 則有

$$\left[\sum_{k=1}^{n} u_k v_k\right]^2 \leq \sum_{k=1}^{n} u_k^2 \sum_{k=1}^{n} v_k^2 \tag{7}$$

並且等號成立的充要條件爲向量 $\vec{u} = (u_1, u_2, \ldots, u_k)$ 與 $\vec{v} = (v_1, v_2, \ldots, v_k)$ 線性相依, 亦即 $\vec{u}$ 與 $\vec{v}$ 有一個可表爲另一個的常數倍。

柯西–施瓦茲不等式有十多種證法, 此地我們透過最小平方法與兩變數的配方法又得到另一種簡潔的證法。

每個不等式的背後都有個等式, 反之亦然。同理, 每一個定理:"若 p 則 q" 的背後都有個逆敘述:"若 q 則 p", 但不一定成立。如果也成立的話, 就成爲"$p \Leftrightarrow q$", 即 p 與 q 等價或互爲充要條件。

在這裡值得特別注意, 有人說柯西–施瓦茲不等式只是內積的簡單推論。理由是, 按內積的定義 $\vec{u} \cdot \vec{v} = \|\vec{u}\| \cdot \|\vec{v}\| \cos\theta$, 其中 θ 爲兩向量的夾角。因爲 $|\cos\theta| \leq 1$, 所以

$$|\vec{u} \cdot \vec{v}| \leq \|\vec{u}\| \cdot \|\vec{v}\|$$

這就證得了柯西–施瓦茲不等式。事實上, 這只說對了一部分。

在兩維空間 $\mathbb{R}^2$ 與三維向量空間 $\mathbb{R}^3$ 的情形沒問題, 因爲此時有自然的幾何角度概念。但是, 當維數 $n \geq 4$ 時, 空間 $\mathbb{R}^n$ 沒有天然的角度概念, 論述就要小心。

假設 $\vec{u} = (u_1, u_2, \ldots, u_k)$ 與 $\vec{v} = (v_1, v_2, \ldots, v_k)$ 爲 $\mathbb{R}^n$ 中的兩個向量。首先, 定義內積爲

$$\vec{u} \cdot \vec{v} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

並且定義向量的長度爲

$$\|\vec{u}\| = \sqrt{x_1^2 + x_2^2 + \cdots x_n^2}$$

接著必須先證明柯西–施瓦茲不等式:

$$|\vec{u} \cdot \vec{v}| \leq \|\vec{u}\| \cdot \|\vec{v}\|$$

再來才定義兩向量 $\vec{u}$ 與 $\vec{v}$ 的夾角 θ 爲:

$$\cos\theta = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \cdot \|\vec{v}\|}$$

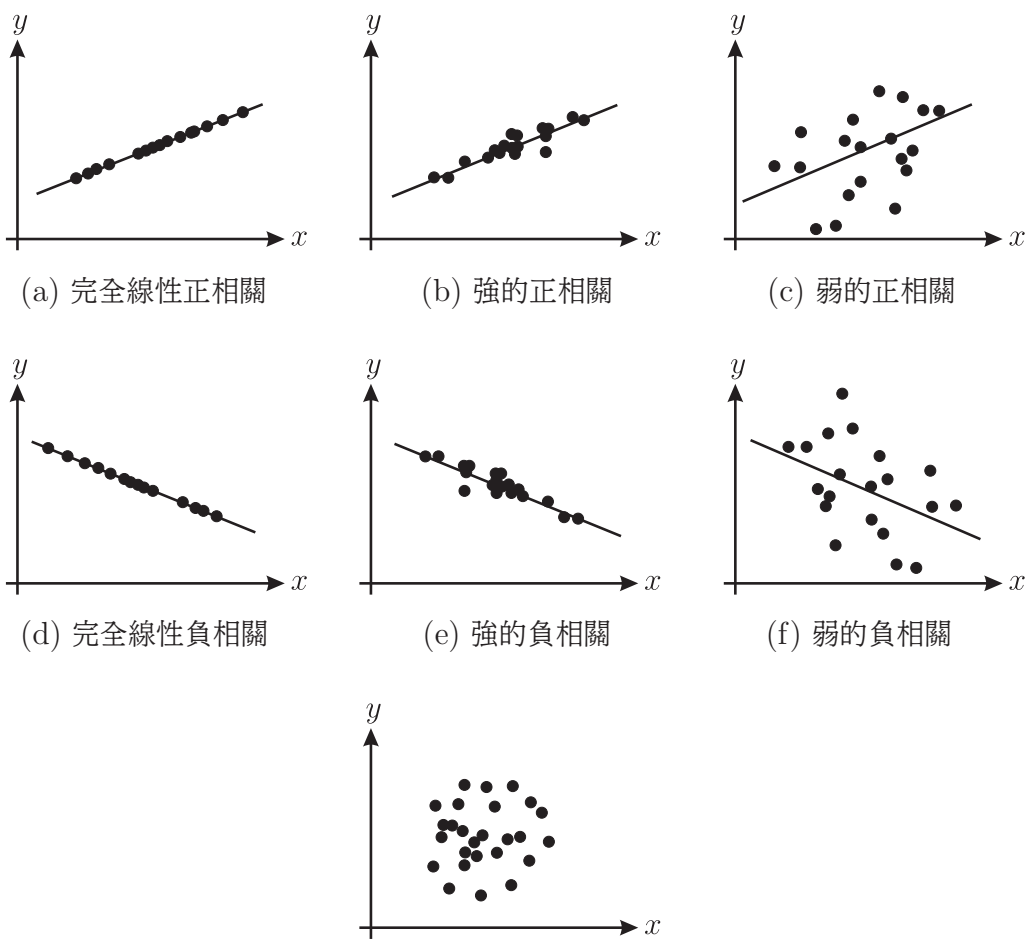
那麼由柯西–施瓦茲不等式知道, 這是適定的 (well-defined)。從而才有兩向量夾角的概念。由此才得到內積的另一種定義:

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \cdot \|\vec{v}\| \cos\theta.$$

對於 $n \geq 4$ 維的空間, 如果我們直接由內積定義與 $|\cos\theta| \leq 1$ 就下結論說柯西–施瓦茲不等式 $|\vec{u} \cdot \vec{v}| \leq \|\vec{u}\| \cdot \|\vec{v}\|$ 成立。這樣的論述是有問題的, 犯了循環論證的邏輯毛病。

四、相關係數

如何用一個數 (當作一根定量的尺度) 來衡量散佈圖上的數據成一直線的趨勢之強弱? 我們先觀察下列七種散佈圖, 直觀感受一下數據落在一直線上趨勢強弱的意思。



(a) 完全線性正相關　(b) 強的正相關　(c) 弱的正相關

(d) 完全線性負相關　(e) 強的負相關　(f) 弱的負相關

(g) 完全不相關

圖 4: (a)–(g)

上述散佈圖 (g) 像一盤散沙, 這是成直線趨勢最弱的情形, 叫做完全不相關。當迴歸直線的斜率爲正時叫做正相關 (x 越大 y 也越大), 如上圖 (a) 至 (c); 當迴歸直線的斜率爲負時叫做負相關 (x 越大 y 越小), 如上圖 (d) 至 (f)。

這一根定量的尺度遠在天邊, 近在咫尺! 定理 1 中的 $\dfrac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}$ 與其平方根 $\dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$ (不加絕對值), 恰好就是我們所要的尺度。令 $r=\dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$, 則 $r^2=\dfrac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}$。

定理 3: r 具有下列的性質:

(i) $|r|\le 1$ 或 $-1\le r\le +1$。

(ii) 當 $r^2=1$, 即 $r=\pm 1$時, 所有的數據 $\{(x_k,y_k):k=1,\dots,n\}$ 完全落在迴歸直線上 (兩條迴歸直線重合)。
當 $r=+1$ 時, 迴歸直線的斜率皆爲正: $a=\dfrac{\sigma_{xy}}{\sigma_x^2}>0$, $c=\dfrac{\sigma_{xy}}{\sigma_y^2}>0$。
當 $r=-1$ 時, 迴歸直線的斜率皆爲負: $a=\dfrac{\sigma_{xy}}{\sigma_x^2}<0$, $c=\dfrac{\sigma_{xy}}{\sigma_y^2}<0$。

(iii) 當 r 越在 0 附近時, 數據分散的程度越大, 數據成直線的趨勢越弱。

(iv) 當 r 越在 ± 1 附近時, 數據越集中在迴歸直線附近, 數據成直線的趨勢越強。

證明:

(i) 因爲 $f(a,b)$ 有最小值爲 $\sigma_y^2\Big(1-\dfrac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}\Big)=\sigma_y^2(1-r^2)$ 並且 $f(a,b)\ge 0$, 所以

$$\sigma_y^2(1-r^2)\ge 0$$

從而

$$r^2\le 1 \quad 或 \quad |r|\le 1.$$

(ii) 當 $r^2=1$ 時, $f(a,b)=\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(y_k-ax_k-b)^2$ 的最小值爲 0, 發生在 $a=\dfrac{\sigma_{xy}}{\sigma_x^2}$ 與 $b=\bar{y}-a\bar{x}$ 這一點上, 此時所有的數據 $\{(x_k,y_k):k=1,\dots,n\}$ 都落在迴歸直線 $y=ax+b$ 上。
當 $r=+1$ 時, 因爲 $\sigma_x>0$ 且 $\sigma_y>0$, 所以 $\sigma_{xy}>0$。從而迴歸直線的斜率 $a=\dfrac{\sigma_{xy}}{\sigma_x^2}>0$。當 $r=-1$ 時, $\sigma_{xy}<0$。從而迴歸直線的斜率 $a=\dfrac{\sigma_{xy}}{\sigma_x^2}<0$。

(iii) 當 r 在 0 附近時, 表示 $f(a,b)$ 的最小值 $\sigma_y^2(1-r^2)$ 越大, 因此數據 $\{(x_k,y_k):k=1,\dots,n\}$ 分散的程度越大。

(iv) 當 r 越在 ± 1 附近時, 表示 $f(a,b)$ 的最小值 $\sigma_y^2(1-r^2)$ 越小, 數據 $\{(x_k,y_k):k=$

$1, \ldots, n\}$ 越集中在迴歸直線上。

總之, r 這個數具有這麼多優良的性質, 簡直是天造地設, 恰好可以用來衡量數據 $\{(x_k, y_k) : k = 1, \ldots, n\}$ 在散布圖上成爲一直線趨勢的強弱指標。因此, 我們自然就要結晶爲如下的定義。

甲、相關係數的定義

定義: 我們稱 $r = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ 稱爲 X 與 Y 的 (線性) 相關係數。

將相關係數用原始數據來表示得到

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^{n} (y_k - \bar{y})^2}} \quad \text{(自動兼顧了正負號)} \tag{8}$$

這是在 1897 年英國統計學家皮爾森 (Karl Pearson, 1857∼1936) 引進的乘積級矩公式 (Product-moment formula)。皮爾森是高爾頓的學生, 是近代統計學的創立者與奠基者。

兩條迴歸直線 $y = ax + b$, $x = cy + d$ 的斜率 a 與 c 叫做迴歸係數, 我們已求得

$$a = \frac{\sigma_{xy}}{\sigma_x^2}, \qquad c = \frac{\sigma_{xy}}{\sigma_y^2}$$

因此兩個迴歸係數的幾何平均就是相關係數:

$$r = \sqrt{ac} = \sqrt{\frac{\sigma_{xy}}{\sigma_x^2} \cdot \frac{\sigma_{xy}}{\sigma_y^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

注意: 按照數學的規約, 開方應該要加絕對值符號 $\sqrt{\alpha^2} = |\alpha|$。但是此地的開方, 不加絕對值, 故 r 可取正負值。

考慮極端情形: 當數據完全落在一水平直線時, 即 $y_1 = y_2 = \cdots = y_n = \bar{y}$, 我們有 $\sigma_y = \sigma_{xy} = 0$, 於是 r 變成不定型 $0/0$, 所以沒有定義。同理, 當數據完全落在一垂縱直線時, 也有相同的結論。兩者都不是 $r = \pm 1$。對於這兩個極端的情形, 若把相關係數看成數據呈現直線趨勢的強弱, 則 x 與 y 爲完全相關。若從無法定義相關係數的眼光來看, 則 x 與 y 爲不相關。事情出現兩極化的說法, 所以這是規約問題。

頭腦的體操: 若兩條迴歸直線的斜率乘積爲 1, 即 $ac = 1$, 證明兩迴歸直線合而爲一, 並且所有數據都落在迴歸直線上。

註: 有數學家曾經這樣說: 統計只是相關係數, 而相關係數只是夾角的餘弦, 所以一切皆顯然! 這當然是偏頗的。

迴歸直線是統計數據的最佳適配直線 (best fitting line), 又可用來估計 (estimate) 未觀察的數據。特別是當 x 是指時間 t, 一個經濟系統的迴歸直線 $y = at + b$ 稱爲時間序列 (time series), 代表著趨勢線, 經常是我們根據過去來預測 (predict or forecast) 未來 y 值的工具。

上述所討論的相關係數其實是線性相關係數, 它是用來衡量數據落在一條直線的趨勢之強弱。通常我們省略掉「線性」這兩個字, 直接稱爲相關係數。請看下例:

例 2: 如果 y 與 x 有 $y = x^2$ 的關係, 並且在我們所取的數據中, 諸 x_k 是正負對稱成對, 那麼很容易算出 $r = 0$ (完全不相關或零相關), 因爲 $\bar{x} = 0$ 並且 $\sum x_k y_k = \sum x_k^3 = 0$。雖然 y 與 x 是函數地相關! 事實上, y 根本是 x 的平方! 這也警告我們, (線性) 相關係數並不是在談論兩個統計變量時, 唯一應該注意的東西。切記, 相關並不是因果關係 (correlation does not imply causation.)。 □

乙、相關係數的使用法

一般而言, 用相關係數 r 來衡量兩變數 x, y 的數據形成直線趨勢的強弱時, 我們採用如下的規約:

(i) 當 $r > 0$ 時, 我們稱兩變數爲正相關, 此時迴歸直線的斜率爲正數。

(ii) 當 $r < 0$ 時, 我們稱兩變數爲負相關, 此時迴歸直線的斜率爲負數。

(iii) 當 $r = 1$ 時, 我們稱兩變數爲完全正相關, 此時迴歸直線的斜率爲正數並且數據完全落在其上。
當 $r = -1$ 時, 我們稱兩變數爲完全負相關, 此時迴歸直線的斜率爲負數並且數據完全落在其上。

(iv) 當 $r = 0$ 時, 我們稱兩變數爲零相關或不相關。

(v) 當 $0.7 \le r < 1$ 時, 我們稱兩變數爲高度正相關。
當 $0.3 \le r < 0.7$ 時, 我們稱兩變數爲中度正相關。
當 $-1 < r \le -0.7$ 時, 我們稱兩變數爲高度負相關。
當 $-0.7 < r \le -0.3$ 時, 我們稱兩變數爲中度負相關。

(iv) 當 $0 < r < 0.3$ 時, 我們稱兩變數爲低度正相關, 數據落在迴歸直線上的趨勢薄弱。
當 $-0.3 < r < 0$ 時, 我們稱兩變數爲低度負相關, 數據落在迴歸直線上的趨勢薄弱。

註: 0.3 與 0.7 純是人爲取定, 所以才叫做規約。

例3: 20 對夫妻的年齡資料如下, 求兩條迴歸直線的方程式與相關係數。

夫 (歲)	22, 24, 26, 26, 27, 27, 28, 28, 29, 30, 30, 30, 31, 32, 33, 34, 35, 35, 36, 37
妻 (歲)	18, 20, 20, 24, 22, 24, 27, 24, 21, 25, 29, 32, 27, 27, 30, 27, 30, 31, 30, 32

解答: 先算出下列表格:

X	Y	XY	X^2	Y^2	X	Y	XY	X^2	Y^2
22	18	396	484	324	30	32	960	900	1024
24	20	480	576	400	31	27	837	961	729
26	20	520	676	400	32	27	864	1024	729
26	24	624	676	576	33	30	990	1089	900
27	22	594	729	484	34	27	918	1156	729
27	24	648	729	576	35	30	1050	1225	900
28	27	756	784	729	35	31	1085	1225	961
28	24	672	784	576	36	30	1080	1296	900
29	21	609	841	441	37	32	1184	1369	1024
30	25	750	900	625					
30	29	870	900	841	600	520	15887	18324	13868

所以 $\bar{x} = 30$, $\bar{y} = 26$,

$$\sigma_x^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - (\bar{x})^2 = \frac{1}{20} \times 18324 - (30)^2 = 16.2$$

$$\sigma_y^2 = \frac{1}{20} \times 13868 - (26)^2 = 17.4$$

$$\sigma_{xy} = \frac{1}{20}\sum_{k=1}^{n} x_k y_k - \bar{x}\bar{y} = \frac{1}{20} \times 15887 - 30 \times 26 = 14.35$$

y 對 x 的迴歸直線爲 $y - \bar{y} = \dfrac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$, 亦即 $y - 26 = 0.886(x - 30)$;

x 對 y 的迴歸直線爲 $x - \bar{x} = \dfrac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$, 亦即 $x - 30 = 0.825(x - 26)$

相關係數爲 $r = \sqrt{0.886 \times 0.825} = \sqrt{0.7310} \fallingdotseq 0.85$ (高度正相關)。 □

頭腦的體操: 某班 10 名學生的數學與物理成績如下:

X (數學成績)	75	80	93	65	87	71	98	68	84	77
Y (物理成積)	82	78	86	72	91	80	95	72	89	74

(i) 求兩條迴歸直線的方程式。

(ii) 求相關係數。

(iii) 當某生數學考 60 分時, 試預測其物理成績。

(iv) 當某生物理考 88 分時, 試預測其數學成績。

參考文獻

1. Stephen Stigler, The History of Statistics. The Measurement of Uncertainty before 1900. Harvard University Press, 1986.
2. Murray R. Spiegel, Statistics, Theory and Problems. McGraw-Hill, 1981.

—本文作者爲台大數學系退休教授—