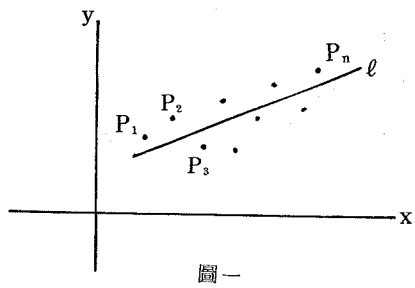
迴歸直線與相關係數

葉東進

國立科學園區實驗高中

科學活動的主要內容在通過某些基本模式的建立，以尋找變量之間的關係或規律，但是有些變量之間並不存在必然的關係或是明顯的規律，譬如人的身高與體重、學生的理化成績與數學成績、產品的廣告費與銷售量等；然而它們變化之間卻又隱約有某個趨勢存在。例如兩個變量 x 與 y 的數據：

x	x_1 , x_2 , x_3 , ····· , x_n
y	y_1 , y_2 , y_3 , ····· , y_n



圖一

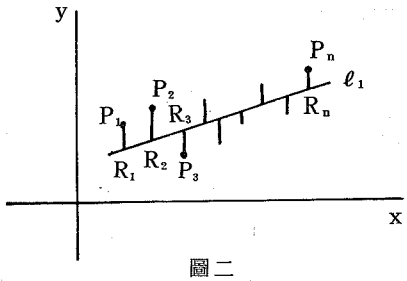
從它們的散佈情形來看（圖一），雖然沒能找到一個基本的模式 f 來直接表出 x 與 y 的關係 $y = f(x)$，但是隱約之間，我們看到由這些數據所建構的點 P_i（x_i , y_i）是有向某一直線靠攏的趨勢。因此，退而求其次，我們設法找出這一直線，並對靠攏的趨勢程度加以量化。

一、迴歸直線

如（圖一）中的直線 ℓ，其意義可以有許多不同的解釋，只要不離開這些點太離譜

，隨便哪一條直線都可以，問題是：什麼方法才能使找到的直線可以最佳地表達出數據間的變化趨勢？「最小方差法」便是其中的方法之一：

首先，考慮諸點 $P_i(x_i, y_i)$（$i = 1, 2, 3, \cdots\cdots, n$）沿縱軸方向在 ℓ_1 上的投影 R_i（圖二），使 $\Sigma\overline{P_iR_i}^2$ 爲最小。



圖二

爲方便處理，取 ℓ_1 的方程式爲　$y = mx + k$

則有 $\Sigma\overline{P_iR_i}^2 = \Sigma(mx_i + k - y_i)^2$（記爲 $f(m, k)$）

由　$\dfrac{\partial f}{\partial m} = \dfrac{\partial f}{\partial k} = 0$

得　$\begin{cases} \Sigma(mx_i + k - y_i)x_i = 0 \\ \Sigma(mx_i + k - y_i) = 0 \end{cases}$

$\Rightarrow \begin{cases} (\Sigma x_i^2)m + (\Sigma x_i)k = \Sigma x_iy_i \\ (\Sigma x_i)m + n \cdot k = \Sigma y_i \end{cases}$

解得　$m = \dfrac{\Sigma x_iy_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$，其中 $\bar{x} = \dfrac{1}{n}\Sigma x_i$，$\bar{y} = \dfrac{1}{n}\Sigma y_i$

由　$k = \dfrac{1}{n}\Sigma y_i - (\dfrac{1}{n}\Sigma x_i)m$

$\Rightarrow$　$k = \bar{y} - m\bar{x}$

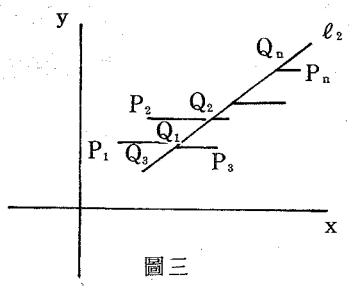
$\therefore$　$\bar{y} = m\bar{x} + k$

此式表示直線 ℓ_1：$y = mx + k$ 通過點 $(\bar{x}, \bar{y})$，因此 ℓ_1 的方程式亦可寫成：

$$\ell_1: \quad y - \bar{y} = \frac{\Sigma x_iy_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}(x - \bar{x}) \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

我們把直線 ℓ_1 稱爲 y 對 x 的迴歸直線。

其次，我們也必須考慮諸點 $P_i(x_i, y_i)$（$i = 1, 2, 3, \cdots\cdots, n$）沿橫軸方向



圖三

在 ℓ_2 上的投影 Q_i（圖三），使 $\Sigma\overline{P_iQ_i}^2$ 爲最小。

取 ℓ_2 的方程式爲

$$x = my + k$$

則有 $\Sigma\overline{P_iQ_i}^2 = \Sigma(my_i + k - x_i)^2$（記爲 $g(m, k)$）

由 $$\frac{\partial g}{\partial m} = \frac{\partial g}{\partial k} = 0$$

得

$$\begin{cases} \Sigma(my_i + k - x_i)y_i = 0 \\ \Sigma(my_i + k - x_i) = 0 \end{cases} \Rightarrow \begin{cases} (\Sigma y_i)^2 m + (\Sigma y_i)k = \Sigma x_i y_i \\ (\Sigma y_i)m + n \cdot k = \Sigma x_i \end{cases}$$

解得 $$m = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma y_i^2 - n\bar{y}^2}$$

由 $$k = \frac{1}{n}\Sigma x_i - \left(\frac{1}{n}\Sigma y_i\right)m$$

$$\Rightarrow \quad k = \bar{x} - m\bar{y}$$

$$\therefore \quad \bar{x} = m\bar{y} + k$$

此式表示直線 ℓ_2：$x = my + k$ 通過點 $(\bar{x}, \bar{y})$，因此 ℓ_2 的方程式亦可寫成：

$$\ell_2: \quad x - \bar{x} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma y_i^2 - n\bar{y}^2}(y - \bar{y})$$

或是 $$y - \bar{y} = \frac{\Sigma y_i^2 - n\bar{y}^2}{\Sigma x_i y_i - n\bar{x}\bar{y}}(x - \bar{x}) \quad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

我們把直線 ℓ_2 稱爲 x 對 y 的迴歸直線。

顯然，ℓ_1 與 ℓ_2 的差異如果愈小，諸點 P_i 靠攏 ℓ_1 與 ℓ_2 的趨向便愈明顯。

二、相關係數

把前述 y 對 x 的迴歸直線(1)及 x 對 y 的迴歸直線(2)的斜率分別記爲 m_1 與 m_2 ：

$$m_1 = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$$

$$m_2 = \frac{\Sigma y_i^2 - n\bar{y}^2}{\Sigma x_i y_i - n\bar{x}\bar{y}}$$

由 $$\begin{aligned}\Sigma(x_i-\bar{x})(y_i-\bar{y}) &= \Sigma x_i y_i - (\Sigma x_i)\bar{y} - (\Sigma y_i)\bar{x} + n\bar{x}\bar{y} \\ &= \Sigma x_i y_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y} \\ &= \Sigma x_i y_i - n\bar{x}\bar{y}\end{aligned}$$

及 $$\begin{aligned}\Sigma(x_i-\bar{x})^2 &= \Sigma(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \Sigma x_i^2 - 2(\Sigma x_i)\bar{x} + n\bar{x}^2 \\ &= \Sigma x_i^2 - 2n\bar{x}\cdot\bar{x} + n\bar{x}^2 \\ &= \Sigma x_i^2 - n\bar{x}^2\end{aligned}$$

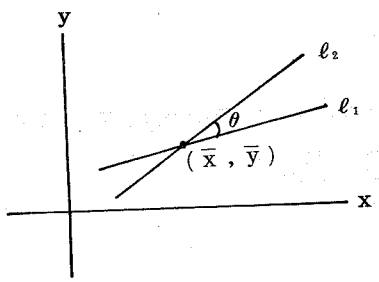
所以 $$m_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$$

$$m_2 = \frac{\Sigma(y_i-\bar{y})^2}{\Sigma(x_i-\bar{x})(y_i-\bar{y})}$$

考慮 $\frac{m_1}{m_2} = \frac{[\Sigma(x_i-\bar{x})(y_i-\bar{y})]^2}{\Sigma(x_i-\bar{x})^2 \cdot \Sigma(y_i-\bar{y})^2}$ (>0) ，不失一般性，底下的討論均假定 $m_1 > 0$ ， $m_2 > 0$ ，而且 $m_2 \geq m_1$ 。

取 $K = \frac{m_1}{m_2}$

可以看出，在 m_1 的值固定的情況下， m_2 與 K 成反比。就是說 m_2 愈小， K 值隨之

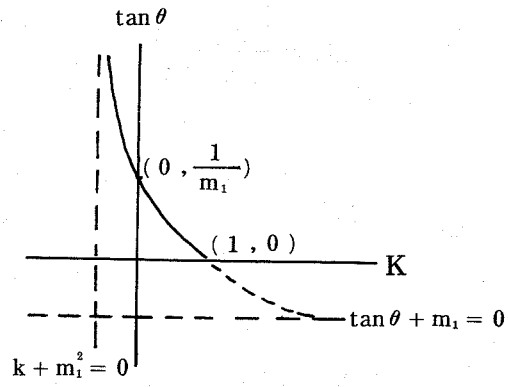


圖四

愈大；m_2 愈大，K值隨之愈小。幾何意義看來，便是 ℓ_1 與 ℓ_2 的銳夾角 θ 愈小，K值隨之愈大；θ 愈大，K值隨之愈小（圖四）。因此，K值的大小能夠適度地反映出 ℓ_2 與 ℓ_1 之間的差異。

事實上，由

$$\tan\theta=\frac{m_2-m_1}{1+m_2m_1}=\frac{1-\frac{m_1}{m_2}}{\frac{1}{m_2}+m_1}=\frac{1-K}{\frac{K}{m_1}+m_1}$$



圖五

因爲 θ 爲銳角，$\tan\theta\geq 0$，即 $0<K\leq 1$。

另外，從 $(\tan\theta+m_1)(K+m_1^2)=m_1+m_1^3$ 可以繪出 $\tan\theta$ 與K之間的變化關係圖形（圖五），從圖形中我們看出 $\tan\theta$ 是隨K值的增大而遞減；隨K值的減小而遞增。也可以說 θ 與K值成遞減關係。

既然K值的大小可以反映出 ℓ_2 與 ℓ_1 間的差異，而 ℓ_2 與 ℓ_1 間的差異又反映出諸點 P_i 靠攏 ℓ_1 與 ℓ_2 的趨向，因此K值的大小便反映出諸點 P_i 靠攏 ℓ_1 與 ℓ_2 的趨向程度。

由 $$K=\frac{m_1}{m_2}=\frac{[\Sigma(x_i-\bar{x})(y_i-\bar{y})]^2}{\Sigma(x_i-\bar{x})^2\cdot\Sigma(y_i-\bar{y})^2}>0$$

取 $$\sigma=\sqrt{K}$$

我們把 σ 稱作是兩個變量 x 與 y 之間的相關係數。

因此，兩個變量 x 與 y 間的相關係數便是反應諸點 $P_i(x_i, y_i)$ 靠攏迴歸直線的程度的一個數量。

參考資料

1. 高中基礎數學第四册，師大科教中心，國立編譯館。
2. 普通數學教程，楊維哲、蔡聰明，文仁。
3. 統計學初階，賴建業譯，中央。