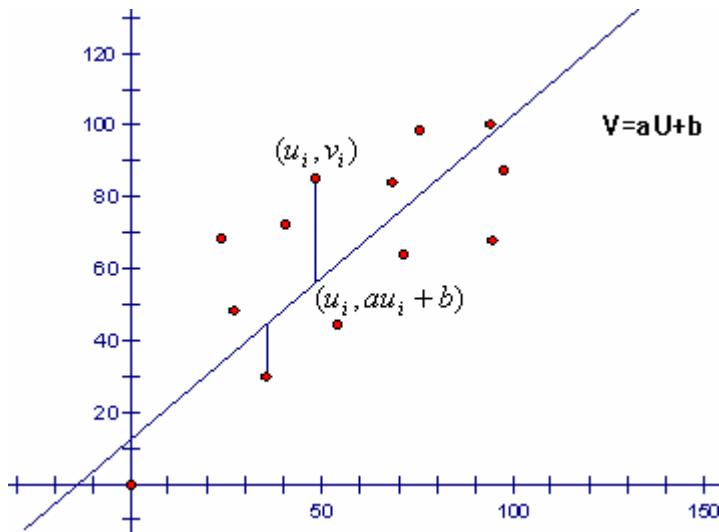


## 相關係數—最小平方法

目的：欲探討兩組資料序對間之關係程度，例如數學成績好，是否物理的成績也會比較好。讀書所花的時間愈多，是否成績也會比較好。社團參加的時數愈多，是否成績也會比較差。

散佈圖：將各資料序對畫在以資料項目為橫座標及縱座標的座標平面上，標出點的位置，以便觀察各資料序對分佈狀況，及兩資料項目間關係的程度。此圖即稱為散佈圖。

方法：希望在散佈圖中找出一條直線，以代表兩資料項目之線性關係，且此直線是最適合的，散佈圖中各點至此直線之鉛直距離平方和為最小，此方法稱為最小平方法。此時該條直線稱為最適合直線（最佳近似直線），亦稱為迴歸直線。



有  $n$  筆原始資料序對  $(x_i, y_i)$ ， $1 \leq i \leq n$ ，其資料項目分別稱為 X 與 Y，且 X 的平均數  $\bar{x}$ ，標準差  $s_x$ ；Y 的平均數  $\bar{y}$ ，標準差  $s_y$ 。

今將 X 以  $s_x$  為單位予以標準化得各資料項分別為  $u_i = \frac{x_i - \bar{x}}{s_x}$ ， $1 \leq i \leq n$ ，且  $\sum_{i=1}^n u_i$

$$= \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} = \frac{1}{s_x} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{s_x} [\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x}] = \frac{1}{s_x} (n\bar{x} - n\bar{x}) = 0$$

將 Y 以  $s_y$  為單位予以標準化得各資料項分別為  $v_i = \frac{y_i - \bar{y}}{s_y}$ ， $1 \leq i \leq n$ ，且  $\sum_{i=1}^n v_i =$

$$\sum_{i=1}^n \frac{y_i - \bar{y}}{s_y} = \frac{1}{s_y} \sum_{i=1}^n (y_i - \bar{y}) = \frac{1}{s_y} [\sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y}] = \frac{1}{s_y} (n\bar{y} - n\bar{y}) = 0$$

今欲求  $(u_i, v_i)$  與  $(u_i, au_i + b)$  鉛直距離平方和  $\sum_{i=1}^n [v_i - (au_i + b)]^2$ ，並使其值為最小

小，那時  $a$  與  $b$  各為多少？找出了  $a$  與  $b$ ，則最佳近似直線  $v = au + b$ （或  $y = ax + b$ ）也就決定了。

$$\begin{aligned}
 & \sum_{i=1}^n [v_i - (au_i + b)]^2 \\
 &= \sum_{i=1}^n [v_i^2 - 2v_i(au_i + b) + (au_i + b)^2] \\
 &= \sum_{i=1}^n v_i^2 - 2\sum_{i=1}^n v_i(au_i + b) + \sum_{i=1}^n (au_i + b)^2 \\
 &= \sum_{i=1}^n v_i^2 - 2a\sum_{i=1}^n u_i v_i - 2b\sum_{i=1}^n v_i + a^2 \sum_{i=1}^n u_i^2 + 2ab \sum_{i=1}^n u_i + \sum_{i=1}^n b^2 \\
 &= \sum_{i=1}^n v_i^2 + nb^2 + a^2 \sum_{i=1}^n u_i^2 - 2a \sum_{i=1}^n u_i v_i \\
 &= \sum_{i=1}^n v_i^2 + nb^2 + \sum_{i=1}^n u_i^2 [a^2 - 2 \frac{\sum_{i=1}^n u_i v_i}{\sum_{i=1}^n u_i^2} a + (\frac{\sum_{i=1}^n u_i v_i}{\sum_{i=1}^n u_i^2})^2] - \frac{(\sum_{i=1}^n u_i v_i)^2}{\sum_{i=1}^n u_i^2} \\
 &= \sum_{i=1}^n v_i^2 + nb^2 + \sum_{i=1}^n u_i^2 (a^2 - \frac{\sum_{i=1}^n u_i v_i}{\sum_{i=1}^n u_i^2})^2 - \frac{(\sum_{i=1}^n u_i v_i)^2}{\sum_{i=1}^n u_i^2}
 \end{aligned}$$

故當  $a = \frac{\sum_{i=1}^n u_i v_i}{\sum_{i=1}^n u_i^2}$  且  $b = 0$  時， $\sum_{i=1}^n [v_i - (au_i + b)]^2$  有最小值。

$$\begin{aligned}
 a &= \frac{\sum_{i=1}^n u_i v_i}{\sum_{i=1}^n u_i^2} \\
 &= \frac{\sum_{i=1}^n (\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})}{\sum_{i=1}^n (\frac{x_i - \bar{x}}{s_x})^2}
 \end{aligned}$$

$$\begin{aligned}
&= \frac{s_x}{s_y} \times \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}}{\sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}}} \times \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}
\end{aligned}$$

故直線  $v = au + b$  即為  $v = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} u$

若定義  $s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ ，則  $a = \frac{s_{xy}}{s_x s_y}$ ，則直線  $v = au + b$  即為  $v = \frac{s_{xy}}{s_x s_y} u$ 。

$\frac{s_{xy}}{s_x s_y}$  即為迴歸直線之斜率，代表意義為：X 變動一標準單位，Y 變動多少標準

單位，可將其視為 Y 對 X 的變動關係，故定義其為 Y 與 X 的相關係數，相關係

$$\text{數 } r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \circ$$

迴歸直線可表示為

$$\frac{Y - \bar{y}}{s_y} = \frac{s_{xy}}{s_x s_y} \left( \frac{X - \bar{x}}{s_x} \right) \text{ 或 }$$

$$Y - \bar{y} = \frac{s_{xy}}{s_x^2} (X - \bar{x}) \text{ 或 }$$

$$Y - \bar{y} = r \times \frac{s_y}{s_x} (X - \bar{x})$$