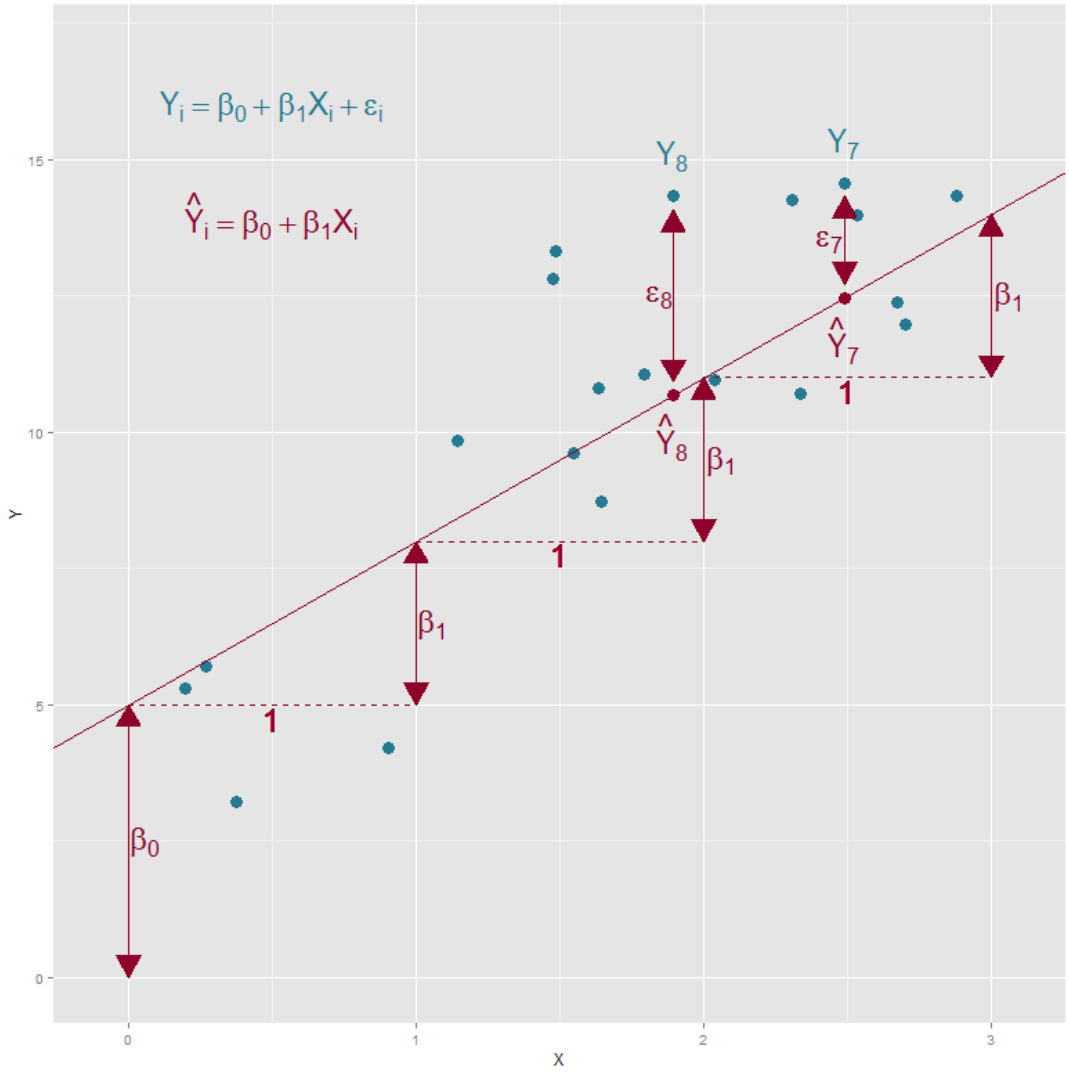
R-web 資料分析應用：迴歸分析

李智慎 副統計分析師



上圖為一模擬資料，可以很直覺的看出 X 與 Y 有正相關，且每個資料點都落在直線附近，在日常生活中許多事物彼此間常常存在著線性關係，如要將變數與變數之間的關係以具體的式子表達，其中一個簡單且常用的方法就是利用簡單線性迴歸模型來分析，兩變項 X 與 Y 關係可表示成$Y = \beta_0 + \beta_1 X$，其中 Y、X 分別稱為依變數(dependent variable)與自變數(independent variable)，由此式子模型可以很明確的從截距項β_0和係數β_1得

知自變數改變時對依變數的影響，當自變數增加 1 單位，依變數則增加β_1單位，但現實生活中實際例子幾乎不存在這種完美的線性關係，會有各種其他因素造成誤差存在，因此會在模型中加入一個隨機誤差項ε_i來表示，完整的簡單線性迴歸模型可表示為

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

隨機誤差項ε_i在統計學上假設為常態分配，且平均數為0變異數為σ^2，可寫成$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$，iid為independent and identically distributed的縮寫，表示誤差項ϵ_i彼此互相獨立且相同分配。迴歸分析中限制依變數需為連續型變數而自變數則無限制連續離散皆可，假如想建造離散型依變數的迴歸模型則可用邏輯斯迴歸，這部份我們將會在下期作介紹。而本期同樣統一使用源自基隆社區為基礎的整合篩選計畫(Keelung Community-based Integrated Screen Program, KCIS)的心血管疾病資料作範例資料檔，有關此資料的詳細資訊及變數定義請參閱首期生統eNews。

➢ 迴歸模型係數的估計-最小平方法

我們知道迴歸模型為一種表示自變數與依變數之間關係的方式，但迴歸係數通常都是未知的，我們該設定係數為多少才是一個好的迴歸模型呢?最簡的的方法就是最小平方法，其精神在於讓迴歸模型的誤差項平方合能最小，即最小化物差平方總合$\sum_{i=1}^{n}(Y_i - \beta_0 + \beta_1 X_i)^2$，可利用微分的方式進而求得預估值$\hat{\beta}_0 = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma_{i=1}^{n}(X_i - \bar{X})^2}$、$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_1 \bar{X}$，而此時誤差項變異數的估計值為$\hat{\sigma}^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$，其中$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$利用到前面所得的係數估計

值$\hat{\beta}$做計算。

而模型適不適用與迴歸係數為不為 0 也有關係，因此可以做迴歸係數是否為 0 的檢定，即虛無假設為$H_0: \beta_1 = 0$，對應的檢定統計量為$t = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$，而$s.e(\hat{\beta}_1) = \hat{\sigma}\sqrt{1/\Sigma_{i=1}^{n}(X_i - \bar{X})}$。當虛無假設為真時，此統計量服從自由度為 n-2 的 t 分佈。

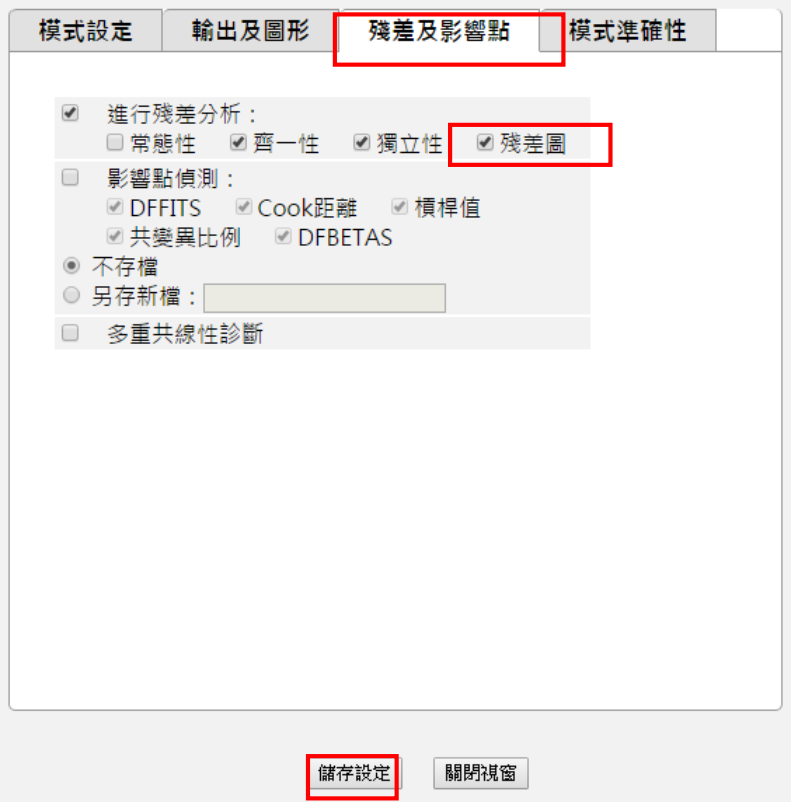
➢ R Web-迴歸分析操作步驟

在初階使用者的模式下，從 R-web 主選單中依序點選【分析方法】→【迴歸模式】→【迴歸分析】即可進入分析頁面。

我們以 CVD 資料”SysBP”（心臟收縮壓）為依變數，”Age”（年齡）為自變數作分析，R-Web 操作圖解如下

第一步，先選擇要進行分析的資料檔，點選”使用者個人資料檔”後選擇”CVD”的檔案（心血管疾病資料），系統將自動帶出參數設定畫面。在步驟二選擇要進行分析的變數，在此設定依變數為”SysBP”（心臟收縮壓）、

自變數為”Age”（年齡）。最後點選【進階選項】將出現選如下圖，



點選上方”殘差及影響點”標籤後勾選”殘差圖”再點擊【儲存設定】，完成設定後即可點擊步驟二下方【開始分析】。

- **迴歸係數估計[I]：**

係數 coefficient	估計值 estimation	標準差 std. err.	t檢定統計量 t-statistic	p值[II] p-value	參數的 95% 信賴區間 95% C.I. for estimations	
					下界 lower	上界 upper
(截距項)	93.7881	0.2639	355.3254	< 2.22e-16 ***	93.2708	94.3054
Age	0.6298	0.0054	116.4961	< 2.22e-16 ***	0.6192	0.6404

I：依變數為SysBP，模式包含常數項

II：顯著性代碼： ‘***’ : < 0.001, ‘**’ : < 0.01, ‘*’ : < 0.05, ‘#’ : < 0.1

上圖為分析結果其中的迴歸係數估計表，當顯著水準為 0.05 時使用信賴區間法或 p 值法得到雙尾檢定結果顯示”SysBP”（心臟收縮壓）與”Age”（年齡）有顯著的相關，且年齡每增加一歲，收縮壓會增加 0.6298，

也就是迴歸模型係數β_1的估計值，而截距項β_0的估計值為 93.7881，帶入係數估計值後我們可得到迴歸預估模型如下

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad \Longrightarrow \quad \hat{Y}_i = 93.7881 + 0.6298 X_i$$

有了此預估模型則可以用來預測依變數，例如有一人的年齡為 27 歲，則套入此預估模型可估計此人心臟收縮壓平均測量值為 110.7927。簡單線性迴歸是直接假設依變數和自變數為線性關係，再對迴歸係數作是否為 0 的檢定，但事實上依變數與自變數之間可能不為線性關係，有各種不同的方法可以做檢測，而簡單線性迴歸模型中最常被使用的是判定係數(coefficient of determination，R^2)，又稱 R 平方值，從定義上來說，R^2可以表示自變數能解釋多少比例的依變數變異，數值會介於 0~1 之間，愈接近 1 代表此模型愈能解釋依變數的變化，其等式為

$$R^2 = \frac{\Sigma_i^n \left(\hat{Y}_i - \bar{Y}\right)^2}{\Sigma_i^n (Y_i - \bar{Y})^2}$$

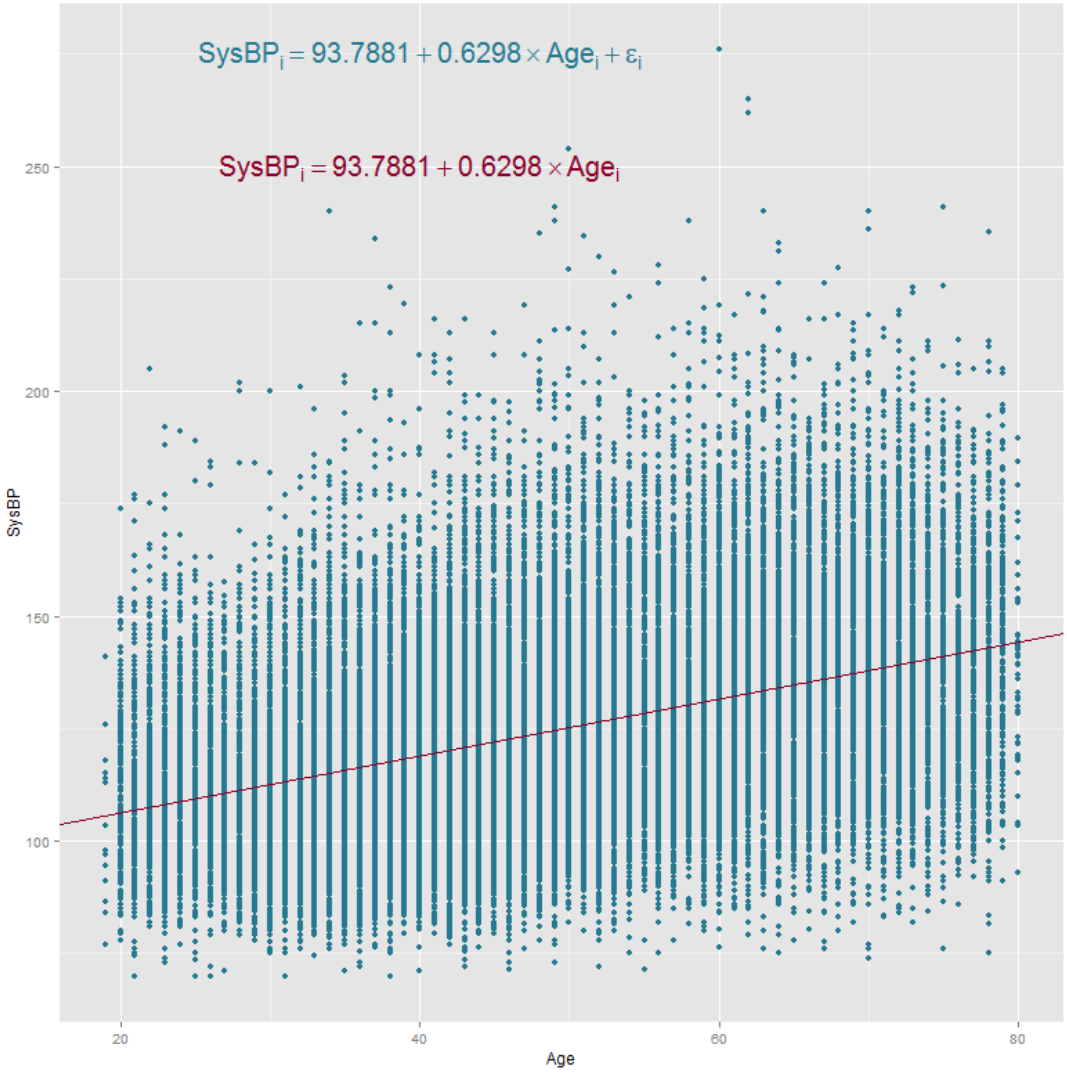
其值可由分析結果"迴歸模式的變異數分析"中得知，如下圖所示

- **迴歸模式的變異數分析：**

虛無假設：迴歸模式不顯著						
來源 source	平方和 sum of squares	自由度 d.f.	均方和 mean square	F檢定統計量 F-statistic	臨界值 F(d.f.1,d.f.2,1-α)	p-值[I] p-value
迴歸 regression	4844368.0094	1	4844368.0094	13571.3512	3.8416	< 1e-04 ***
誤差 error	22577076.3777	63249	356.9555			
總和 total	27421444.3871	63250				
判定係數(R-square)：17.67 %						
調整判定係數(adjusted R-square)：17.67 %						

I：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

由結果可得知所得到的迴歸預估模型判定係數為 0.1767，表示使用此預估模型自變數對於解釋依變數變異的能力不是很好。



上圖為”Age”（年齡）與”SysBP”（心臟收縮壓）的散佈圖，藍色點代表各個實際資料點，而紅色線為依照迴歸預估模型$\hat{Y}_i = 93.7881 + 0.6298X_i$所得的迴歸線，可看出年齡與心臟收縮有線性關係但並不非常的明顯，且資料分佈的位置並沒有明顯向迴歸線集中，與R^2值 0.1767 相符合。

因為在此預估迴歸模型下，自變數 Age 並不能充分解釋依變數 SysBP 的變異，且並無非常明顯的線性關係，建議可以換個變數試試，以下我們選擇”DiaBP”（心臟舒張壓）為自變數且重複與之前同樣的步驟，可得到

新的迴歸分析結果如下

- **迴歸模式的變異數分析：**

虛無假設：迴歸模式不顯著						
來源 source	平方和 sum of squares	自由度 d.f.	均方和 mean square	F檢定統計量 F-statistic	臨界值 F(d.f.1,d.f.2,1-α)	p-值[I] p-value
迴歸 regression	15045376.0369	1	15045376.0369	77879.2002	3.8416	< 1e-04 ***
誤差 error	12210101.011	63203	193.1886			
總和 total	27255477.0478	63204				
判定係數(R-square)：55.2 %						
調整判定係數(adjusted R-square)：55.2 %						

I：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

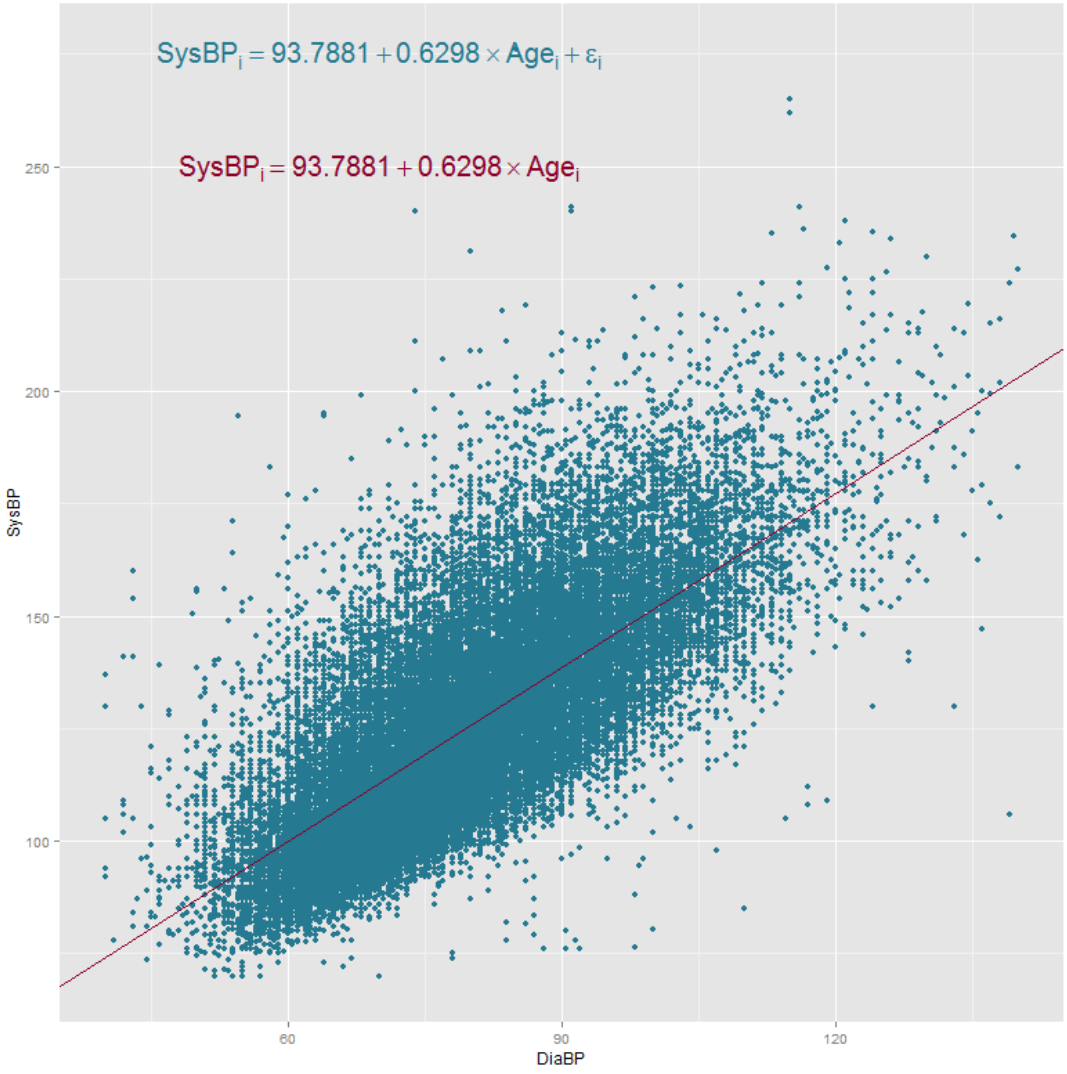
- **迴歸係數估計[I]：**

係數 coefficient	估計值 estimation	標準差 std. err.	t檢定統計量 t-statistic	p值[II] p-value	參數的 95% 信賴區間 95% C.I. for estimations	
					下界 lower	上界 upper
(截距項)	22.5819	0.3649	61.885	< 2.22e-16 ***	21.8667	23.2971
DiaBP	1.2892	0.0046	279.0685	< 2.22e-16 ***	1.2801	1.2982

I：依變數為SysBP，模式包含常數項

II：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

由上圖可得依變數”SysBP”（心臟收縮壓）與自變數”DiaBP”（心臟舒張壓）的簡單線性迴歸結果，顯著水準為 0.05 時，雙尾檢定結果顯示有顯著的關係，且判定係數R^2為 0.552 與自變數為”Age”（年齡）時的迴歸模型比較起來，”DiaBP”（心臟舒張壓）更能解釋依變數”SysBP”（心臟收縮壓）的變異，下圖為依變數”SysBP”（心臟收縮壓）與自變數”DiaBP”（心臟舒張壓）的散佈圖



由此圖可發現兩者有較為明顯的線性關係，且資料也較向迴歸線集中，因此可判斷依變數”SysBP”（心臟收縮壓）與自變數”DiaBP”（心臟舒張壓）有更高度的線性相關。

本期生統 eNews 的介紹到此告一段落，此次介紹了迴歸分析中的簡單線性迴歸方法，希望大家能從圖表中更能理解簡單線性迴歸模型與資料間的關係。下一期的生統 eNews 將為大家介紹迴歸模式的分析方法—『邏輯斯迴歸分析』，敬請期待！