

迴歸 (Regression)

筆記參考資料：

1. Regression (迴歸) Dr. Tun Wen Pai

線性回歸

線性迴歸簡介

一種統計學上分析資料的方法，目的在於了解多個獨立變數與一個應變數的關係。

通常來說，迴歸存在的意義，是造出一條曲線盡可能地滿足這些資料，以達到預測、了解關係的目的。

我們可以使用相關係數來比較這個造出來的曲線的好與壞。

相關係數

相關係數 r ，用來評估兩個變數之間的關係是否相關，其定義如下：

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

定義 \bar{X} 、 \bar{Y} 為變數 X 、變數 Y 的平均值。

其中 $-1 \leq r \leq 1$ ，其中若 $r = 0$ 時則代表完全無相關， $|r| = 1$ 時則代表完全相關， $0 < |r| < 1$ 時則代表存在一定的線性相關。

決定係數

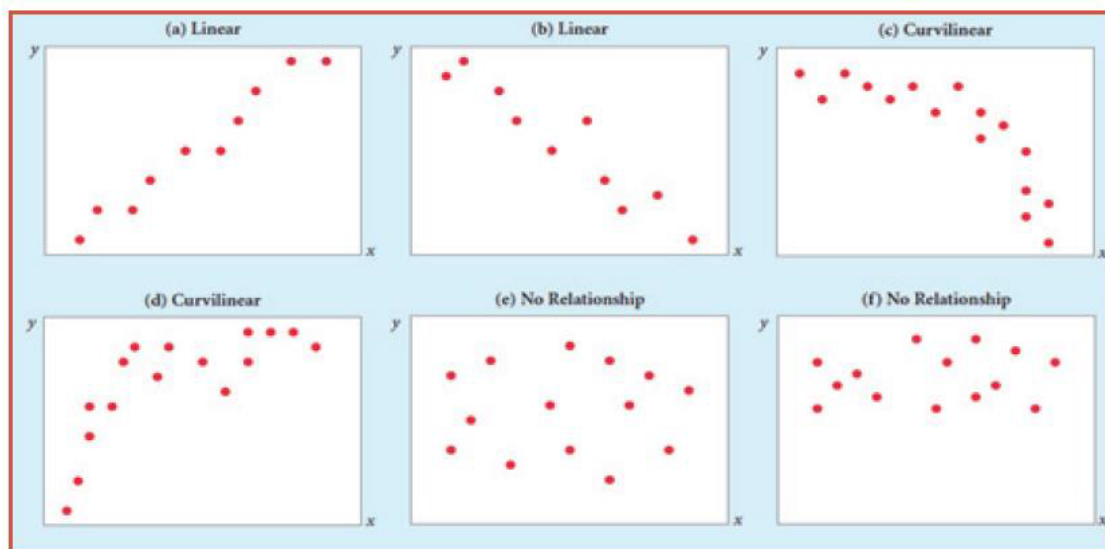
Reference : <https://www.youtube.com/watch?v=2AQKmw14mHM>

決定係數 r^2 ，用來判斷回歸模型的解釋力，可以將相關係數平方，得到決定係數。

決定係數可以更好的幫助我們判斷兩個變數之間的關係，可以知道選擇的兩個變數能夠解釋多少比例的資料變異。

雙變數的圖形呈現

對於一個雙變數的資料集，我們可以畫出一張二維的散佈圖，可透過散佈圖來觀察出變數之間的關係。



▲ 圖 7.1：顯示兩變數間各種關係類型的散佈圖（來源：Groebner 等人，2013）

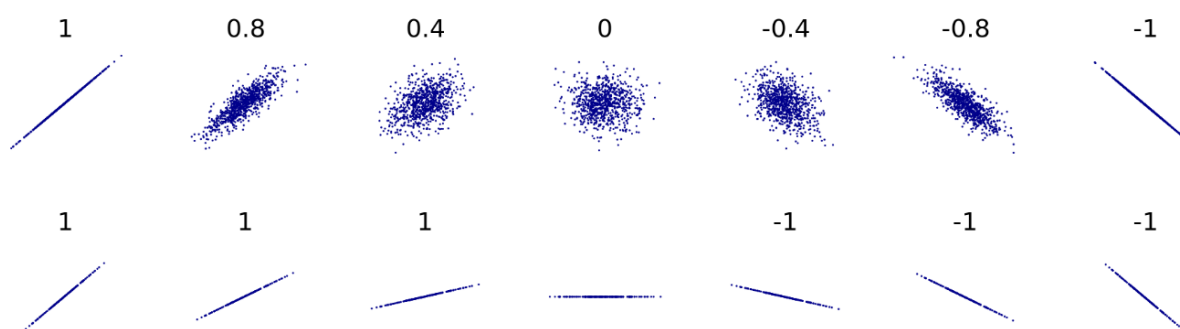
可以發現，(a) 與 (b) 的圖形可以畫上一條直線，資料大部分都在這條線附近，所以是線性相關。

(c) 與 (d) 可以畫上一條曲線，資料大部分都在這條曲線附近，所以是曲線相關。

(e) 與 (f) 找不到直線、曲線能夠含括大部分的資料，故為無相關。

結合前一個小節所講的相關係數，我們可以從散佈圖上來找出與相關係數的關係，可以發現資料越散，取絕對值後的相關係數越小。

資料越集中於一條線，取絕對值後相關係數的係數越大。



線性迴歸的推導

待補

邏輯回歸

待補