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：

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

定義 $\overline{X}$、$\overline{Y}$ 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

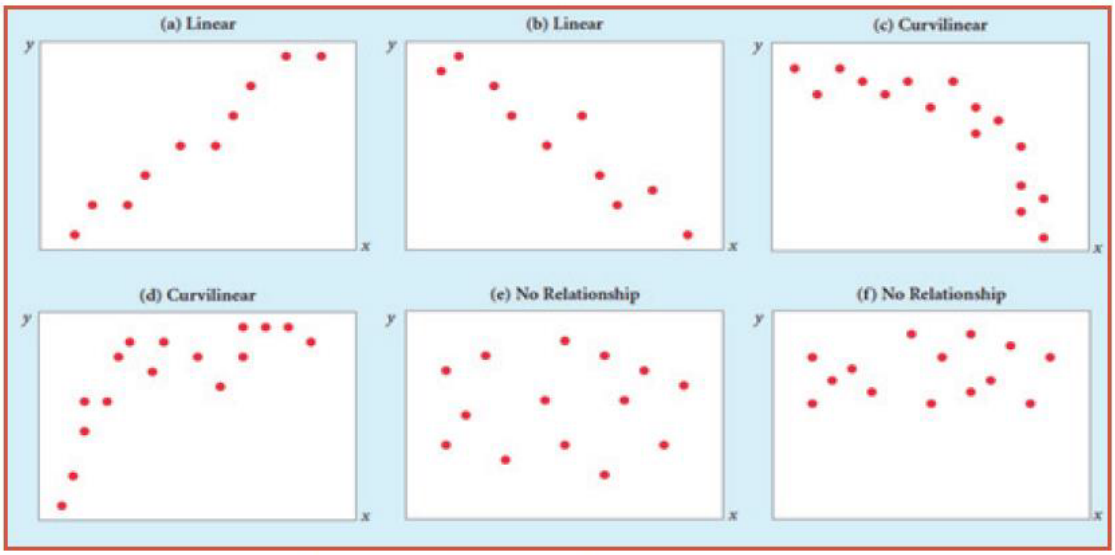
> Reference：https://www.youtube.com/watch?v=2AQKmw14mHM

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）

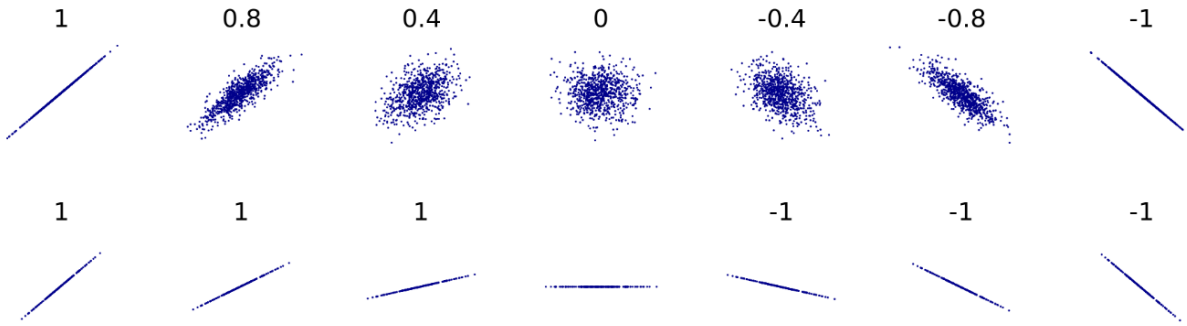
可以發現，(a) 與 (b) 的圖形可以畫上一條直線，資料大部分都在這條線附近，所以是線性相關。

(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。

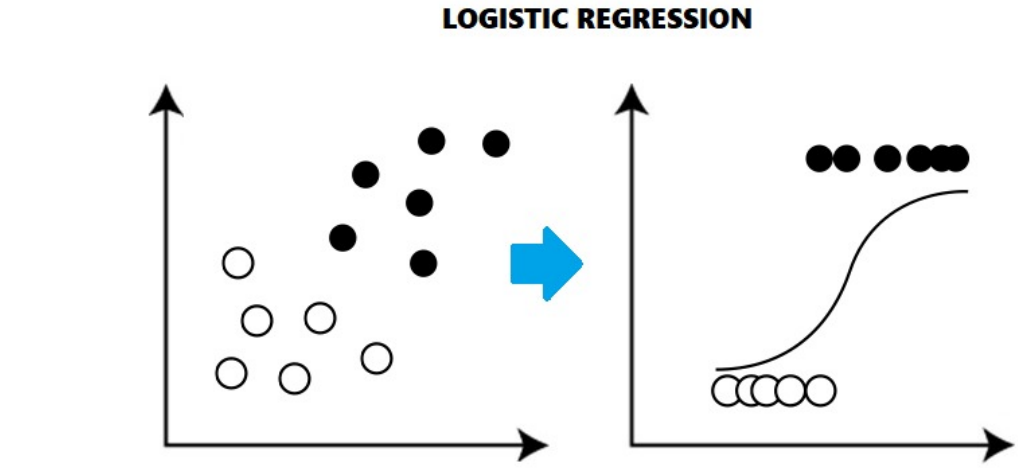
結合前一個小節所講的相關係數，我們可以從散佈圖上來找出與相關係數的關係，可以發現資料越散，取絕對值後的相關係數越小。

資料越集中於一條線，取絕對值後相關係數的係數越大。



羅吉斯迴歸

羅吉斯迴歸



> Image from equiskill

羅吉斯迴歸（Logistic Regression），是一種**對數分類器**而不是一種迴歸模型。

而羅吉斯迴歸又分成二元、多項式羅吉斯迴歸等等。

這邊主要介紹二元羅吉斯迴歸，也就是對於一個變量，評估結果是否發生。

羅吉斯迴歸的圖形

二元羅吉斯迴歸可以用一張二維平面圖來作圖。

其中，x 軸是一種變量，而 y 軸則是對於導致這個結果產生的機率，機率介於 0 到 1 之間。

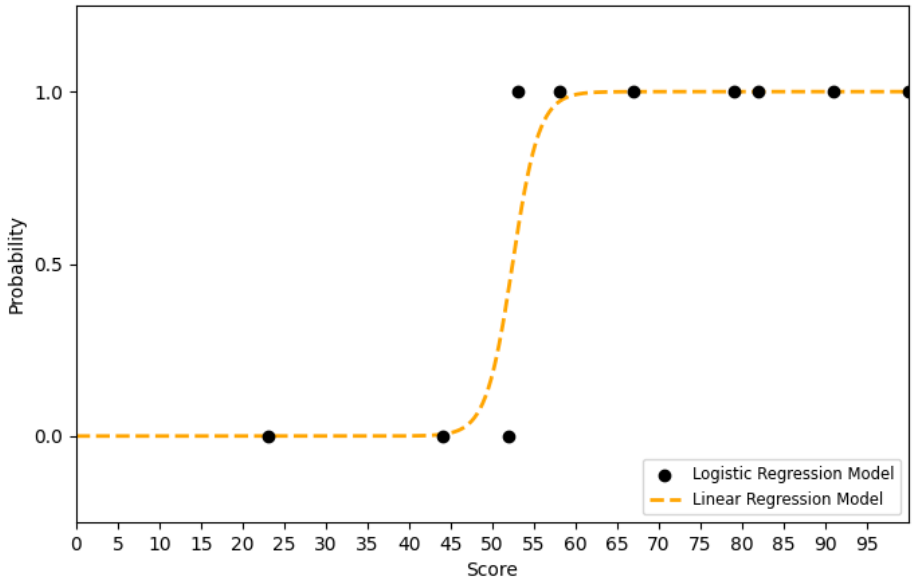
圖上的點密集分布在 $y=0$ 與 $y=1$ 上，代表著對應的變量是否發生結果。

而圖上的的線即為線性迴歸模型，我們可以根據這個曲線對應到的變量，來預測對於一件事情的發生機率為何。

舉個例子，我們有一組資料為學生期中成績與通過課程與否的資料。

期中考成績	44	53	23	67	79	91	100	58	82	52
通過	0	1	0	1	1	1	1	1	1	0

我們可以根據這個資料，做一張 Logistic Regression 的圖。



> Image from me with matplotlib and scikit-learn

圖中的橘線即為擬合後的線性迴歸模型，而點則是期中考成績對應到的通過與否。

勝算

我們定義勝算 (odds) 為，一件事情發生的機率，除以一件事情不發生的機率，也就是 $\frac{p}{1-p}$

$\log(odds)$ 可以有效的讓勝算的起點設定在數線的 0 上，並且隨著數值的大與小，顯示在數線上距離 0 有多遠。

羅吉斯迴歸的圖形轉換

上面的圖似乎可以非常棒的看出對應的成績所造成的通過與否了，不過我們可以將這個圖形做一點轉換。

我們考慮將這條線性迴歸模型拉直，也就是轉成 $y = ax + b$ 的形式。

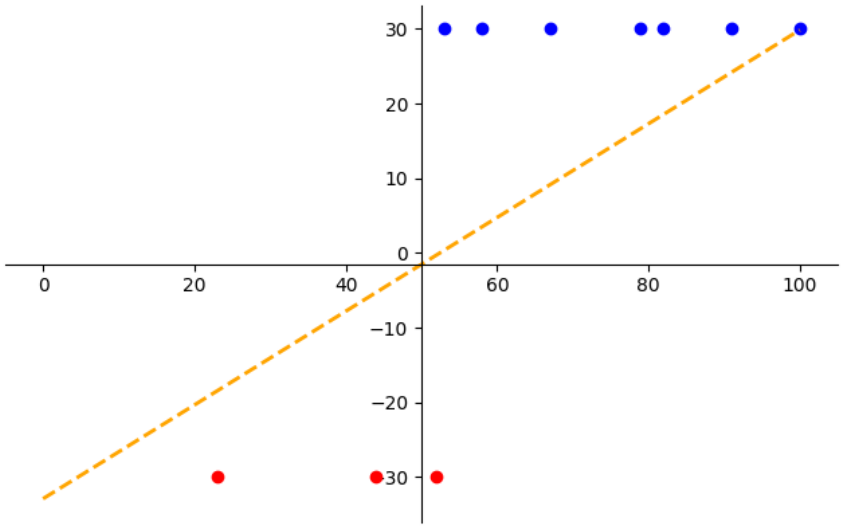
為此，我們可以考慮將 y 軸每一點的機率值，映射到一個定義域為 $(-\infty, \infty)$ 的 y 軸上

我們可以使用 Logit Function，定義為 $\ln(odds) = \ln(\frac{p}{1-p})$ ，來對 y 軸進行映射，

再套用訓練出來的參數，將這個羅吉斯迴歸的圖形造成一條直線。

例如以上圖來說，可以將上圖轉成以下的平面。

請注意：圖形上的點應在 $y = -\infty$ 或 $y = \infty$，因坐標軸無法顯示，故以最大值表示。



> Image from me with matplotlib and scikit-learn