二維數據分析

卓永鴻 提供

■1 簡介

某日在課堂上，數學老師勸大家少玩手機遊戲，玩得越多，就越影響課業表現。這時大家在底下議論紛紛。

甲：「玩遊戲歸玩遊戲，跟讀書考試有什麼關係？」

乙：「可是如果太沉迷於遊戲，多少還是會影響課業的吧？」

丙：「那個 XXX，還不是每天都玩，他成績就蠻好的啊！你很少玩還不是都考很爛！」

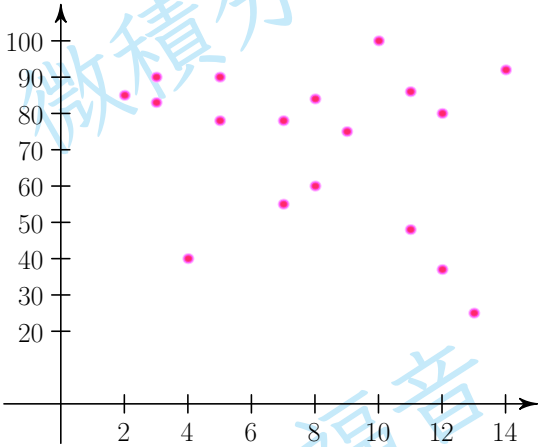
此時在台上的老師，作為一個二十一世紀有學識、講求科學證據的優秀老師，他決定給同學一個有說服力的說法，來讓同學相信玩太多遊戲真的會對課業有負面影響。

老師實際作了調查，訪問了 18 位同學，得到如下資料：

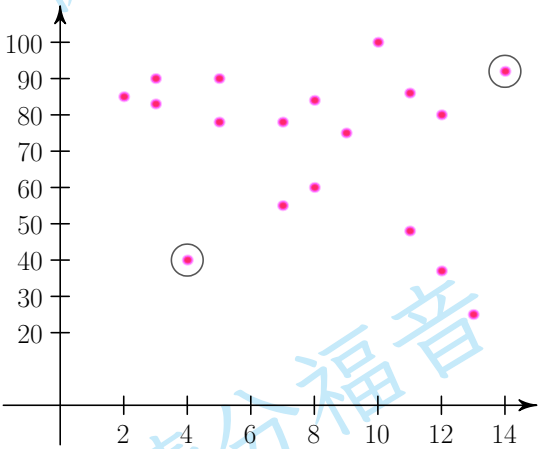
每週遊戲時間	8	2	10	9	12	5	11	5	14
數學成績	60	85	99	75	80	90	86	78	92
每週遊戲時間	12	3	7	7	11	13	3	4	8
數學成績	37	83	55	78	48	25	90	40	84

微積分福音

老師：「這樣看，可能還是不夠有感覺。接著我將這些資料畫出來吧！」
於是老師將每一筆數據都畫到**散 圖**上，得到下圖：



老師：「這樣一來，就稍微看得出下降趨勢了吧[①]！剛剛丙同學所犯的謬誤，是他只看兩種極端例子，但我們必須從整體資料去看，才可以去描述一般趨勢。」



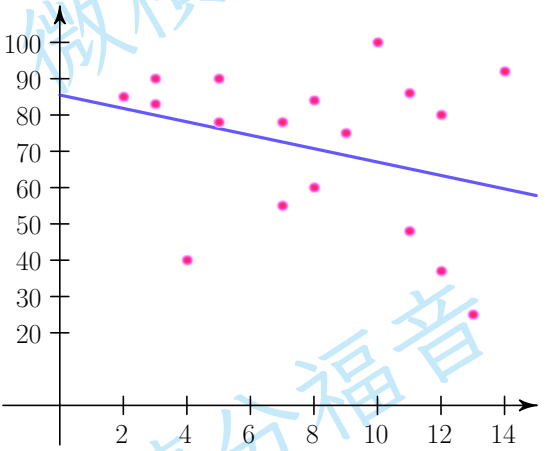
老師：「我們看上面這張圖，丙同學所說的，可能就是

[①] x, y 方向的單位長不一樣大，所以實際上的下滑趨勢會比這個圖大！

圈起來的這兩個資料。想想看，我們真的可以只看這兩個極端案例，卻無視整體趨勢嗎？」

丁：「可是老師，雖然你現在把資料都點在圖上面，比起看表格有感覺多了，但我還是覺得這上面好多點點，要我看趨勢實在有點吃力！」

老師：「好，那麼我便用一條直線，來代表全體趨勢。」

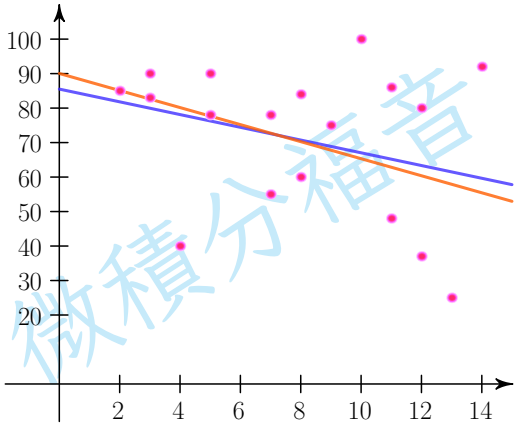


老師：「我現在這樣說，我可以用這條直線，來粗略地代表全體的趨勢。並不是說全體資料會在這條直線上，而是用它來粗略地描述全體情況。」

老師：「如果現在看這條直線，是不是更明顯可以感受到，手遊玩得越多，成績就越下滑的趨勢呢？」

戊：「可是老師，我覺得我看起來的趨勢不像這樣耶。」

說完，戊同學上台又畫了一條直線。



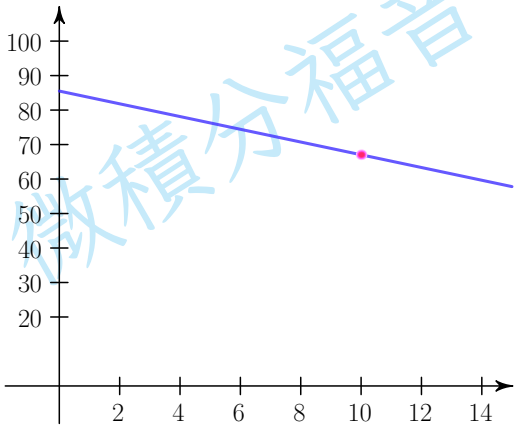
戊:「我覺得應該是這樣呀!這條線比較像全體的趨勢!」

老師:「這問題不錯!現在我們有了爭議,我們各自認為可代表全體的直線是不同條,所以我們必須有一個公認的辦法,來由原始數據求出這條代表性的直線。」

老師:「正好,我們接下來要上的主題,就是二維數據分析,我們就多介紹點這方面的概念吧!」

老師:「等我們學會如何求出這條代表全體的直線後,以後遇到任何新的統計資料,我們都可以套用這個固定模式來求出直線。」

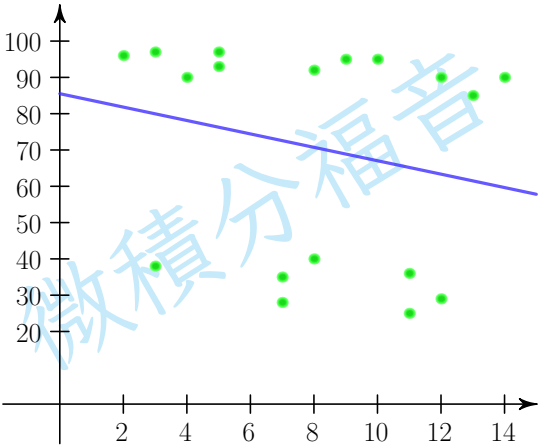
老師:「求出直線以後,現在如果再問一位同學,得知他每週玩手遊 10 個小時,我們想猜猜看他的成績,要怎麼猜呢?」



老師:「因為這條直線已經用來代表全體,所以我們自然就會猜在直線上。當然我們很可能會猜錯,但是雖不中,亦不遠矣,往直線上去猜,是最可能猜得準的了!」

老師:「所以我們要如何求出這條代表全體的直線,就是要求出一條會讓我們盡可能猜得準的直線!猜得準,我們才敢說這條直線可代表全體。因此,這條直線我們稱之為 **歸直線**,又叫**最適直線**,因為它是與這筆資料最適配的直線。」

老師:「不過也有一個問題,就是當求出迴歸直線以後,這條直線的代表性如何?」



老師：「我們看這筆新的資料，它與剛剛的資料有相同的迴歸直線。可是很明顯，這筆資料較為分散，所以如果我們要用來猜某同學的成績，就比較可能會猜得不準。這也就是說，這條迴歸直線的代表性較弱！」

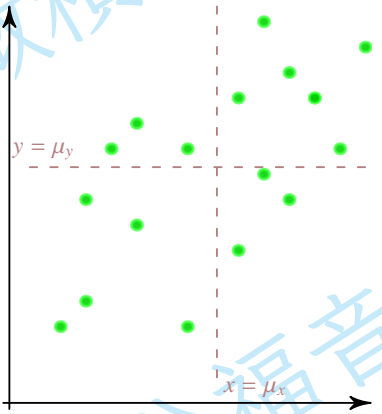
老師：「二維數據分析，就是圍繞在這樣的課題。當我們針對手遊時間與數學成績作調查，得到一筆資料，我們如何從中解讀手遊時間與數學成績之間的關係？其大致趨勢為何？這個趨勢的代表性是強是弱？」

老師：「這些概念不能仰賴直觀去看，否則每個人的感受都不太一樣。以下我們就來討論如何用數學式子來表達！」

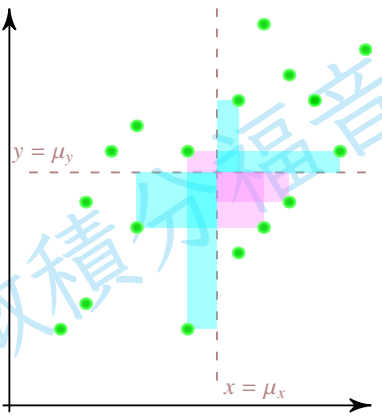
微積分福音

■ 2　相關係數

調查出一筆資料後，作出散佈圖，並畫出 $x=\mu_x$，$y=\mu_y$ 兩條直線。以這兩條直線作為新的 x 軸與 y 軸，將散佈圖分為第一到第四象限。



接著將每一個數據 (x_i,y_i)，把它的 x 座標 x_i，減掉 x 的平均 μ_x；同時 y 座標 y_i 也減掉 y 的平均 μ_y，然後兩者相乘，得到 $(x_i-\mu_x)(y_i-\mu_y)$。這樣子等於算出每個點拉到兩個新座標軸的矩形「面積」，而這個「面積」是有正負號的，因為如果是在第二象限的點，$(x_i-\mu_x)$ 就會是負的、第四象限的點，$(y_i-\mu_y)$ 是負的、第三象限的點，$(x_i-\mu_x)$，$(y_i-\mu_y)$ 都是負的，乘起來變正的。



我們把這每一個「面積」加起來，得到

$$\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y) \tag{1}$$

圖中顯然第一、三象限的點多，第二、四象限的點少，這表示大體來說，x 增加的話 y 也跟著增加，我們稱之為正相關。因為第一、三象限的點明顯較多，所以「面積」加起來後會是正的。反過來說，若大體而言，x 增加的話 y 會隨之減少，我們稱之為負相關。則圖中第一、三象限的點少，第二、四象限的點多，所以「面積」加起來後會是負的。

目前這個式子 (1)，可以幫助我們判斷正相關與負相關。但光是這樣還不夠好用，我們希望可以再進一步判斷相關程度的高低，就是說數據比較集中在迴歸直線附近還是較為分散，這就是**相關係數**的概念。相關係數是正的，即為正相關；相關係數是負的，則為負相關。相關係數的絕對值越大，就代表相關程度越高。但式子 (1) 顯然不能拿來當作相關係數的定義，因為如果我計算一個班級學生座號與成績的相關係數，而隔壁班正好也有一模一樣的成績分布，我現在兩班合併後再算一次相關係數，照理說相關程度應該不變，但拿上面那個式子來算會變兩倍！因此現在將式子略作修改，變成

$$\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{n} \tag{2}$$

將算「面積」總和，改為算「面積」平均，如此便能消除資料數 n 的干擾。但這還是不能拿來當作相關係數的定義，因為它會受到數據尺度的影響。舉例來說你調查氣溫 (°C) 與跑一百米秒數的關係，如果氣溫由攝氏改用華氏 (°F)，式子 (2) 算出來的值就會不一樣。可是資料根本還是同一組，只不過作了溫標轉換，相關係數算出來卻不一樣，這也太荒謬了！

在一維數據分析中，我們便已學過一種消除數據尺度的手法，就是**標準化**！對於資料 X，先將每一筆

數據 x_i 都減去 x 的平均，減完之後再除以 x 的標準差，如此操作則得到標準化後的數據 Z：

$$z_i = \frac{x_i - \mu_x}{\sigma_x}$$

標準化後的數據，其平均為 0、標準差為 1，如此則消弭了數據尺度這一因素。所以我們現在將資料 X 與 Y 都先進行標準化，得到 $\frac{x_i-\mu_x}{\sigma_x}$ 與 $\frac{y_i-\mu_y}{\sigma_y}$，然後再代回式子 (2)

$$\frac{\sum_{i=1}^{n}\left(\frac{x_i-\mu_x}{\sigma_x}-0\right)\left(\frac{y_i-\mu_y}{\sigma_y}-0\right)}{n}$$

標準化後的數據的平均為 0，所以減去平均那裡都變成減 0。再繼續化簡：

$$\frac{\sum_{i=1}^{n}\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right)}{n} = \sum_{i=1}^{n}\frac{(x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\sigma_y}{n}$$

$$= \frac{\frac{1}{\sigma_x\sigma_y}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{n} = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{n\sigma_x\sigma_y}$$

這就是相關係數的定義了！還可以再套用標準差的定義，寫成

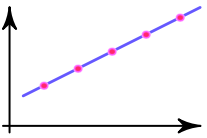
$$\frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\not{n}\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu_x)^2}{\not{n}}}\sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu_y)^2}{\not{n}}}}$$

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

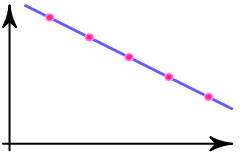
性質 1　相關係數的性質

1. 相關係數 r 滿足 $-1 \le r \le 1$

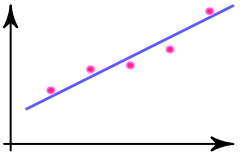
2. 當 $r = 1$，所有資料都在斜率為正的迴歸直線上，稱為**完全正相關**



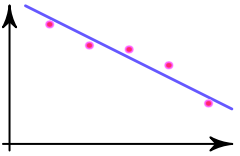
3. 當 $r = -1$，所有資料都在斜率為負的迴歸直線上，稱為**完全負相關**



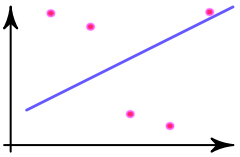
4. 當 r 很接近 1，資料集中在斜率為正的迴歸直線附近，稱為**高度正相關**



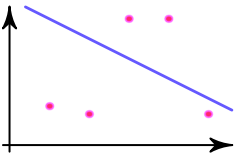
5. 當 r 很接近 -1，資料集中在斜率為負的迴歸直線附近，稱為**高度負相關**



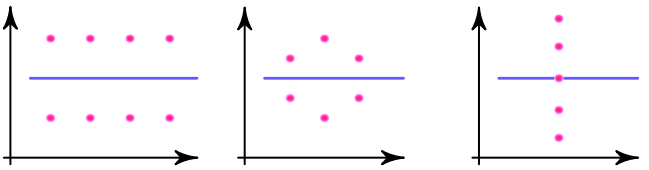
6. 當 r 正但很接近 0，迴歸直線斜率為正，但資料很分散，稱為**低度正相關**



7. 當 r 負但很接近 0，迴歸直線斜率為負，但資料很分散，稱為**低度負相關**



8. 當 $r = 0$，迴歸直線斜率為 0，此時稱為**零相關**



9. 若對於資料 X, Y 皆進行線性變換：$X' = aX + b$, $Y' = cY + d$，則新資料的相關係數 $r(X', Y')$ 與原資料的相關係數 $r(X, Y)$ 之間的關係為

$$r(X', Y') = \begin{cases} r(X, Y) & , ac > 0 \\ -r(X, Y) & , ac < 0 \end{cases}$$

線性變換以後相關係數的絕對值不變，只有在伸縮的 a, c 異號時才會多出負號。

注意

(1) 關於相關係數 r 為什麼會介於 -1 到 1 之間，一個看法是利用柯西不等式，寫

$$\begin{aligned} &\left((x_1 - \mu_x)^2 + \cdots + (x_n - \mu_x)^2\right) \\ &\quad \times \left((y_1 - \mu_y)^2 + \cdots + (y_n - \mu_y)^2\right) \\ &\geq \left((x_1 - \mu_x)(y_1 - \mu_y) + \cdots + (x_n - \mu_x)(y_n - \mu_y)\right)^2 \end{aligned}$$

即

$$\begin{aligned} &\left(\sum_{i=1}^{n} (x_i - \mu_x)^2\right)\left(\sum_{i=1}^{n} (y_i - \mu_y)^2\right) \\ &\geq \left(\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)\right)^2 \end{aligned}$$

移項後

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

微積分福音

接著再開根號即得

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

另一個看法比較抽象一點，是看成有兩個 n 維的向量

$$\vec{a} = (x_1-\mu_x, x_2-\mu_x, \cdots, x_n-\mu_x)$$
$$\vec{b} = (y_1-\mu_y, y_2-\mu_y, \cdots, y_n-\mu_y)$$

設 $\vec{a}$, $\vec{b}$ 夾角為 θ，利用向量的夾角公式

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\left|\vec{a}\right|\left|\vec{b}\right|}$$

就可以寫出

$$\cos\theta = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i-\mu_y)^2}}$$

因此相關係數可以看成是 $\vec{a}$ 與 $\vec{b}$ 的夾角取餘弦值，所以就會介於 -1 到 1 之間。$\cos\theta = 1$ 表示兩向量同向，為完全正相關；$\cos\theta = -1$ 表示兩向量反向，為完全負相關；$\cos\theta$ 越接近 0，就是兩向量越接近垂直，相關程度越低。

(2) 不可將零相關理解成兩資料完全沒有關聯，正確來說是「沒有線性相關」，即無法找到斜率不為零的直線來表示它們的趨勢，但有可能可以找到曲線來代表它們的趨勢。例如上面零相關的圖中有一個是數據都在一個圓上，這樣怎能說是 X, Y 完全無關呢？另外一例是設 $X = -1, 0, 1$，$Y = X^2$，即有三

筆資料 $(-1,1),(0,0),(1,1)$，則 $\sum_{i=1}^{3}(x_i-\mu_x)(y_i-\mu_y)=(-1)\cdot\frac{1}{3}+0+1\cdot\frac{1}{3}=0$，故 $r=0$。然而從 $Y=X^2$，已說明兩資料是有曲線關係的。

(3) 關於線性變換後的相關係數，只要直接將 $X'=aX+b$, $Y'=cY+d$ 代入相關係數的公式，就會有

$$\begin{aligned}
&r(X',Y')\\
&=\frac{\sum_{i=1}^{n}(x_i'-\mu_{x'})(y_i'-\mu_{y'})}{\sqrt{\sum_{i=1}^{n}(x_i'-\mu_{x'})^2}\sqrt{\sum_{i=1}^{n}(y_i'-\mu_{y'})^2}}\\
&=\frac{\sum_{i=1}^{n}\big[(ax_i+b)-(a\mu_x+b)\big]\big[(cy_i+d)-(c\mu_x+d)\big]}{\sqrt{\sum_{i=1}^{n}\big[(ax_i+b)-(a\mu_x+b)\big]^2}\sqrt{\sum_{i=1}^{n}\big[(cy_i+d)-(c\mu_x+d)\big]^2}}\\
&=\frac{\sum_{i=1}^{n}\big[a(x_i-\mu_x)\big]\big[c(y_i-\mu_y)\big]}{\sqrt{a^2\sum_{i=1}^{n}(x_i-\mu_x)^2}\sqrt{c^2\sum_{i=1}^{n}(y_i-\mu_y)^2}}\\
&=\frac{ac}{|ac|}r(X,Y)
\end{aligned}$$

若 a,c 同號則 $\frac{ac}{|ac|}=1$；若 a,c 異號則 $\frac{ac}{|ac|}=-1$。

也可以簡單用想的，相關係數是先將兩資料標準化後再代回式子 (2)。如果伸縮係數 a 是正的，則 $X'=aX+b$ 的標準化數據與 X 的標準化數據是一樣的；如果伸縮係數 a 是負的，則 $X'=aX+b$ 的標準化數據與 X 的標準化數據差負號。所以若 a,c 皆正，則 X' 與 Y' 的標準化數據都與 X 和 Y 的標準化數據一樣，那麼再代回式子 (2) 當然不會變；若 a,c 一正一負，標準化後其中一個差負號，再代回式子 (2) 就多個負號；若 a,c 皆負，標準化後兩個都差負號，再代回式子 (2) 就不變號。

例題 1 令 X 代表每個高中生平均每天研讀數學的時間（以小時計），則 $W=7(24-X)$ 代表每個高中生平均每週花在研讀數學以外的時間。令 Y 代表每個高中生數學學科能力測驗的成績。設 X, Y 之相關係數 R_{XY}，W, Y 之相關係數為 R_{WY}，則 R_{XY} 與 R_{WY} 兩數之間的關係，下列選項何者為真？

(A) $R_{WY}=7(24-R_{XY})$　(B) $R_{WY}=7RXY$

(C) $R_{WY}=-7R_{XY}$　(D) $R_{WY}=R_{XY}$

(E) $R_{WY}=-R_{XY}$

90 學測

解

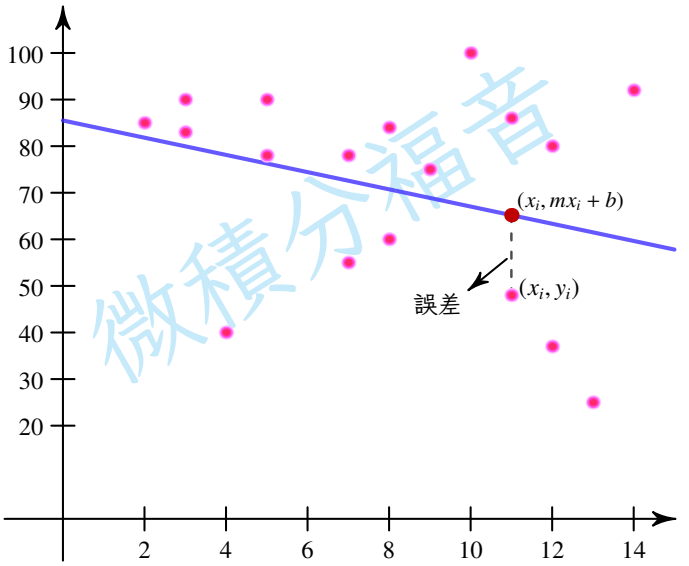
W 為將 X 作線性變換，其伸縮係數 -7 是負的，故 $R_{WY}=-R_{XY}$。

微積分福音

■ 3　迴歸直線

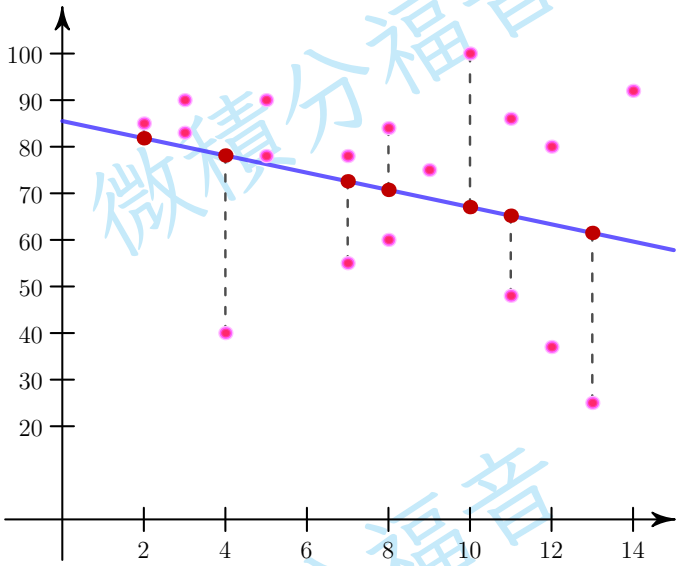
Y 對 X 的迴歸直線，是指在散佈圖中以直線粗略代表整體趨勢，並且當我們由 X 數據去猜測 Y 數據時，是最可能猜得準的。所謂的最可能猜得準，換句話說，如果我們計算估測值與實際值之間的誤差，我們希望大量、長期猜下來誤差是比較小的。

先作一條直線 $y=mx+b$，若由 $x=x_i$ 去估測 y 值，就把 $x=x_i$ 代到直線 $y=mx+b$ 中，得到我們估測 $y=mx_i+b$。而實際上的數據，則是 (x_i,y_i)，如下圖所示。估測的誤差，就是估測值 $y=mx_i+b$ 與實際值 $y=y_i$ 之間的差，也就是圖中兩點的鉛直距離。



這個距離就是 $\left|y_i-(mx_i+b)\right|$，如果我們能決定出 $y=mx+b$ 的係數 m 與 b，使得所有的距離加起來要最小，我們可以相信這條直線符合我們需求。但是實際運算會超級麻煩，我們居然要把一堆絕對值相加，還要求怎樣的 m,b 會使這個加總極小！為了運算上簡便，我們將「所有誤差的絕對值總和」改成「所有誤差的平方和」。這是因為平方和比起絕對值的和好處理多了，這跟標準差的定義也是取平方和是一樣的道理。所以我們現在改這樣說，如果我們能決定出係數 m 與 b，使得所有的誤差平方 $\big(y_i-(mx_i+b)\big)^2$ 加起來要最小，這

樣的 $y=mx+b$ 就是迴歸直線。這個方法叫做 **最小平方法**，而這樣求出的迴歸直線又可稱 **最小平方直線**。



先 將 Y 與 X 都先標準化，得到 Y' 與 X'，則 Y' 對 X' 的最小平方直線，經過一番很複雜的運算後，可得到簡單的結果

$$y'=r\cdot x'$$

數據標準化後迴歸直線斜率恰為相關係數。接著再代 $x'=\dfrac{x-\mu_x}{\sigma_x}$，$y'=\dfrac{y-\mu_y}{\sigma_y}$，得到

$$\frac{y-\mu_y}{\sigma_y}=r\cdot\frac{x-\mu_x}{\sigma_x}$$

再移項後就有

$$y-\mu_y=r\cdot\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

這就是 Y 對 X 的迴歸直線公式。

注意

(1) 迴歸直線必過 (μ_x,μ_y)。這條直線既然能代表全體，當然也通過具代表性的平均數。

(2) 若是想改由 Y 來估測 X，就要改求 X 對 Y 的迴歸直線，這樣通常會是不同條直線，因為這樣估測誤差就會變成水平距離而非鉛直距離。

(3) 套用相關係數與標準差的公式，寫成

$$y-\mu_y=r\cdot\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

$$y-\mu_y=\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum(x_i-\mu_x)^2}\sqrt{\sum(y_i-\mu_y)^2}}\cdot\frac{\sqrt{\frac{\sum(y_i-\mu_y)^2}{n}}}{\sqrt{\frac{\sum(x_i-\mu_x)^2}{n}}}\cdot(x-\mu_x)$$

$$\Rightarrow y-\mu_y=\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{\sum(x_i-\mu_x)^2}\cdot(x-\mu_x)$$

這樣許多時候算起來更快。

(4) 物理的運動學中有個公式 $v=v_0+at$。如果我們做個實驗，固定了初速 v_0 與加速度 a，然後測量在不同的時間 t 時的速度 v。因為測量數據難免有些許誤差，所以將數據描點出來後，不會那麼完美地成一直線。但如果我們求迴歸直線，就會很接近 $v=v_0+at$。換句話說，想透過做實驗來驗證已知公式或者猜測未知公式，在實驗難免有誤差的情況下，求迴歸直線是個重要手段。

(5) 為何叫「迴歸」直線？因為統計學家 Galton 發現，父母的身高越高，子代通常較父母矮而較普通人高，

是為「迴歸平庸」(Regression towards mediocrity)。後來為紀念 Galton，便沿用了迴歸直線 (regression line) 這一名稱。

例題 2 設 $(x_1,y_1)=(2,3)$，$(x_2,y_2)=(3,1)$，$(x_3,y_3)=(4,2)$，$D=(y_1-a-bx_1)^2+(y_2-a-bx_2)^2+(y_3-a-bx_3)^2$。試求出實數 a 與 b 使得 D 有最小值。

解

若直線 L 方程為 $y=bx+a$，則點 (x_i,y_i) 鉛直對應到 L 線上的點就是 (x_i,bx_i+a)。所以 $D=\sum_{i=1}^{3}(y_i-(bx_i+a))^2$ 的意義就是鉛直距離的平方和。這題問的就是最小平方法!

先求出 $\mu_x=3,\mu_y=2$，再利用公式

$$y-\mu_y=\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{\sum(x_i-\mu_x)^2}\cdot(x-\mu_x)$$

就可得到

$$y-2=\frac{(-1)(-1)+0+0}{(-1)^2+0+1^2}\cdot(x-3)$$

$$\Rightarrow y-2=\frac{1}{2}\cdot(x-3)$$

微積分福音

例題 3 經濟學者分析某公司服務年資相近的員工之「年薪」與「就學年數」的資料，得到這樣的結論：『員工就學年數每增加一年，其年薪平均增加 8 萬 5 千元』。試問上述結論可直接從下列哪些選項中的統計量得到？

(1) 「年薪」眾數與「就學年數」眾數
(2) 「年薪」全距與「就學年數」全距
(3) 「年薪」平均數與「就學年數」平均數
(4) 「年薪」與「就學年數」之相關係數
(5) 「年薪」對「就學年數」之迴歸直線斜率

98 數乙

解

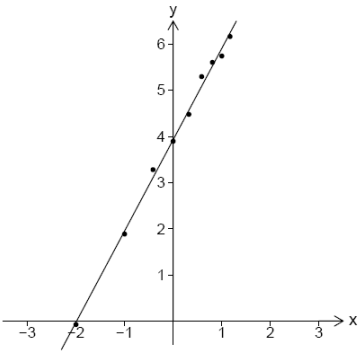
只要你明白什麼叫迴歸直線，這題就是來送分的，一看就是 (5)，別的都不可能。

例題 4 某人進行實驗確定某運動之距離 d 與時間 t 的平方或立方成正比，所得數據如下：

時間 t (秒)	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25
距離 d (呎)	0.95	3.69	9.71	14.88	22.32	39.34	48.68	53.65	71.79

為探索該運動的距離與時間之關係，令 $x=\log_2 t$，$y=\log_2 d$，即將上述的數據 (t,d) 分別取以 2 為底的對數變換，例如：$(2,53.65)$ 變換後成為 $(1,5.74)$。已知變換後的數據 $(x_1,y_1),(x_2,y_2),\cdots(x_9,y_9)$ 之散佈圖及最小平方法所求得變數 y 對變數 x 的最適合直線（或稱迴歸直線）為 $y=a+bx$，如下圖所示

微積分福音



試問下列哪些選項是正確的?

(1) 若 $d=14.88$，則 $3<\log_2 d<4$
(2) x 與 y 的相關係數小於 0.2
(3) 由上圖可以觀察出 $b>2.5$
(4) 由上圖可以觀察出 $a>2$
(5) 由上圖可以確定此運動之距離與時間的立方約略成正比

97 數甲

解

(1) ○ $3=\log_2 8<\log_2 14.88<\log_2 16=4$

(2) × 由圖看來為高度相關

(3) × 由圖觀察斜率約為 $\frac{4}{2}=2$

(4) ○ a 為迴歸直線的 y 截距，大約為 4

(5) × 承 (3)，因斜率大約為 2，故運動距離約與時間平方成正比。

注意

此題便用物理實驗入題，並演示了公式並不一定要是線性的，如果有幾次方的正比（或反比）關係，只要取個對數就變成線性關係了。這同時亦說明了對數的用處。

例題 5 已知以下各選項資料的迴歸直線 (最適合直線) 皆相同且皆為負相關，請選出相關係數最小的選項。

(1)

x	2	3	5
y	1	13	1

(2)

x	2	3	5
y	3	10	2

(3)

x	2	3	5
y	5	7	3

(4)

x	2	3	5
y	9	1	5

(5)

x	2	3	5
y	7	4	4

102 學測

解

觀察每個選項的 X 資料皆相同，每個選項的 Y 資料總和亦皆相同（所以平均也相同）。而每個選項的迴歸直線 $y-\mu_y=r\cdot\frac{\sigma_y}{\sigma_x}(x-\mu_x)$ 相同，若 σ_y 越小，r 便負得越多（相乘是定值），也就是越小。所以要找相關係數最小，只須判斷哪個選項的 σ_y 最小。而顯然 (5) 的 Y 的標準差是最小的（最集中），故選 (5)。

例題 6 統計某一公司在過去 10 年中，每年廣告費 (X) 與營業額 (Y) 的資料為：(x_1,y_1), (x_2,y_2), (x_3,y_3), $\cdots$, (x_{10},y_{10})。計算得到平均數、標準差與相關係數如下：$\mu_X=20$ 萬元，$\mu_Y=1500$ 萬元，$\sigma_X=4$ 萬元，$\sigma_Y=50$ 萬元，$r=0.8$。若今年要有 2000 萬元的營業額，則約需花 _____ 萬元的廣告費。

解

微積分福音

利用迴歸直線公式，寫出

$$y-1500=0.8\cdot\frac{50}{4}(x-20)$$

再代 $y=2000$，得到 $x=70$。

像這樣寫，就錯啦！仔細看清題目，所求是利用營業額 (Y) 估測廣告費 (X)，所以應該寫 X 對 Y 的迴歸直線

$$x-20=0.8\cdot\frac{4}{50}(y-1500)$$

再代 $y=2000$，得到 $x=52$。

千萬要注意是誰對誰的迴歸直線

例題 7 設兩變數 X 與 Y 的 n 筆資料為 $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$，且 $Y=-X+7$，若變數 $X'=\dfrac{X-\mu_x}{\sigma_x}$，$Y'=\dfrac{Y-\mu_y}{\sigma_y}$，其中 μ_x,μ_y 分別為 X,Y 的算術平均數，σ_x,σ_y 分別為 X,Y 的標準差，請選出下列正確選項：

(A) $r_{XY}=1$

(B) $r_{X'Y'}>r_{XY}$

(C) Y' 對 X' 的迴歸直線為 $y'=-x'$

(D) Y' 對 X' 的迴歸直線斜率 m' 等於 Y 對 X 的迴歸直線斜率 m

(E) 若 $\mu_X=0$，則 Y' 對 X' 的迴歸直線與 Y 對 X 的迴歸直線是同一條直線。

解

$Y=-X+7$ 的意思是兩資料完全符合這關係。換句話說，$y_i=-x_i+7$ 對於 $i=1,\cdots,n$ 皆成立。所以所有資料都完全在直線 $y=-x+7$ 上，而這條正是迴歸直線。

(1) $\times$ 所有資料完全在斜率為負的直線上，是為完全負相關，故 $r_{XY}=-1$

(2) $\times$ 標準化不改變相關係數

(3) $\bigcirc$ 標準化數據 Y' 對 X' 的迴歸直線為 $y'=rx'$

(4) $\bigcirc$ 承 (3)，$m=m'=-1$

(5) $\times$ 我們前面已分析出這兩條迴歸直線，無論 μ_x 值為何都不改變。

微積分福音

例題 8 某校高三共有 300 位學生，數學科第一次段考、第二次段考成績分別以 X，Y 表示，且每位學生的成績用 0 至 100 評分。若這兩次段考數學科成績的相關係數為 0.016，試問下列哪些選項是正確的？

(1) X 與 Y 的相關情形可以用散佈圖表示

(2) 這兩次段考的數學成績適合用直線 $X = a + bY$ 表示 X 與 Y 的相關情形
（a, b 為常數，$b \neq 0$）

(3) $X+5$ 與 $Y+5$ 的相關係數仍為 0.016

(4) $10X$ 與 $10Y$ 的相關係數仍為 0.016

(5) 若 $X' = \dfrac{X - \overline{X}}{S_X}$、$Y' = \dfrac{Y - \overline{Y}}{S_Y}$，其中 $\overline{X}$、$\overline{Y}$ 分別為 X、Y 的平均數，S_X、S_Y 分別為 X、Y 的標準差，則 X' 與 Y' 的相關係數仍為 0.016

96 指考甲

解

(1) ○ 當然可以，怎麼不可以？

(2) × 相關係數 0.016 太低，故不適合

(3) ○ $r(X+5, Y+5) = r(X, Y) = 0.016$
資料平移不影響相關係數

(4) ○ $r(10X, 10Y) = r(X, Y) = 0.016$
資料伸縮且伸縮係數同號，不影響相關係數

(5) ○ 標準化不影響相關係數

■ 4 總結

如何記憶相關係數公式?

1. 計算數據標準化後的「面積」平均，即 $\dfrac{\sum x_i' \cdot y_i'}{n}$

2. 代回原數據，得

$$\frac{\sum\left(\frac{x_i-\mu_x}{\sigma_x}\right)\left(\frac{y_i-\mu_y}{\sigma_y}\right)}{n}=\underbrace{\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{n\sigma_x\sigma_y}}_{\text{公式1}}$$

3. 套用標準差公式，得

$$\underbrace{\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum(x_i-\mu_x)^2}\sqrt{\sum(y_i-\mu_y)^2}}}_{\text{公式2}}$$

如何記憶迴歸直線公式?

1. 數據標準化後，Y' 對 X' 的迴歸直線為

$$y_i' = r \cdot x_i'$$

2. 代回原數據，得

$$\frac{y_i-\mu_y}{\sigma_y} = r\left(\frac{x_i-\mu_x}{\sigma_x}\right)$$

$$\Rightarrow \underbrace{y_i-\mu_y = r\cdot\frac{\sigma_y}{\sigma_x}(x_i-\mu_x)}_{\text{公式1}}$$

3. 套用標準差公式，得

$$\underbrace{y-\mu_y=\frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{\sum(x_i-\mu_x)^2}\cdot(x-\mu_x)}_{\text{公式2}}$$

微積分福音