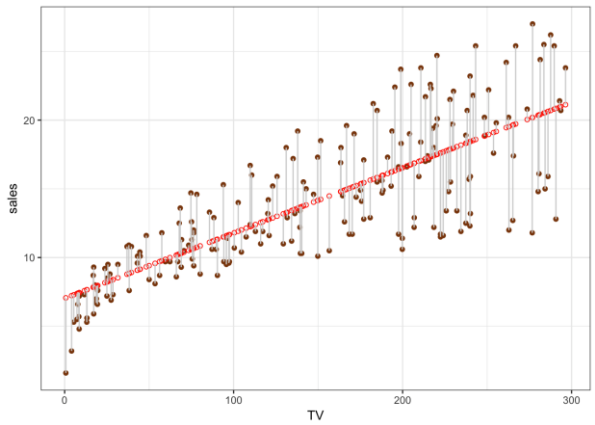
迴歸分析---最小平方法

前面已提及，Galton 把身高跑回平均數的傾向稱為「迴歸」，亦即迴歸於平均數之意。而「散佈圖」可以呈現變數之間的關聯、「相關係數」可以度量直線關聯的強度和方向。如果散佈圖呈現出很強的直線關聯，是不是可以用直尺隨手畫出一條直線來描述，這條線就是「迴歸直線」，表示此線的方程式就稱為「迴歸方程式」。我們可以用迴歸直線來描述當解釋變數 x 變化時，反應變數 y 會怎麼跟著變。迴歸和相關係數最大的不同，就是迴歸必需區分何者為解釋變數、何者為反應變數。Galton 發現在許多自然現象和社會現象的演變中，都有向平均數迴歸的時性。從此以後，即使所研究的現象不一定含有向平均數迴歸的原意，在統計分析中對於估計線和估計方程式的統計技術均稱為「迴歸分析」。

通常，一條直線最容易適配(Fit)，而且在一定範圍內直線常能把一組資料配合得相常良好，因此直線應用最廣。當然如果不太講究，研究人員可以直接在散佈圖上「隨手描繪」(Free-hand Method) 一條近似的直線即可，講究一點拿個直尺來畫會更直，這樣雖可以很容易的在圖上得到直線，但是沒有直線的方程式，而且也沒有辦法保證這條就是「最好」的直線。

統計技巧可以根據資料找到最好的迴歸直線方程式，這當中最常用到的技巧就是用「最小平方法」(Least Square Method) 找出一條最佳的「最小平方迴歸線」(Least Square Line)！所謂的最小是指所有的資料點距離直線的鉛垂距離平方和為最小的那一條直線，如下圖。



幾乎所有的統計軟體及某些計算機都可以幫助我們很容易的得到最小平方

迴歸直線方程式，重點是我們要怎麼樣理解及應用！以下是直線方程式的通式，x 代表解釋變數，y 代表反應變數：

$$y = a + bx$$

b 是直線的斜率(slope)，就是 x 增加一個單位時 y 的改變量。 a 是截距(intercept)，是當 x = 0 時 y 的值。要利用這個方程式做預測時，只要我們把 x 代入方程式，計算出 y 值即可。使用電腦作計算再方便不過，再大筆的資料也沒問題，可是電腦不會自己決定哪一個是解釋變數，哪一個是反應變數，偏偏這是很重要的，因為變換了 x，y ，算出來的迴歸方程式是不一樣的。還有一點要注意的是：最小平方迴歸直線會受到少數極端值的嚴重影響，所以我們一定要事先選好解釋變數、反應變數，並事先檢視散佈圖確實有直線的型態並排除特別的離群值。

還有，雖然迴歸方程式可以用解釋變數預測反應變數，但並不是表示其間就一定有因果關係，例如計算美國股票指數與台股指數的迴歸方程式，然後我們可以用美股漲跌來預測台股漲跌，至於二者是否有因果關係，好像有影響又不必然，你說呢？

適配直線對預測有多大作用是和相關強度有關的，而相關係數 r 就是直線關聯的一種量度。x 和 y 之間的關係只能解釋 y 的部份變異，也就是當 x 移動時，y 值沿著迴歸直線跟著移動的部份，剩下的則是資料點距直線的距離。相關係數的平方(即 r^2)剛好是 y 值的總變異當中，可以用 y 對 x 的最小平方迴歸來解釋的部分所占的比例。

我們在作迴歸分析時，通常也會將 r^2 的值一併列出，作為用迴歸直線來解釋反應變數程度的指標。當我們看到相關係數時，也應該關心相關係數平方，以掌握關聯的強度。相關係數 r = 0.7 差不多是介於 r = 0 和 r = 1 的中間(不是 r = 0.5)，因為 $(0.7)^2 = 0.49$。一般來說，在自然科學以外的許多領域，高相關係數都是不常見的！

當沒有任何一個單一解釋變數和反應變數有高相關時，我們也可以同時用好幾個解釋變數來預測反應變數，這就是多元迴歸(multiple regression)。多元迴歸下的相關係數稱為「複相關係數」(multiple correlation coefficient)，代號同樣也是用 r，複相關係數的計算相當複雜，好在電腦可以代勞，且複相關係數的平方 r^2 也和一個解釋變數時意義相同。例如美國法學院入學測驗的成績和入學後的成績之間的相關係數大約 r = 0.36，也就是說只能解釋 13%(因為 $r^2 = 0.1296$)，可是如果再加上大學成績，複相關的相關係數就可提高到 0.45，入學

後成績的解釋度就可提高到約 20%(因為 $r^2 = 0.2025$)，雖然還是不理想。

利用迴歸方程式來預測不能太過自信，尤其是變數很多的多元迴歸，因為解釋變數之間常有強的關聯，造成複雜的關係和陷阱，經由實驗來作解釋還是比較恰當的方法。