國立政治大學應用數學系

數學教學碩士在職專班

碩士學位論文

相關係數面面觀

Several Ways to Look at the Correlation Coefficient

碩專班學生： 江美菊 撰

指導教授： 江振東 博士

中華民國 102 年 01 月 09 日

摘 要

相關係數r是一個用以描述兩變量之間線性關係程度的指標。它的值域範圍介於1到–1之間，正、負號表示兩變量之間的相關方向，而$|r|$的大小則表示兩變量間相關程度的強弱。本文主要從皮爾森積差相關係數的概念下手，從不同向度切入來探討兩變量間的線性相關性，提供多樣面向的兩變量相關強弱程度的解釋與演繹計算的方法。

Abstract

The correlation coefficient r is an indicator of the degree of linear relation between two variables. Its values range between 1 and -1; positive signs and negative signs indicate the directions of correlations between two variables. Its absolute value indicates the strength of correlation between two variables. This study starts with the Pearson product-moment correlation coefficient and explores the linear correlation between the two variables from different aspects. We then provide various explanations for the strength of linear correlation between two variables and its calculation methods.

目　錄

第一章 研究動機 1

第二章 皮爾森積差相關係數 2

第一節 圖形散佈 2

第二節 r 的值域 4

第三章 資料標準化線性關係不變 6

第一節 變數資料標準化後，線性關係不變 6

第二節 公式推導 7

第四章 標準化共變異數 8

第一節 共變異數 8

第二節 相關係數與共異變數關係 8

第五章 相關係數與標準化迴歸直線的斜率關係 11

第一節 簡單迴歸分析 11

第二節 皮爾森積差相關係數主要是用於直線關係。 13

第三節 相關係數 r 的絕對值等同兩迴歸直線斜率的幾何平均數 14

第六章 方向餘弦與相關係數 15

第一節 向量內積 15

第二節 歐氏空間 $\Re^n$ 16

第七章 兩個標準化迴歸直線的夾角與相關係數關係 18

第一節 相關係數不因觀測變數角度不同改變 18

第二節 兩直線銳角夾角與皮爾森相關係數 r 的關係 19

第三節 公式推導 19

第八章 判定係數與相關係數 21

第一節 判定係數 R^2 21

第二節 判定係數 R^2 = 相關係數 r^2 22

第九章 結論與建議 24

參考文獻 25

附錄 26

相關名詞解釋 26

圖檔 28

第一章 研究動機

我們常常想要了解兩個變數間是否具有某種關聯性，比方說如果其中一個變數增加時，另一個變數是否也會增加？或者是其中一個增加時，另一個反而減少？

畢竟一大堆資料是沒辦法隨身帶著走的。想要形容一組資料中兩變數間的關聯性時，如果可以藉由一個簡單的指標來做描述，自然是再好不過了。因此究竟有沒有辦法可以藉由一個簡單的統計量來描述兩個變數之間的關聯性呢？

統計學上，用來描述兩個變數間線性相關性強弱的一個工具就是「相關係數」(Correlation Coefficient)。基本上就是希望設法用一個『數』來表示兩者之間線性關聯程度的大小，同時說明關聯的方向(亦即其中一個變數變大時，另一個變數會隨之變大或變小)。

本研究主要是以高中課程中所介紹的相關係數定義方式為出發點，希望藉由不同向度來切入了解相關係數 r 的可能意涵。主要參考文獻為 Joseph Rodgers 及 W. Alan Nicewander 的論述. *Thirteen Ways to Look at the Correlation Coefficient*.（Rodgers and Nicewander (1988)），擷取其中適合高中學生程度的觀點來進行探討。

第二章為皮爾森積差相關係數 r 的定義介紹。首先以皮爾森積差相關係數 r 來說明兩變數間資料的散佈關係，再利用柯西不等式討論 r 的值域範圍。第三章為第二章的觀念延伸，但是改將變數的資料先行標準化後，再來看兩變數間的相關性；藉由數理觀點驗證兩變數間的線性關係並不會因為標準化後而改變其原有的線性關係。第四章以共變異數的向度來看相關係數，這是因為共變異數標準化後，去除單位的因素干擾，焦點將著重在於測量兩變數線性關係強弱的判斷。第五章，藉由直線的線性關係和斜率，以迴歸直線向度切入來看相關係數。由線性關係來看，標準化迴歸直線的斜率與兩標準化迴歸直線斜率的幾何平均數均可用以表示相關係數。觀看兩標準化兩迴歸直線與兩軸的夾角相等關係，可以得出變數間相關性的強弱不會因為端看資料的角度不同而改變。第六章以向量的方向餘弦角度切入看相關係數；第七章以兩迴歸直線交角向度看相關係數。第八章以判定係數 R^2 為結尾，利用判定係數 R^2 和相關係數 r 的關係，計算求得 r 值。第九章則為結論與建議。

第二章 皮爾森積差相關係數

皮爾森積差相關係數(Pearson Product Moment Correlation Coefficient)r探究的是兩個變數之間的線性關係，其中這兩個變數在本質上必須是連續的，換句話說，這些變數理論上可以取用某個連續區間中的任何數值，例如身高、年齡、考試成績或收入。至於其他不連續的隨機變數，例如種族〈如白人和黑人〉、社會階級〈如高和低〉和政治背景〈如民進黨或國民黨〉等，則不在探討的範圍內。

令(X_1,Y_1)、(X_2,Y_2)、…、(X_n,Y_n) 為n筆資料。就此資料而言，皮爾森積差相關係數r的定義如下：

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}}$$

其中，$\bar{X}$ 、$\bar{Y}$分別為變數X 、Y的算術平均數。

第一節　圖形散佈

利用散佈圖來觀看兩變數資料的關係，很有感覺。不過，兩變數資料的相關程度可以用數值加以表現嗎？這是一個很有趣的問題。

算術平均數(簡稱平均值)稱為資料的中心點。簡單來說，當每一個資料都減去平均值時，這些數的總和就是0，亦即$\sum\left(X-\bar{X}\right)=0$、$\sum\left(Y-\bar{Y}\right)=0$；其中$\left(X-\bar{X}\right)$、$\left(Y-\bar{Y}\right)$為離均差，也就是變數資料$X$ 、Y和平均值$\bar{X}$ 、$\bar{Y}$的差。因為離均差總和為0，所以我們稱平均值為資料的中心點，因此平均值是非常重要的「資料代表數」。

以$\left(\bar{X},\bar{Y}\right)$為座標點原點來觀看資料分散的情形，恰可將散佈圖分割為四個象限。第一象限內各點之$\left(X-\bar{X}\right)$及$\left(Y-\bar{Y}\right)$均為正，其乘積$\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)$為正；第二象限內各點之$\left(X-\bar{X}\right)$為負，而$\left(Y-\bar{Y}\right)$為正，其乘積為負；第三象限內各點之$\left(X-\bar{X}\right)$及$\left(Y-\bar{Y}\right)$均為負，其乘積為正；第四象限內各點之$\left(X-\bar{X}\right)$為正，而$\left(Y-\bar{Y}\right)$為負，其乘積為負。如下圖 2.1-1 所示：

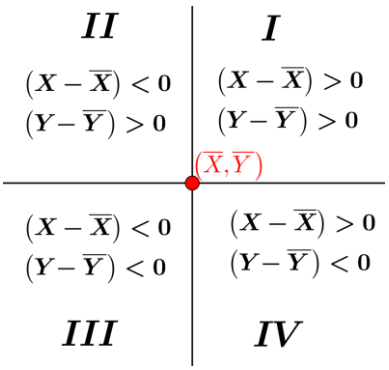


圖 2.1-1

圖 2.1-2 顯示，資料落於第一、三象限的數量明顯多於第二、四象限，兩變數間似乎呈現出正比的關係。反之，圖 2.1-3 則顯示資料落於第二、四象限數量明顯多於第一、三象限，兩變數間似乎呈現出反比的關係。

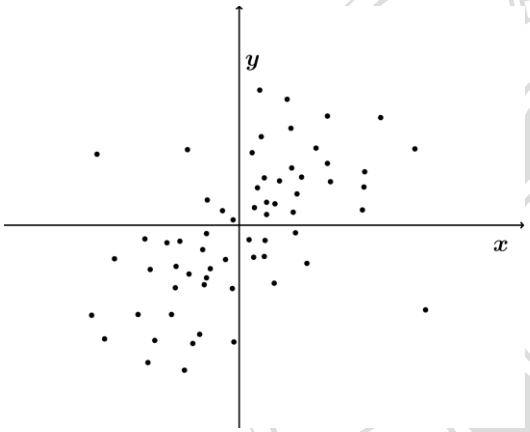


圖 2.1-2

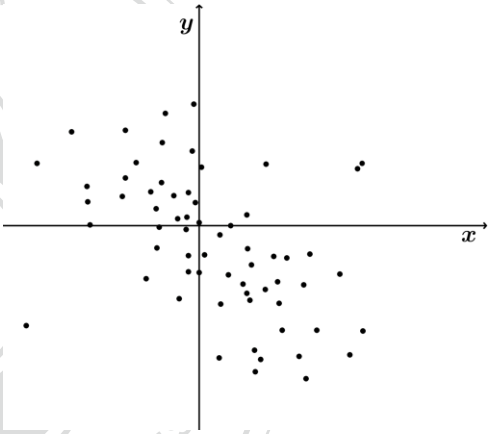


圖 2.1-3

由於兩變數 X 與 Y 若呈現正比關係時，也就是說 X 值上升(下降)時，Y 值也呈現上升(下降)的趨勢，則多半的 $(X-\overline{X})(Y-\overline{Y})$ 均為正，所以其和 $\sum(X-\overline{X})(Y-\overline{Y})$ 也為正，當關聯程度愈高時，則資料散佈在第一、第三象限的內的點數也會愈多，因此 $\sum(X-\overline{X})(Y-\overline{Y})$ 之值必為較大正數。反之，若兩變數成反比關係時，也就是說 X 值上升(下降)時，Y 值則呈現下降 (上升)的趨勢，則多半的 $(X-\overline{X})(Y-\overline{Y})$ 為負，因此 $\sum(X-\overline{X})(Y-\overline{Y})$ 為負，同樣地，關聯程度愈高，則密集於第二、第四象限內的點數愈多，因此 $\sum(X-\overline{X})(Y-\overline{Y})$ 之值必為愈小的負數。

藉由前述的說明可以發現，由於 $\sum(X-\overline{X})(Y-\overline{Y})$ 之正負與大小，恰好可以反映出變數間究竟呈現出正比或反比的關係以及關係的強弱，因此 $\sum(X-\overline{X})(Y-\overline{Y})$ 可作為測量兩變數 X 與 Y 的相關程度

的一個指標。

綜合上述的觀察，我們可以發現下列結果：

1. 當$r>0$時，X與Y成正比關係，也就是說X值上升(下降)時，Y值也呈現上升(下降)的趨勢。反之，當$r<0$時，X與Y成反比關係，也就是說X值上升(下降)時，Y值則呈現下降 (上升)的趨勢。

2. $\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)>0$，$r>0$； $\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)<0$，$r<0$。

因此皮爾森積差相關係數r主要著眼於說明變數X、Y間所呈現出的線性關係，其中正比關係$(r>0)$我們稱之為正相關；反比關係$(r<0)$我們則稱之為負相關。

第二節 r的值域

皮爾森積差相關係數r的值域為$-1\le r\le 1$；這點可藉由柯西不等式的觀點來作說明。

引理：柯西不等式

設a_1、a_2、…、a_n，b_1、b_2、…、b_n為$2n$個實數，則

$$\left(a_1^{\,2}+a_2^{\,2}+\cdots+a_n^{\,2}\right)\left(b_1^{\,2}+b_2^{\,2}+\cdots+b_n^{\,2}\right)\ge\left(a_1b_1+a_2b_2+\cdots+a_nb_n\right)^2$$

等號"="成立的時機為$\frac{a_1}{b_1}=\frac{a_2}{b_2}=\frac{a_3}{b_3}=\cdots=\frac{a_n}{b_n}$

令 $\begin{cases}a_i=X_i-\overline{X}\\ b_i=Y_i-\overline{Y}\end{cases}$

則 $$r^2=\frac{\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)\right]^2}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}=\frac{\left[\sum_{i=1}^{n}a_ib_i\right]^2}{\sum_{i=1}^{n}a_i^{\,2}\sum_{i=1}^{n}b_i^{\,2}}$$

根據柯西不等式，可得$r^2\le 1$，亦即$-1\le r\le 1$。因此相關係數r的值恆介於-1與$+1$之間。

一般而言相關係數主要著眼於資料是否近似線型結構，或者是散成一團？當然，實際上的資料的呈現並不是這麼單純的兩極化。如圖 2.2-4，相關係數的值會是什麼樣，要看資料的散佈狀況。

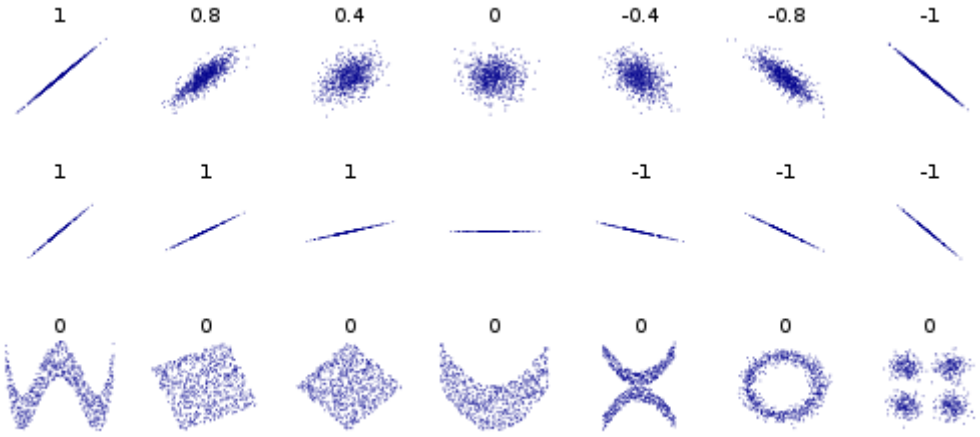


圖 2.2-1　當中數字是該圖象代表的資料之相關係

圖相來源：http://en.wikipedia.org/wiki/Correlation

當資料散得越亂的時候，相關係數會接近於零。當資料的組成越接近一條斜率為正的直線時，相關係數就會越接近於 1；反之如果越近似一條斜率為負的直線時，則相關係數就會越接近於 -1。不過，要注意的是就算資料能夠聚成為一條直線，但是如果這條直線的角度越接近於水平或是垂直時相關係數也會越接近於零，這表示 X 的變動與 Y 的變動之間並沒有關係。經驗法則告訴我們，相關係數 r 值的大小與兩變數相關程度強弱的界定，大致如下表所示：

相關係數絕對值	相關程度
約=1	完全相關 (Perfect Correlated)
0.7~0.99	高度相關 (Highly Correlated)
0.4~0.69	中度相關 (Moderately Correlated)
0.1~0.39	低度相關 (Modestly Correlated)
0.01~0.09	接近無相關 (Weakly Correlated)
約=0	無相關

表 2.2-1

第三章 資料標準化線性關係不變

皮爾森積差相關係數r，亦可透過下列方式來計算：

$$r = \frac{\sum Z_X Z_Y}{n} \qquad (3.1)$$

其中Z_X為X變項標準化過後的Z分數、Z_Y為Y變項標準化過後的Z分數、n為樣本總數。亦即，將變數先行標準化後，相關係數恰為其內積的平均值。

第一節 變數資料標準化後，線性關係不變

因為大都數資料多是散亂不一致的。所以將資料標準化的用意在為以整組數據資料的標準差為新的度量單位，計算每筆數據距離平均值有幾個標準差。標準化後的資料將是一個無單位相對數值。這在直覺上和實務上對於若碰上兩變數資料單位不同而欲觀看這兩變數之間的關係時，將會變得使人更容易分辨。標準化資料除了有這個特性，另外還有一個好處，就是它們的平均值都是0。把標準化的變數資料放到座標平面上，它們的中心點就會落在原點上；如圖 3.1-1 所示。

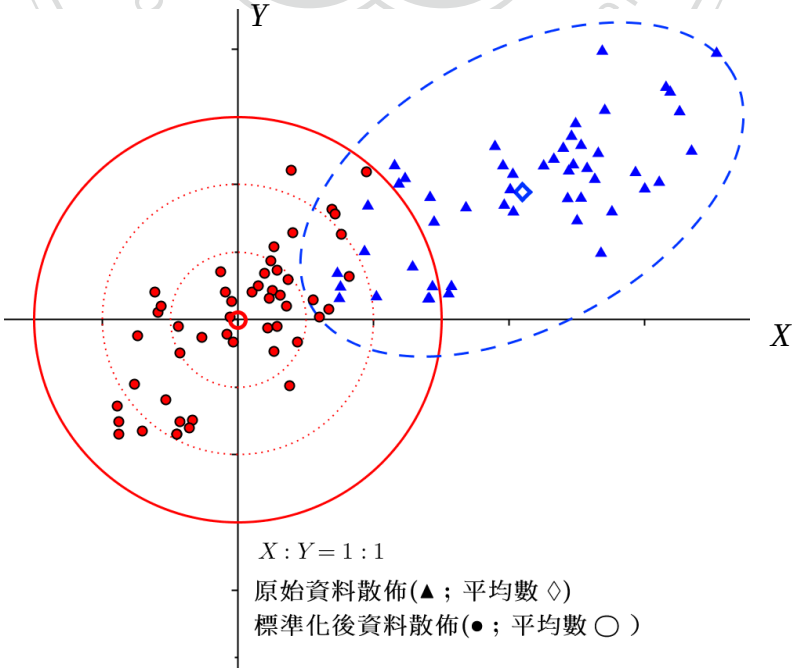


圖 3-1-1 資料標準化後$S_{Z_X} = S_{Z_Y} = 1$

將變數 X 、 Y 標準化後可得：

$$Z_X = \frac{\left(X - \overline{X}\right)}{S_X} \quad ; \quad Z_Y = \frac{\left(Y - \overline{Y}\right)}{S_Y}$$

明顯可以看出 $Z_X \bullet Z_Y$ 與 $\left(X - \overline{X}\right) \bullet \left(Y - \overline{Y}\right)$ 同構。由於 $Z_X \bullet Z_Y$ 的正負值與 $\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)$ 相同，若 $\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right) > 0$，則 $\sum Z_X \bullet Z_Y > 0$，亦即 $r > 0$。反之，若 $\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right) < 0$，則 $\sum Z_X \bullet Z_Y < 0$，亦即 $r < 0$。因此，將兩變數標準化後，並不會改變其原來關係，即線性關係不變。另外，標準化後的變數無單位，故皮爾森積差相關係數 (r) 不受討論變數單位不同所影響，重心擺在變數間線性關係的評估。

第二節　公式推導

由於：

$$S_X = \sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n}} \quad ; \quad S_Y = \sqrt{\frac{\sum\left(Y - \overline{Y}\right)^2}{n}}$$

因此皮爾森積差相關係數 r 也可改寫如下：

$$r = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{\sqrt{\sum\left(X - \overline{X}\right)^2} \bullet \sqrt{\sum\left(Y - \overline{Y}\right)^2}}$$

$$= \frac{\frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{n}}{\sqrt{\frac{\sum\left(X - \overline{X}\right)^2}{n}} \bullet \sqrt{\frac{\sum\left(Y - \overline{Y}\right)^2}{n}}}$$

$$= \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{nS_X S_Y} = \frac{\sum \frac{\left(X - \overline{X}\right)}{S_X}\frac{\left(Y - \overline{Y}\right)}{S_Y}}{n} = \frac{\sum Z_X \bullet Z_Y}{n}$$

第四章 標準化共變異數

共變異數(Covariance)與相關係數相同，都可用於測量變數間線性關係。所以在計算相關係數 r 時，我們也可藉由共變異數來求得，亦即：

$$r = \frac{S_{XY}}{S_X S_Y}$$

其中 S_{XY} 為共變異數， S_X 為 X 的標準差， S_Y 為 Y 的標準差。

第一節 共變異數

X 與 Y 的共變異數的定義如下：

$$S_{XY} = \frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{n}$$

X 與 Y 的共變異數可以藉由 X 、 Y 兩變數扣除個別平均值後的乘積和除以樣本總數後計算得出。由於 S_{XY} 的正負值取決於 $\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)$，所以在判斷兩變數之間的正負相關性時， S_{XY} 似乎不失為一項可以用來作為依據的一項指標。

不過通常我們在判斷兩變數間線性關係的強弱時，並不會直接引用它來做為判斷依據，這是因為共變異數的大小易受變數單位的不同而影響，很難藉以說明兩變數間的相關程度到底是大還是小，因此單憑共變異數是不足以描述兩個變數間線性相關性的強弱。

第二節 相關係數與共異變數關係

因共變異數易受變數單位不同的干擾，故欲將共變異數測量的焦點轉移到兩變數線性關係強弱的判斷時，去除變數單位的這一干擾因素是必然的步驟。由於一個隨機變數的標準差可以視為一種測度標準，且標準差與原始變數具有相同的單位，因此直覺的想法是如果將共變異數除以兩變數個別的標

準差，亦即$S_{XY}/(S_X S_Y)$，應該可以解決當變數單位不同時，共變異數無法用來測量兩變數線性關係強弱的缺失。事實的確也是如此。

由前一章式(3-1)已知，在計算相關係數r時，若不使用原始資料，我們可以利用X、Y變數的Z分數來計算相關係數。變數經標準化後，線性關係並不會因此而改變。

由於 $r=\dfrac{\sum Z_X Z_Y}{n}$

$$=\frac{\sum\dfrac{\left(X-\overline{X}\right)}{S_X}\dfrac{\left(Y-\overline{Y}\right)}{S_Y}}{n}=\frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{nS_X S_Y}=\underbrace{\frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{n}}_{S_{XY}(\text{共變異數})}\cdot\frac{1}{S_X S_Y}$$

所以 $r=\dfrac{S_{XY}}{S_X S_Y}$。

資料若先行標準化後，$S_{Z_X}=S_{Z_Y}=1$，因此$r=S_{z_X z_Y}/\left(S_{z_X}\bullet S_{z_Y}\right)=S_{z_X z_Y}$，亦即相關係數$r$＝標準化後的共變異數$\left(S_{z_X z_Y}\right)$。

此外，若將式子改寫為$S_{XY}=r(S_X S_Y)$，分別以$S_X S_Y$為橫坐標、S_{XY}為縱座標，則$S_{XY}=r(S_X S_Y)$可以視為通過原點，斜率為r的直線方程式；這個式子也提供我們另一種探討相關係數特性的方向：

(狀況一) 當$X=Y$時，$S_{XY}=\dfrac{\sum\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n}=\dfrac{\sum\left(X_i-\overline{X}\right)^2}{n}=S_X{}^2=S_Y{}^2=S_X S_Y$

(狀況二) 當$X=-Y$時，$S_{XY}=\dfrac{\sum\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n}=\dfrac{-\sum\left(X_i-\overline{X}\right)^2}{n}=-S_X{}^2=-S_Y{}^2=-S_X S_Y$

上述兩變數關係狀況，可視為$S_{XY}=r(S_X S_Y)$的兩個特例。

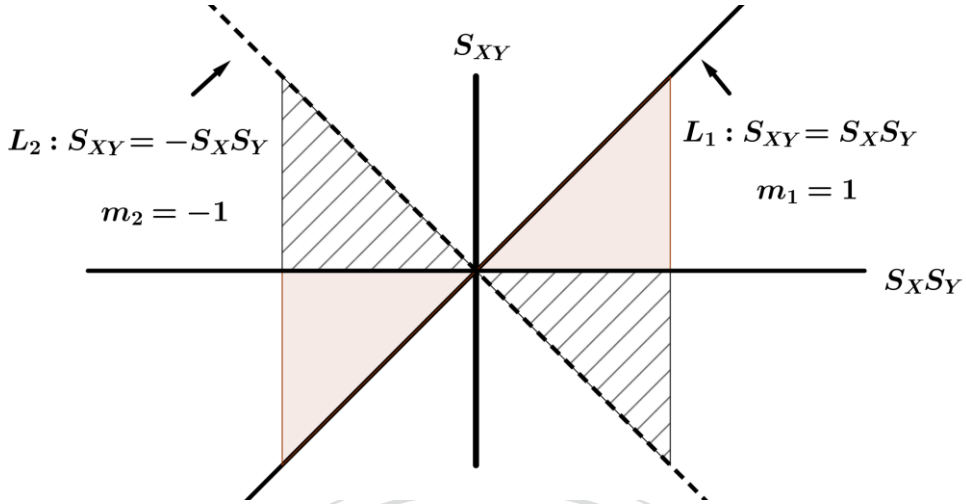


圖 4.2-1

如圖 4.2-1 所示，直線$S_{XY} = r(S_X S_Y)$分別為對應於斜率$m = \pm 1$的線型函數。將狀況一、二代入(3-1)式，恰可得$r = \pm 1$，此值恰為相關係數(r)值的極大值與極小值。因此可知兩變數相關性最大時，資料落在L_1或L_2直線上。

第五章 相關係數與標準化迴歸直線的斜率關係

在計算相關係數 r 時，亦可利用迴歸直線的斜率求得。實際上，若將標準化後的資料進行迴歸分析，所得到的迴歸直線的斜率即為相關係數。關係式如下：

$$\hat{Z}_Y = r \bullet Z_X$$

其中 Z_Y 、Z_X 分別為 Y 與 X 標準化過後的 Z 分數。

第一節 簡單迴歸分析

簡單線性迴歸(Simple Linear Regression)分析是一種統計方法，主要是用於探討兩個變數間是否存在線性關聯，透過適當數學模型的建立來描述變數之間的關係。

1. 簡單線性迴歸： $E(Y) = \beta_1 X + \beta_0$

假設有 n 對樣本資料 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$；我們希望利用一條直線來描述變數 X 與變數 Y 之間的關係，這樣的一條線我們稱之為簡單線性迴歸線。問題是，我們應該如何取得這條線呢？

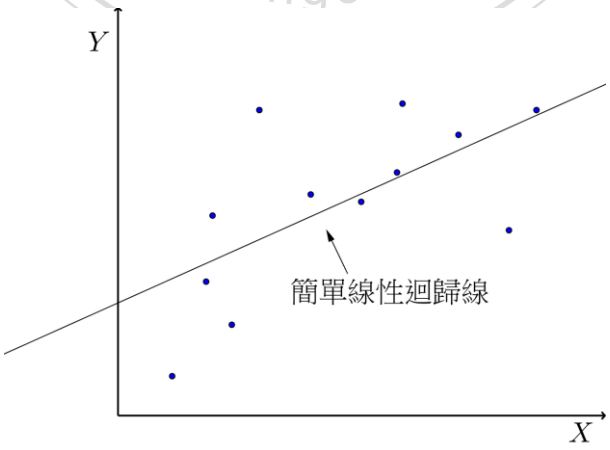


圖 5.1-1 以線性模型來描述變數 X 與變數 Y 之線性關係

2. 最小平方法

當收集到一組資料後，迴歸分析的第一個步驟就是要估計 β_1 與 β_0 。最常用的方法即是最小平方法。所謂的最小平方法就是找使散佈圖中各點至此直線之鉛直距離平方和(即誤差平方和)為最小的那條直線。下圖虛線部分即為誤差。(有關最小平方法視覺化說明請參考附錄。)

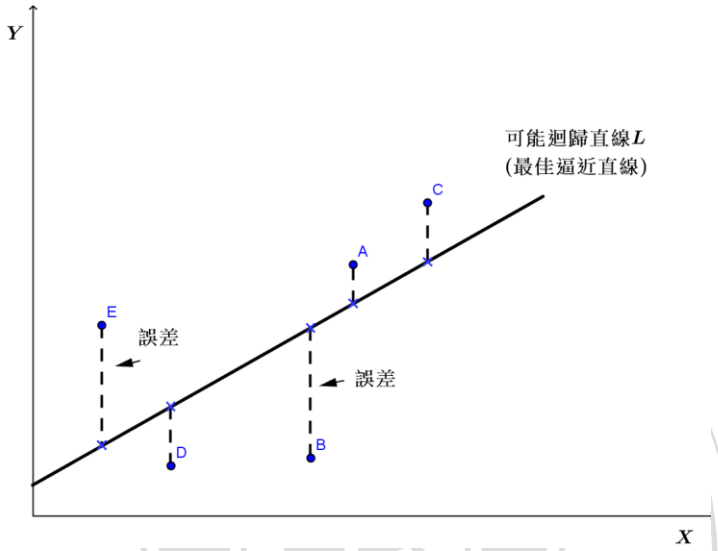


圖 5.1-2

令 $Q=\sum_{i=1}^{n}(Y_i-(\beta_0+\beta_1X_i))^2$ ，分別對 β_0 與 β_1 微分後，並令其值為 0。可得：

$$\frac{\partial Q}{\partial \beta_0}=-2\sum_{i=1}^{n}[Y_i-(\beta_0+\beta_1X_i)]=0$$

$$\frac{\partial Q}{\partial \beta_1}=-2\sum_{i=1}^{n}[Y_i-(\beta_0+\beta_1X_i)]X_i=0$$

亦即 $\sum_{i=1}^{n}Y_i=n\beta_0+\sum_{i=1}^{n}X_i\beta_1$ ； $\sum_{i=1}^{n}X_iY_i=\sum_{i=1}^{n}X_i\beta_0+\sum_{i=1}^{n}X_i^2\beta_1$

解聯立方程式，可得 β_1 與 β_0 兩參數估計式如下：

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} \quad ; \quad \hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}$$

由上式可明顯知道 $\hat{\beta}_1$ 的正負值會跟隨 $\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)$ 而變動。將 $\hat{\beta}_1$ 進一步作推導，可得到和相關係數 r 有下列關係： $\hat{\beta}_1=r\bullet\frac{S_Y}{S_X}$ 。過程如下：

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$$

$$=\underbrace{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}}_{\text{皮爾森積差相關係數}\quad r}\cdot\frac{\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}$$

$$=r\cdot\frac{\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}}}=r\cdot\frac{S_Y}{S_X}$$

明顯的，相關係數 r 與迴歸估計線 $L:\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$ 的斜率有密切關係。

第二節 皮爾森積差相關係數主要是用於直線關係。

將原始資料 X 、Y 標準化為 Z 分數後，標準化過後的新迴歸斜率 $\hat{\beta}$ 即為相關係數 r (黃富廷，民 93)。

1. 迴歸直線 $L:E(Y)=\beta_1X+\beta_0$ ，以離均差($X-\overline{X}$ 或 $Y-\overline{Y}$)來計算迴歸係數時，截距 β_0 將會消失不見。

   令　　$X'=X-\overline{X}$ ； $Y'=Y-\overline{Y}$ 。由於 $\overline{X}'=\frac{1}{n}\sum(X-\overline{X})=0$ ； $\overline{Y}'=\frac{1}{n}\sum(Y-\overline{Y})=0$ ，

   因此　　$\hat{\beta}_0=\overline{Y}'-\hat{\beta}_1\overline{X}'=0$

   從上式可見，截距 β_0 已消失。

2. 標準化資料

   進一步，以 Z 分數來計算迴歸係數時，其迴歸估計式可表示成： $\hat{Z}_Y=\hat{\beta}\bullet Z_X$ 。

因為 $\hat{\beta} = r \bullet \frac{S_{Z_Y}}{S_{Z_X}}$，且 $S_{Z_Y} = S_{Z_X} = 1$，因此 $\hat{\beta} = r$，亦即 $\hat{Z}_Y = r \bullet Z_X$。

故標準化後的新斜率即為相關係數。

第三節 相關係數 r 的絕對值等同兩迴歸直線斜率的幾何平均數

令 $L_1 : \hat{Y} = \hat{\beta}_1 X + \hat{\alpha}_1$、$L_2 : \hat{X} = \hat{\beta}_2 Y + \hat{\alpha}_2$，則 $\hat{\beta}_1 = r \bullet \frac{S_Y}{S_X}$, $\hat{\beta}_2 = r \bullet \frac{S_X}{S_Y}$ 亦即 $r = \hat{\beta}_1 \frac{S_X}{S_Y}$，且 $r = \hat{\beta}_2 \frac{S_Y}{S_X}$。所以 $r^2 = \hat{\beta}_1 \bullet \hat{\beta}_2$，即 $r = \pm\sqrt{\hat{\beta}_1 \bullet \hat{\beta}_2}$。也就是說，$|r| = \sqrt{\hat{\beta}_1 \bullet \hat{\beta}_2}$。

第六章 方向餘弦與相關係數

若將資料先行中心化後，相關係數 r 亦可視為 n 度空間中 X 、Y 兩對應向量的方向餘弦值，亦即：

$$r = \cos\theta$$

其中，θ 為 X 、Y 兩對應向量的夾角。

第一節 向量內積

『向量』是數學、物理和工程等多個自然科學中的基本概念，指的是一個同時具有大小和方向的幾何量。現行高中數學課程中所談的『向量內積』，是將其中一個向量平移，使兩向量的起點相同。如下圖所示，向量 $\vec{v}$ 與向量 $\vec{u}$ 的夾角為 θ。

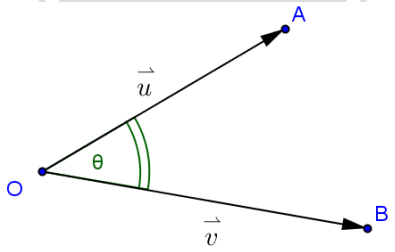


圖 6.1-1 $0^0 \le \theta \le 180^0$

向量 $\vec{v}$ 和 $\vec{u}$ 的內積定義為 $\vec{v}$ 和 $\vec{u}$ 兩向量的大小與其夾角的餘弦函數 $\cos\theta$ 的乘積：

$$\vec{v} \bullet \vec{u} = \left|\vec{v}\right| \bullet \left|\vec{u}\right| \bullet \cos\theta$$

第二節 歐氏空間 $\Re^n$

(X_1,Y_1)、(X_2,Y_2)、…、(X_n,Y_n)為n筆資料；令X分量與Y分量的樣本平均數分別為$\overline{X}$與$\overline{Y}$。

定義向量$\vec{v}$與$\vec{u}$分別代表資料中心化後的向量：

$$\vec{v}=\left(X_1-\overline{X},...,X_n-\overline{X}\right)\quad;\quad \vec{u}=\left(Y_1-\overline{Y},...,Y_n-\overline{Y}\right)$$

向量$\vec{v}$、$\vec{u}$的長度為：

$$\left|\vec{v}\right|=\sqrt{\left(X_1-\overline{X}\right)^2+\cdots+\left(X_n-\overline{X}\right)^2}=\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$$

$$\left|\vec{u}\right|=\sqrt{\left(Y_1-\overline{Y}\right)^2+\cdots+\left(Y_n-\overline{Y}\right)^2}=\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}$$

令兩向量的夾角為θ。因為向量$\vec{v}$和$\vec{u}$的內積為：$\vec{v}\bullet\vec{u}=\left|\vec{v}\right|\bullet\left|\vec{u}\right|\bullet\cos\theta$。

將左式展開得 $\vec{v}\bullet\vec{u}=\left(X_1-\overline{X}\right)\left(Y_1-\overline{Y}\right)+\cdots+\left(X_n-\overline{X}\right)\left(Y_n-\overline{Y}\right)=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$

將右式展開得 $\left|\vec{v}\right|\bullet\left|\vec{u}\right|\bullet\cos\theta=\sqrt{\left(X_1-\overline{X}\right)^2+\cdots+\left(X_n-\overline{X}\right)^2}\bullet\sqrt{\left(Y_1-\overline{Y}\right)^2+\cdots+\left(Y_n-\overline{Y}\right)^2}\bullet\cos\theta$

所以 $$\cos\theta=\frac{\vec{v}\bullet\vec{u}}{\left|\vec{v}\right|\bullet\left|\vec{u}\right|}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\bullet\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}=r$$

$\cos\theta$的值域為$-1\sim1$，即：$-1\le\cos\theta\le1$，這也與相關係數r的值域範圍相同。所以相關係數r也可視為兩變項X、Y方向餘弦值。即：$r=\cos\theta$。

如圖 6.2-1，$\vec{v}$ 為 $\vec{u}$ 在 $\vec{OA}$ 方向的投影分量(正射影)，$\vec{u}$ 與 $\vec{OA}$ 的夾角為 θ，其中 $\vec{OA}$ 上的單位向量為 $\frac{\vec{OA}}{\left|\vec{OA}\right|}$。

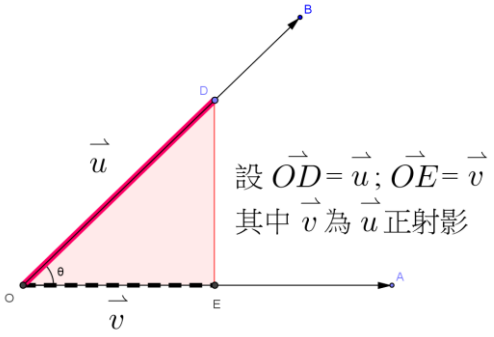


圖 6.2-1

$\vec{u}$ 在 $\vec{OA}$ 方向投影亦可表示為 $\vec{v}=\left(\frac{\vec{u}\bullet\vec{OA}}{\left|\vec{OA}\right|\times\left|\vec{u}\right|}\times\frac{\left|\vec{u}\right|}{\left|\vec{OA}\right|}\right)\bullet\vec{OA}$。

令 $\vec{OA}=X-\overline{X}$，$\vec{OB}=Y-\overline{Y}$，

則 $\vec{OB}$ 在 $\vec{OA}$ 方向的投影又可改寫成下列關係：$Y-\overline{Y}\equiv r\times\frac{S_Y}{S_X}\bullet\left(X-\overline{X}\right)$。

以正射影角度更能明確來看出兩變數間的相關性，這結果正呼應方向餘弦 $\cos\theta$ 的值域變換與兩向量夾角角度 θ 關係。當角度 θ 愈小，當 $\vec{OB}$ 在 $\vec{OA}$ 的投影量愈大，$\vec{OB}$ 與 $\vec{OA}$ 的關係愈密切。此時用 $\vec{OA}$ 去推測 $\vec{OB}$ 的解釋度亦相對提高；反之，當角度 θ 愈大，當 $\vec{OB}$ 在 $\vec{OA}$ 的投影量愈小，$\vec{OB}$ 與 $\vec{OA}$ 的關係將愈疏離。此時用 $\vec{OA}$ 去推測 $\vec{OB}$ 的解釋度相對降低。換句話說，兩變項的相關性高低會跟隨夾角的大小決定。所以以向量的觀點看相關相關係數，相關係數 r 可視為中心化資料的向量方向餘弦值。即，相關係數 r 具有向量方向餘弦性質(陳順宇、鄭碧娥 民 87)。

第七章 兩個標準化迴歸直線的夾角與相關係數關係

藉由兩條標準化迴歸直線相交所成的角度大小，亦可用來表示變數間相關程度的強弱，亦即相關係數 r 亦可藉由下面方式取得：

$$r = \sec\alpha \pm \tan\alpha$$

其中 α 為兩標準化迴歸直線的銳角夾角。

第一節 相關係數不因觀測變數角度不同改變

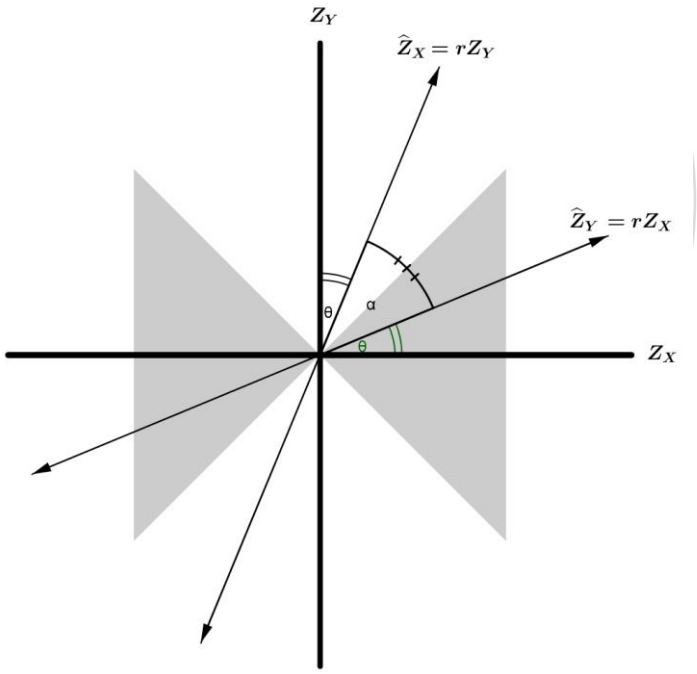


圖 7.1-1 兩標準化迴歸直線和兩軸夾角相等

觀察圖 7.1-1：將資料標準化後，不管是直線 $\hat{Z}_Y = r \bullet Z_X$ 或 $\hat{Z}_X = r \bullet Z_Y$，顯示出 $\hat{Z}_Y = r \bullet Z_X$ 與 X 軸、$\hat{Z}_X = r \bullet Z_Y$ 與 Y 軸的夾角 θ 相同。所以不管是由 X 的觀點來看 Y，或是由 Y 的觀點來看 X，相關性強弱不會因為端看資料的角度不同而有所改變。

第二節　兩直線銳角夾角與皮爾森相關係數 r 的關係

觀察圖 7.2-1、7.2-2，可發現兩迴歸直線間的銳角夾角 α 大小和兩變數間的相關強弱具奇妙的關聯性。即當 α 愈小， L_1 與 L_2 越靠近。當 α 愈大， L_1 與 L_2 疏遠。所以，我們可藉由兩迴歸直線相交時銳角夾角角度數的大小，來看兩變數間的相關性強弱。

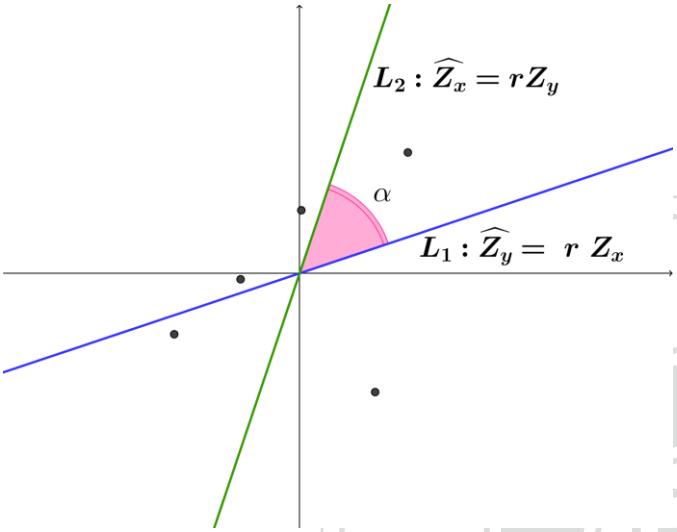


圖 7.2-1

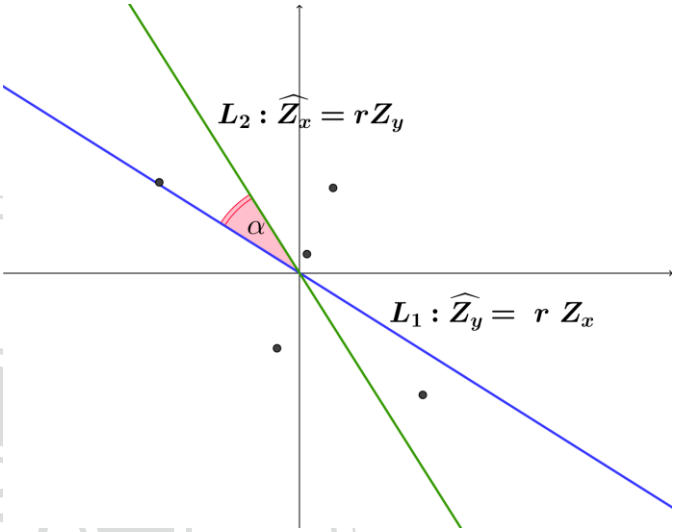


圖 7.2-2

當 L_1 與 L_2 距離愈近，即相關係數 $|r|$ 越大； L_1 與 L_2 距離愈遠，兩變數的相關性相對的也就越小，相關係數 $|r|$ 也越小。

第三節　公式推導

觀看右圖 7.3-1。兩標準化迴歸直線 L_1 、 L_2 與 Z_x 和 Z_y 兩軸的夾角 θ ，兩直線的銳角夾角為 α 。

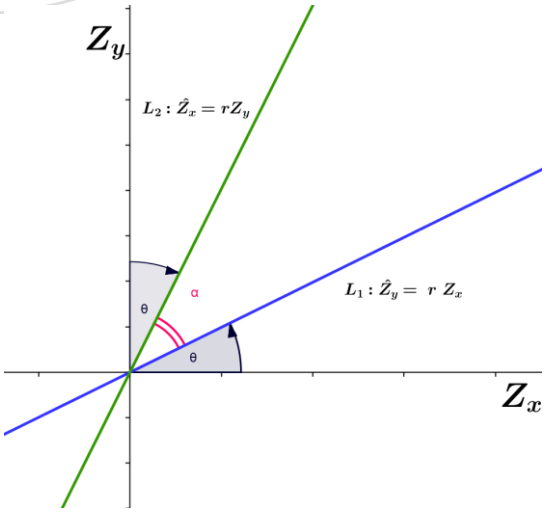


圖 7.3-1

考慮θ在下面兩種狀況：

1. 當$0 \le \theta \le \frac{\pi}{4}$時，因　　$L_1 : Z_y = rZ_x$

   所以正切函數$\tan\theta = r$，且$\tan 2\theta = \frac{2r}{1-r^2}$，

   因此$\tan\alpha = \tan\left(\frac{\pi}{2} - 2\theta\right) = \cot 2\theta = \frac{1-r^2}{2r}$，且　$\sec\alpha = \frac{1+r^2}{2r}$

   由於$\sec\alpha - \tan\alpha = \frac{1+r^2}{2r} - \frac{1-r^2}{2r} = r$

   所以$r = \sec\alpha - \tan\alpha$

2. 當$-\frac{\pi}{4} \le \theta \le 0$，因　　$\alpha = \frac{\pi}{2} + 2\theta$，所以

   $\tan\alpha = \tan\left(\frac{\pi}{2} + 2\theta\right) = -\cot 2\theta = -\frac{1-r^2}{2r} = \left(\frac{1-r^2}{-2r}\right)$

   且　$\sec\alpha = \sec\left(\frac{\pi}{2} + 2\theta\right) = -\csc 2\theta = -\left(\frac{1+r^2}{-2r}\right) = \frac{1+r^2}{2r}$

   將上述兩式相加可得　$\sec\alpha + \tan\alpha = \frac{1+r^2}{2r} + \left(\frac{1-r^2}{-2r}\right) = r$

   所以$r = \sec\alpha + \tan\alpha$

根據 1、2 結果，在看兩變數間的相關性強弱，我們亦可藉由兩迴歸直線銳角夾角度數的大小變換來看變數間的相關性；即，若兩迴歸直線交角α，其中$0^0 \le \alpha \le 90^0$，則α和相關係數r之間的關係式為：$r = \sec\alpha \pm \tan\alpha$。

考慮$r = \sec\alpha - \tan\alpha$，將此關係式改寫成$r = (1-\sin\alpha)/\cos\alpha$；當$\alpha$趨近於$0^0$，$r$的值將愈接近1；這代表兩迴歸直線將愈靠近角平分線$L$(圖 7.3-2)。而當$\alpha$趨近於$90^0$，$r$的值會越接近於零，表示$X$的變動與$Y$的變動之間並無關係。

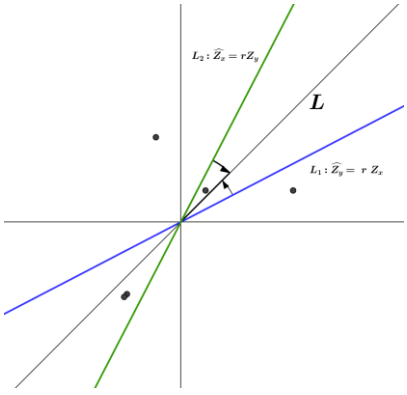


圖 7.3-2

第八章 判定係數與相關係數

在簡單迴歸中，相關係數 r 和判定係數 R^2 的關係為：

$$r = R\text{的正或負平方根}$$

所以，相關係數 r 亦可藉由計算判定係數 R^2 的方式取得。

第一節 判定係數 R^2

相關分析是利用相關係數 r 來衡量兩變數 X 、Y 之間的直線關係強度與相關方向；而迴歸分析是根據依變數 Y 與自變數 X 的關係，求出一個迴歸模型，再利用此迴歸模型，用自變數 X 去預測依變數 Y 。當依變數 Y 與自變數 X 之間的關係可以用一迴歸模型來解釋時，模型解釋能力的程度大小，或者迴歸方程式的配適度如何，是藉由判定係數 R^2 來作描述。

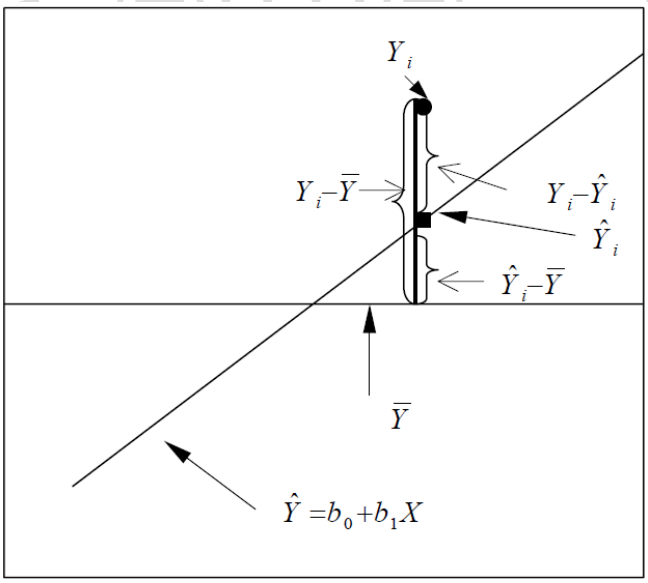


圖 8.1-1 迴歸分析中總變異量成份的解析

因為 $Y_i-\overline{Y}=\left(\hat{Y}_i-\overline{Y}\right)+\left(Y_i-\hat{Y}_i\right)$，且 $\sum\left(Y_i-\overline{Y}\right)^2=\sum\left(\hat{Y}_i-\overline{Y}\right)^2+\sum\left(Y_i-\hat{Y}_i\right)^2$

其中

$\sum\left(Y_i-\overline{Y}\right)^2=$總變異$=sum\ \ of\ \ squares\ \ due\ \ to\ \ total\ \ (SSTO)$

$\sum\left(\hat{Y}_i-\overline{Y}\right)^2=$可解釋變異$=sum\ \ of\ \ squares\ \ due\ \ to\ \ regression(SSR)$

$\sum\left(Y_i-\hat{Y}_i\right)^2=$不可解釋變異$=sum\ \ of\ \ squares\ \ due\ \ to\ \ error\ \ (SSE)$

所以 $SSTO=SSR+SSE$ 。

判定係數 R^2 是依變數 Y 的變異中可以被自變數 X所解釋的比例。定義為：

$$R^2=\frac{SSR}{SSTO}=1-\frac{SSE}{SSTO}$$

判定係數 R^2 之值介於 0 與 1 之間。若將判定係數 R^2 以百分比表示時，R^2 可視為總變異可用估計迴歸方程式解釋的程度，也就是 Y 可以被 X 解釋的程度。判定係數 R^2 愈高，代表自變數 X 越能解釋依變數 Y (估計線性迴歸方程式配適度愈好。)。

第二節 判定係數 R^2 ＝相關係數 r^2

變異的計算公式如下： $SSTO=\sum Y_i^2-n\overline{Y}^2$ ； $SSR=\frac{\left[\sum X_iY_i-n\overline{X}\overline{Y}\right]^2}{\sum X_i{}^2-n\overline{X}^2}$

又因為 $SSR=SSTO-SSE$ ，因此

$$R^2=\frac{SSR}{SSTO}$$

$$=\frac{\left[\sum X_iY_i-n\overline{X}\overline{Y}\right]^2\Big/\sum\left(X_i-\overline{X}\right)^2}{\sum\left(Y_i^2-n\overline{Y}^2\right)}=\frac{\left(\frac{\sum X_iY_i-n\overline{X}\overline{Y}}{n}\right)^2}{\frac{\sum\left(Y_i-\overline{Y}\right)^2}{n}\cdot\frac{\sum\left(X_i-\overline{X}\right)^2}{n}}=\frac{S_{XY}^2}{S_X^2S_Y^2}=r_{XY}^2=r^2$$

所以，判定係數 R^2 ＝相關係數 r^2 。兩個變數共用的特徵越多，它們就越相關。

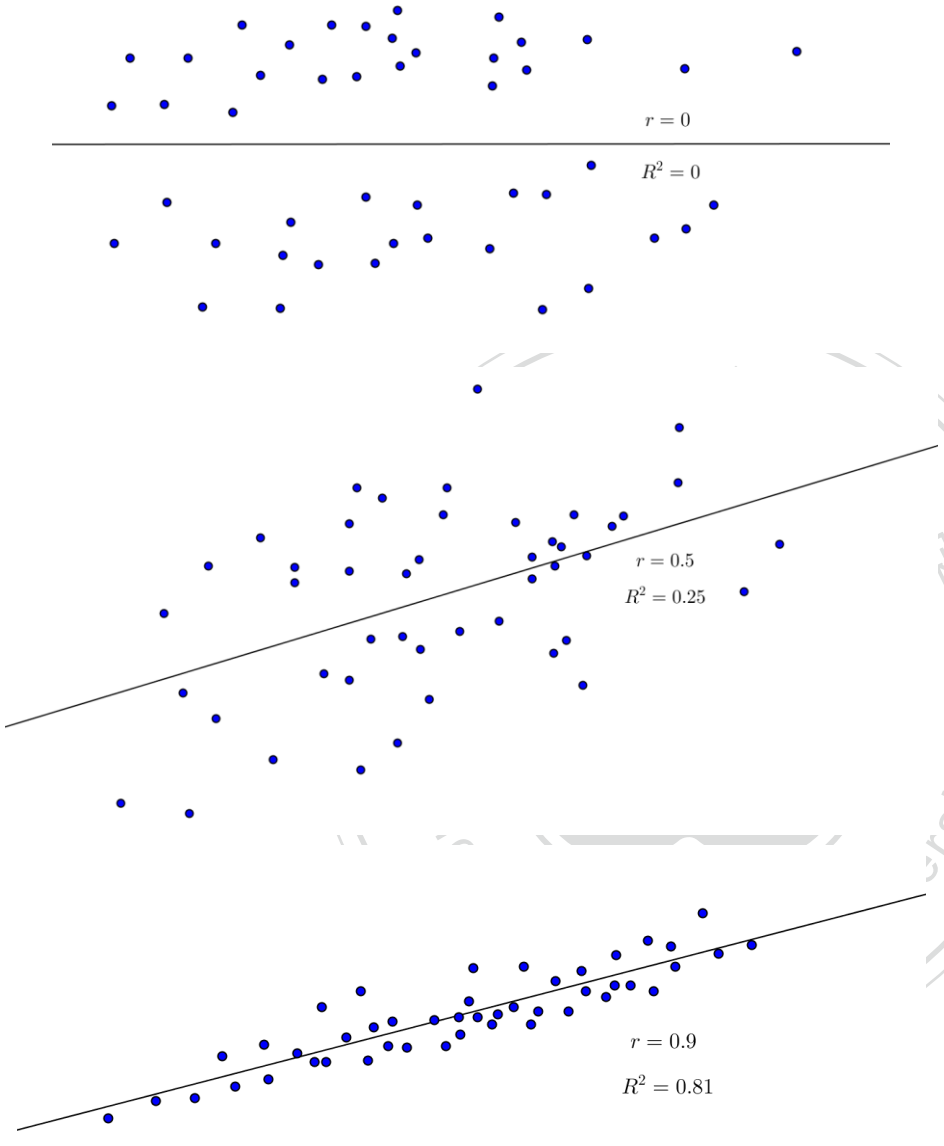
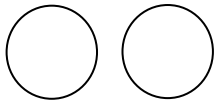
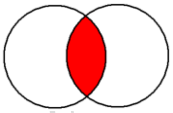
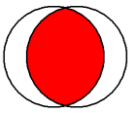
判定係數 $R^2 = r^2$ 相關係數	變數 X 和變數 Y 關係
	 共用 0%  共用 25%  共用 81%

表 8.2-1

表 8.2-1，例舉判定係數 R^2 與相關係數 r 的對應關係。資料點散佈愈緊密於迴歸模型，則模型解釋能力的程度愈高。再將自變數 X 解釋依變數 Y 的比例與相關係數值的大小關係以圖文氏圖形圖表示；每個陰影區域越大(兩個變數共用變異量就越大)，這兩個變數就越高度相關。

第九章 結論與建議

相關係數的說明煩瑣，光是解讀就煞費苦心；尤其在教學時，面對統計觀念尚未熟稔的新生，尤其棘手。本文所談到的不同向度相關係數解釋，希望有助於教師在從事教學活動時，對於相關係數的介紹和使用，能夠更多樣化。

如在標準化系統下，以 S_X 當作 X 座標的單位長， S_Y 當作 Y 座標的單位長，定義相關係數 $r=\frac{\sum Z_X Z_Y}{n}$ ，是比較直觀、易懂、方便教學的作法(第三章)。而在原始資料中，若將變數 X 、 Y 以向量的觀點來看(θ 為 X 、 Y 對應向量的夾角)，則 $r=\cos\theta$ ，原相關係數定義 $r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}$ 中， $-1\le r\le 1$ 之所以成立，也就變得平易近人，學生也較容易接受和記得。

在推導出迴歸直線的係數時，可以正射影作為輔助，用簡單的直線方程式 $\hat{Z}_Y=rZ_X$ ，來聯想 $\frac{\hat{Y}-\overline{Y}}{S_Y}=r\left(\frac{X-\overline{X}}{S_X}\right)$ ，推得 Y 對 X 的迴歸直線 $L:\hat{Y}=\overline{Y}-r\frac{S_Y}{S_X}\overline{X}+r\frac{S_Y}{S_X}X$ ，就容易許多。而且很明顯的可以看出這方程式具由 X 估計 Y 的功能。學生這樣也會比容易接受理解 Y 對 X 的最適合直線涵義。(李政豐,民 99)

電腦與通訊科技的突飛猛進，在邁入網路學習的現今，傳統的教育學習方式，已無法抵擋網路社會的巨大變化。借用資訊科技，利用電腦快速處理大量圖形及超強計算能力，我們可以動態方式呈現各種軌跡圖形、抽象概念以具體圖像來呈現等。透過 GGB、GSP、Minitab、EXCEL 等數學軟體將相關係數相關議題視覺化，將傳統繁瑣的徒手演算，改以直接報導答數和進行結論說明，學生將更容易明白所學為何；學習態度亦會變成主動且熱衷。

參考文獻

英文部分

Rodgrs and Nicewander.(1998). “Thirteen Ways to Look at the Correlation Coefficient.” *The American Statistician*, 42, 59-66.

GeoGebra (GGB) 。 http://www.geogebra.org/cms/

中文部分

黃富廷。(2004)。皮爾遜積差相關之數學原理：線性代數觀點。台東特教第 19 期。

李政豐(2010)。視覺化的相關係數最小平方法與迴歸直線。教育部高中數學學科中心資訊融入教學工作坊。

陳順宇、鄭碧娥 (1998)。統計學，台北市：華泰書局。

附錄

相關名詞解釋

1. 相關係數另外有下列幾種計算相關係數的方法：

(1). 等級相關係數(Rank correlation coefficient)

有時給出的變數值不方便、不經濟，甚至不可能，只能給出變數的等級。在這種情況下，必須利用等級相關係數。等級相關係數可能也是用於變數之間出現非線性關係的情況。

(2). 肯道耳等級相關係數(Kendall's coefficient of rank correlation)

這種相關係數記作 τ ，可利用 n 對數據通過下面公式計算： $\tau = \dfrac{S}{\frac{1}{2}n(n-1)}$

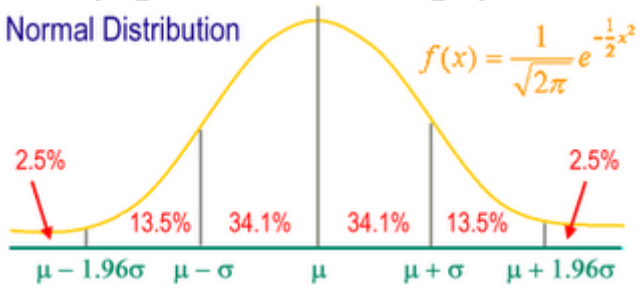
(3). 史匹曼等級相關係數(Spearman's coefficient of rank correlation)

這種相關係數記作 ρ ，其計算公式如下： $\rho = 1 - \dfrac{6\sum D^2}{n(n^2-1)}$

2. 資料中心化

(1). 常態分佈

又稱高斯分佈(Gauss Distribution)。是統計學中最重要的分配之一，其圖形呈鐘形，稱為常態曲線。一般研究變數常會呈現常態分佈或近似常態分佈，如身高、體重、收入、支出、意見程度、評量誤差(error of measurement)。如下圖所示：



中心點位置其數值出現的頻率(次數)最多，離中心點位置左右(可延伸到無窮大±∞)的數值出現頻率漸少，曲線左右對稱，即大於平均值和小於平均值的出現頻率相等。統計學上所謂的標準常態分布是指將觀測資料標準化，使其平均數為 $\mu = 0$ ，標準差為 $\sigma = 1$ 。

(2). 中央極限定理

當樣本越大時，樣本平均值的分布越接近常態分布，且向平均值 μ 集中。以下網址提

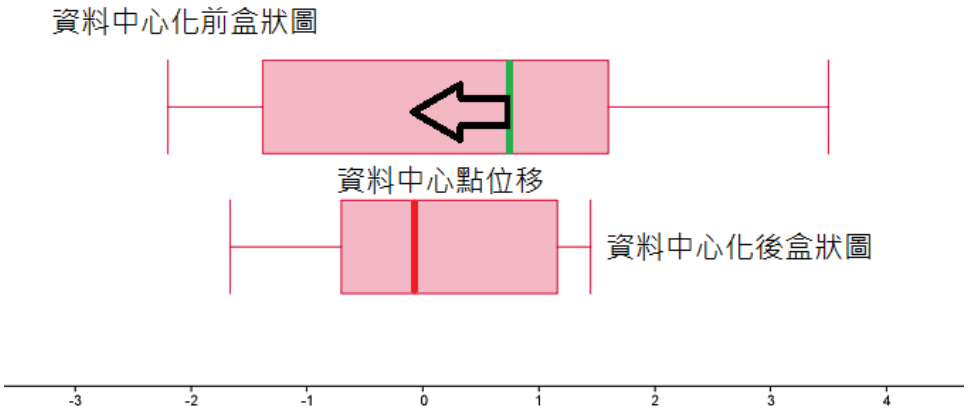
供網路互動式模擬程式，可從互動的實驗中理解中央極限定理的基本概念。

參考網址:

http://www.math.nsysu.edu.tw/StatDemo/CentralLimitTheorem/CentralLimit.html#five

(3). 資料中心化

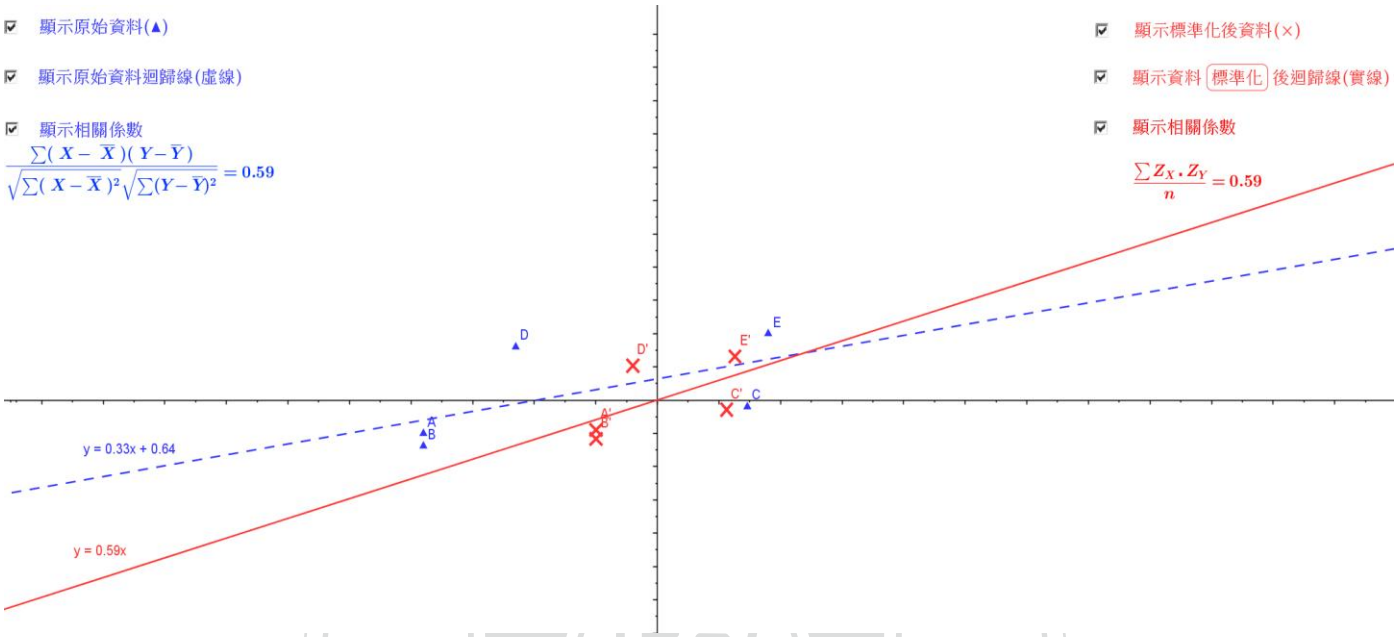
所謂的『資料中心化』，是將觀測資料知原始平均數朝原點來移動。即是將原始資料的平均數化為0 。資料經過中心化後，任何數值等於離均差；任何數值的絕對值義等於離均差長度。



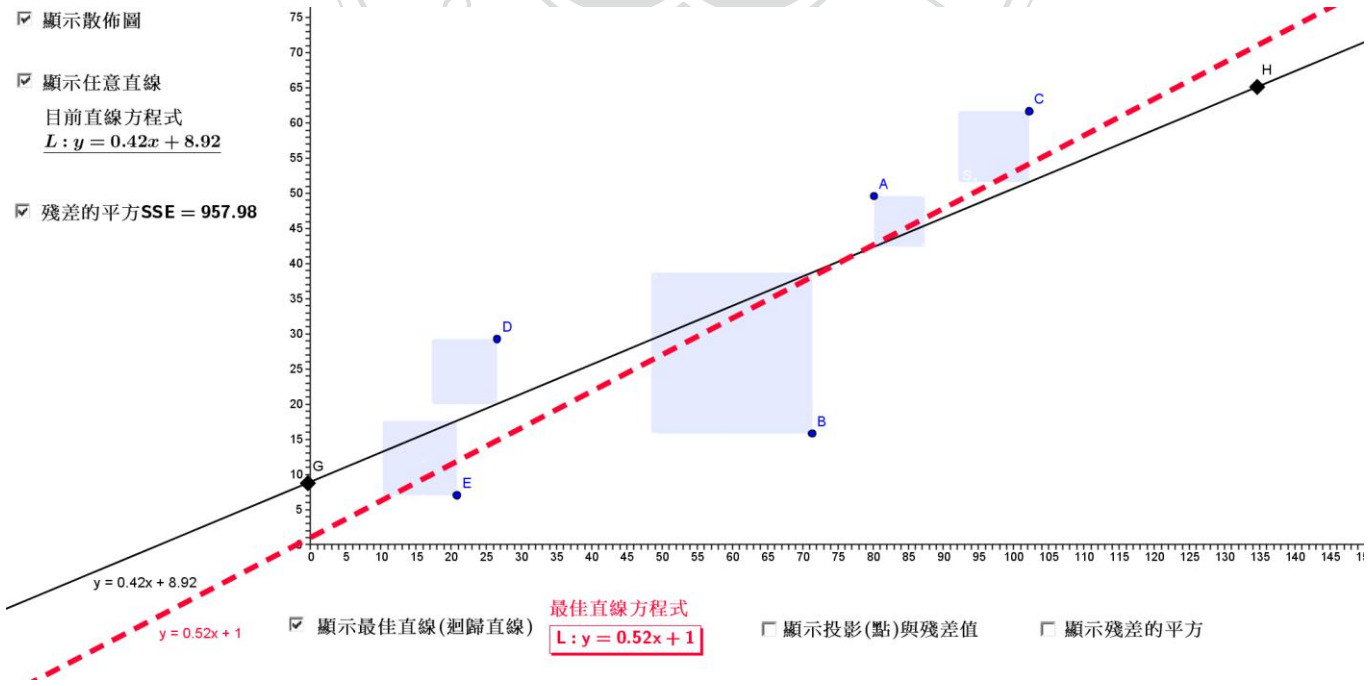
可參考『第三章 資料標準化線性關係不變』。

圖檔

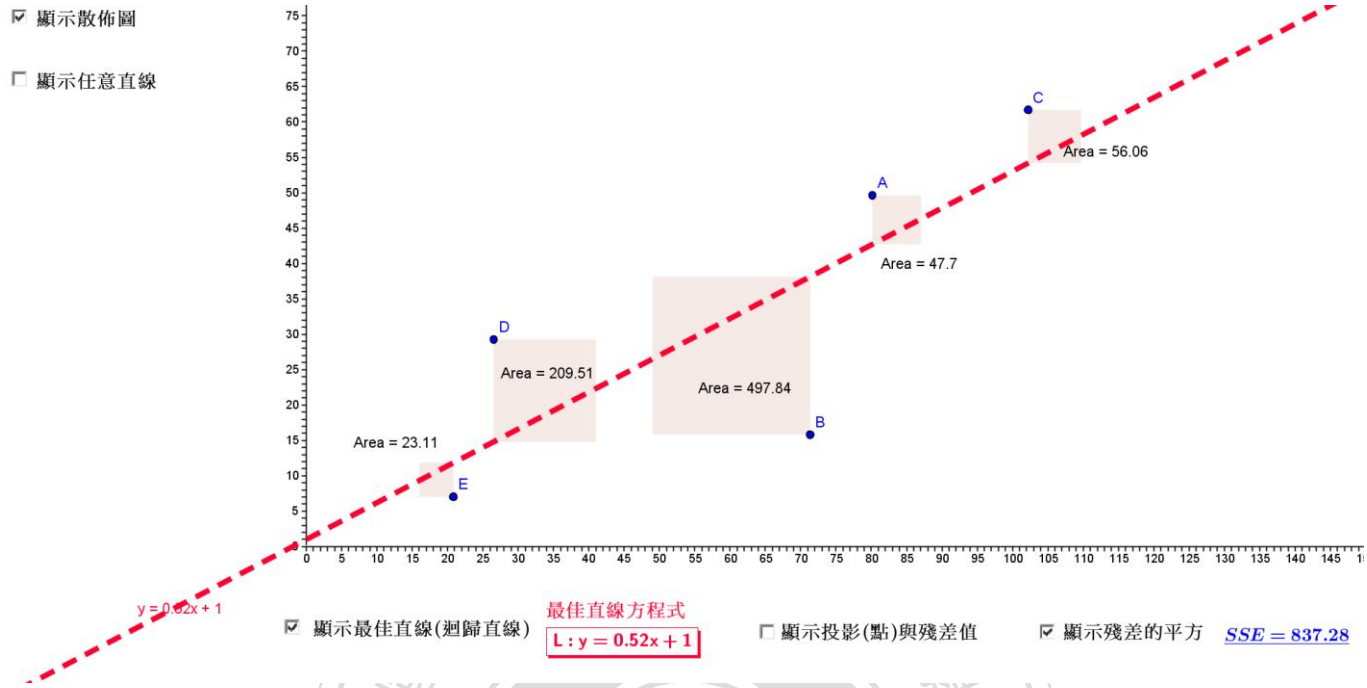
1. 資料標準化後，相關性不變；相關係數等於內積的平均值



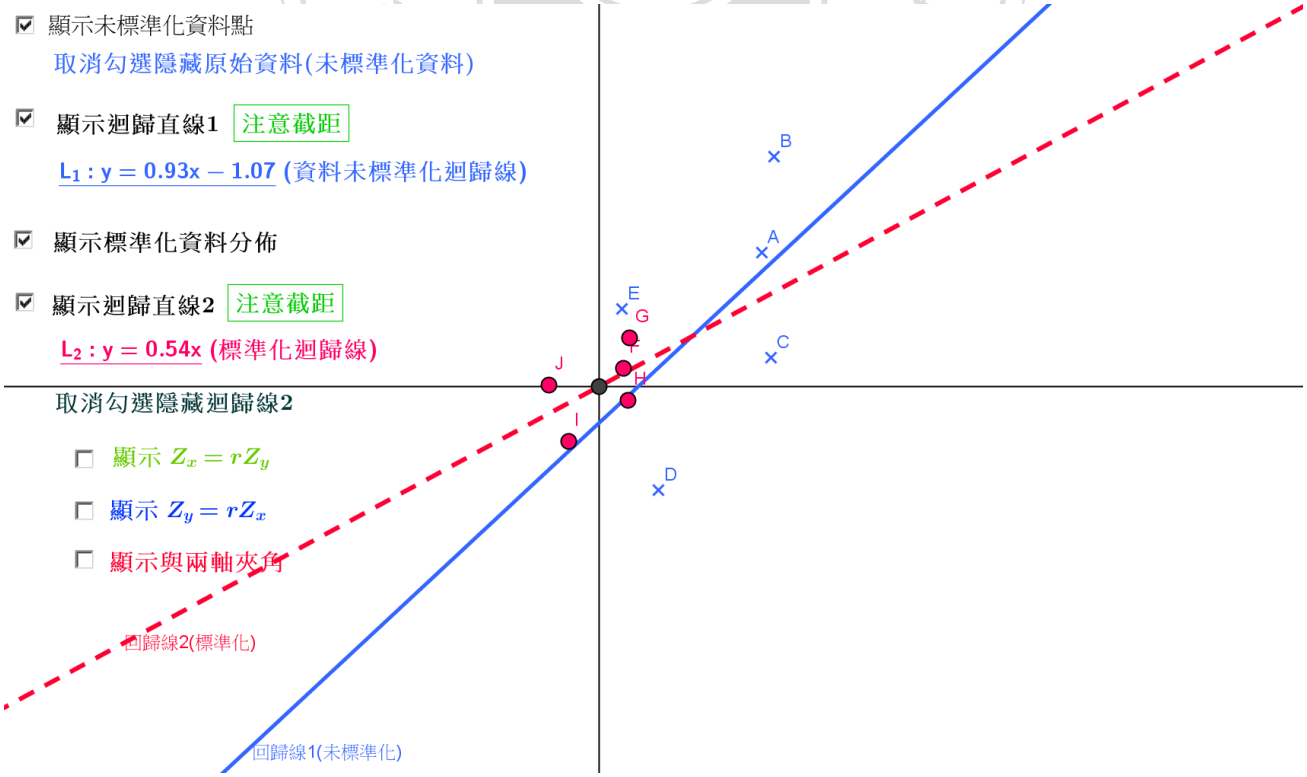
2. 最佳直線(1)



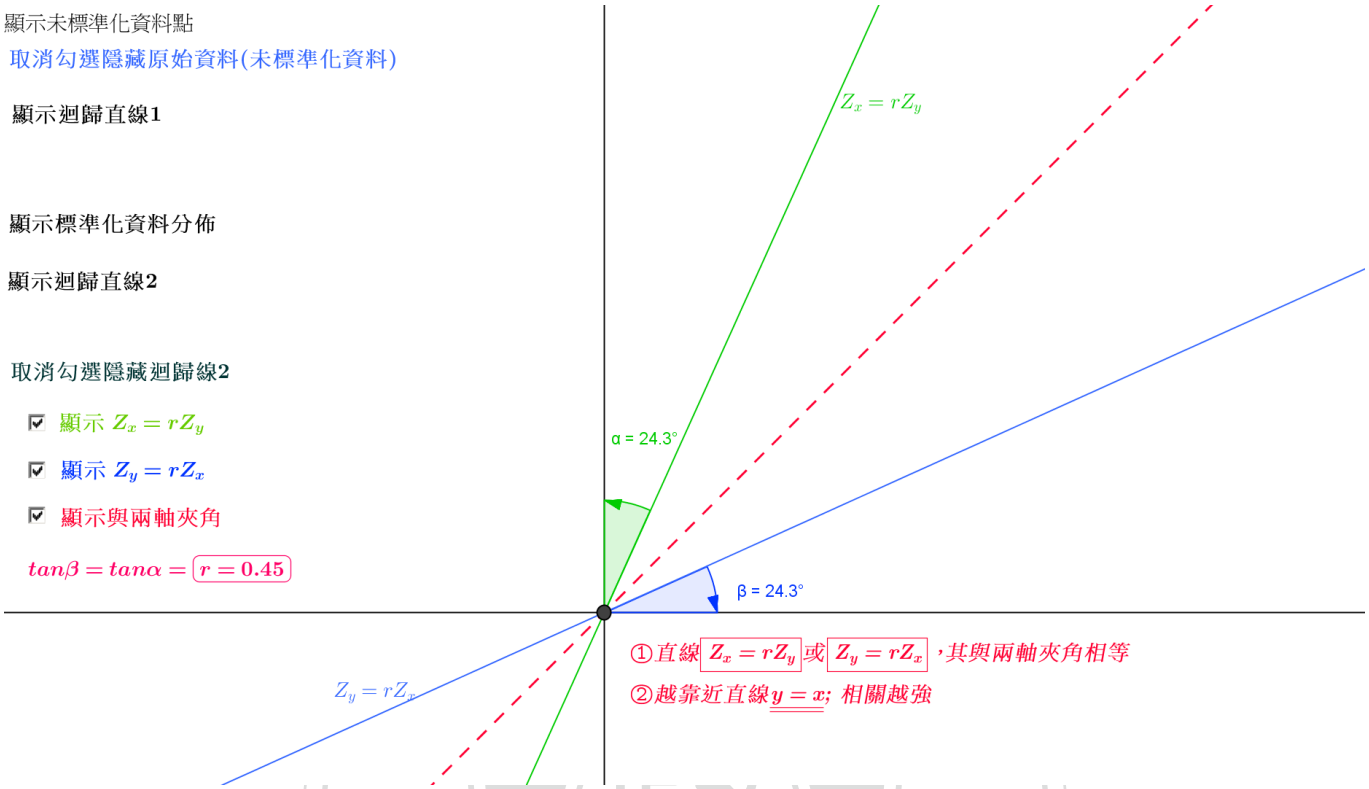
3. 最佳直線(2)



4. 標準化迴歸直線



5. 兩標準化迴歸直線和兩軸夾角相等



6. 正射影