

107相關與回歸分析研習會



相關與回歸之正確使用

呂秀英 場長

iying@mdais.gov.tw

苗栗區農業改良場

內容

1. 基本概念
2. 相關分析
 - 2.1 簡單相關係數
 - 2.2 相關的濫用及誤用
 - 2.3 相關分析結果發表的正確呈現
3. 回歸分析
 - 3.1 簡單(直線)回歸
 - 3.2 曲線(非線性)回歸
 - 3.3 複回歸
 - 3.4 逐步回歸
 - 3.5 回歸的濫用及誤用
 - 3.6 回歸分析結果發表的正確呈現

1.基本概念

資料如何獲得？

❖ 觀測(調查)研究

- 只在不干擾情形下觀察或測量目標對象的特性(稱為變數)以蒐集資訊，如問卷調查、倉庫害蟲調查等。
- 觀測研究之目的，是用來描述一個團體或一種情況。
- 取樣調查是觀測研究中很重要的一種。

❖ 試(實)驗

- 會對某個體作某件事情(處理)，然後觀測個體如何反應，如施以不同藥劑對種子發芽影響、不同飼料對牛體重影響等。
- 實驗之目的，通常要了解某種處理是否會確實對所測量性狀(即變數)引起某種反應。
- 試驗設計重於分析。

農業試驗研究資料分析之主要目的

❖ 比較某變數在不同處理間的平均值差異性

⇒ 變方分析、多重比較等

❖ 探討兩變數間的關聯性

⇒ 相關分析等 相關係數 $r=0.89^{**}$

❖ 探求反應變數受到影響因子的影響程度

⇒ 回歸分析等

回歸式 $Y=2.2 + 5.5 X$, $R^2=0.84$, $\text{Prob}=0.001$

變方分析表(ANOVA)

Source	DF	MS
Treatment	3	1350 **
Error	8	56

處理平均值多重比較表

T1	T2	T3	T4
3.5 b	4.2 b	9.0 a	1.5 c

分析資料集的來源與形式

✚ 變方分析/多重比較

處理	重複	變數
T1	1	
T1	2	
T1	3	
T1	4	
T2	1	
...	...	
...	...	
...	...	
T4	3	
T4	4	

- 資料取得之前要先經適當的**試驗設計**
- 分析時僅針對**個別變數**
- 一定要有**重複**
- 分析變數要符合變方分析**使用前提之隨機常態資料**
- 處理資料形式可為量化(如濃度)或僅是名義(如品種名稱)，使用統計軟體分析時一定要指定處理欄位

✚ 相關分析 ✚ 回歸分析

處理或取樣點	變數1	變數2
T1		
T2		
T3		
T4		
T5		
...		
...		
...		

- 資料來源為**試驗或取樣調查**
- 分析時針對**兩個或兩個以上變數**
- 若原始資料有重複，分析時**通常取其平均值**
- 分析變數皆為符合相關或回歸**使用前提之可量化資料**
- 軟體分析時處理的資料形式不重要，且無須指定

- 探求作物某兩性狀間
- 雜草種子數目與成長日數間
- 乳牛體重與泌乳量間
- 膠質蛋白溶解度與pH值間

....兩變數關係是否存在

相關分析

- 乾物重受溫度
- 生長速率受施肥用量
- 昆蟲族群致死率受殺蟲劑濃度
- 反應物含量受合成時間

.....之影響程度

回歸分析

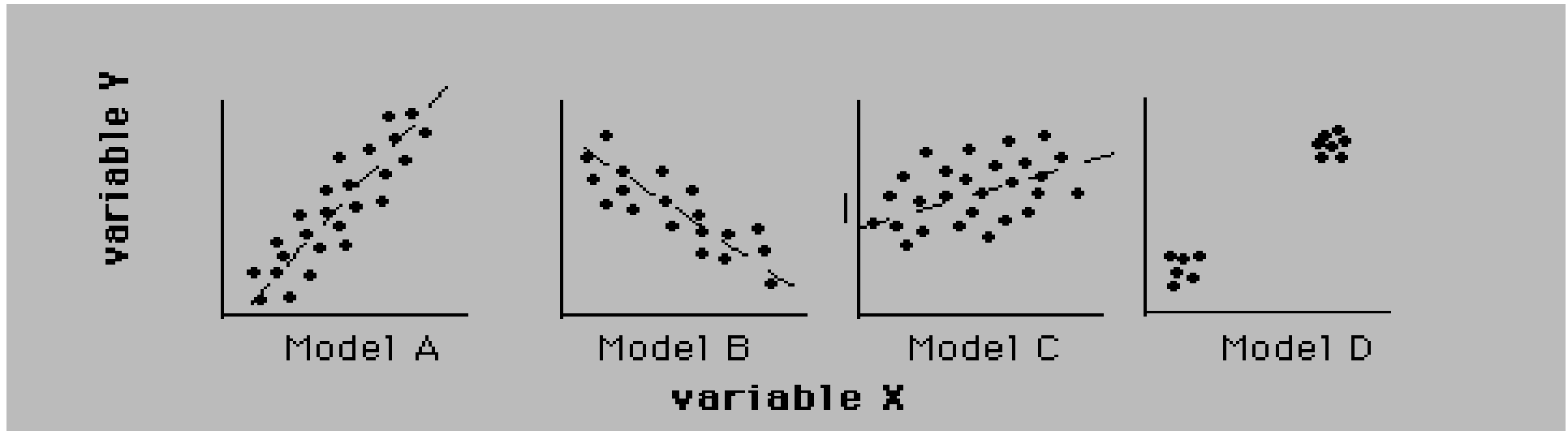
探討變數間的變異程度是否存在某種關係或其強度，未必等於因果關係→最好都不要將變異和因果混為一談

以下是一些**相關回歸不等於因果**的實例：

- 美國超商發現尿布與啤酒的銷售量有高度正相關→買尿布(因)導致愛喝啤酒(果) **X**
- 睡眠與心智健康存在關聯性→睡眠時數異常(因)導致失智症(果) **X**
- 吃口香糖比率與犯罪率間存在正相關→吃口香糖(因)導致犯罪(果) **X**
- 作物乾物重受溫度之影響呈現直線關係→溫度是乾物重增加之原因 **X**

在進行相關或回歸分析前，需先瞭解變數間 呈現何種關係--散佈圖(scatter plot)

各種可能散佈圖：



A：X,Y有正向關係，Y隨X增加而增加

B：X,Y有負向關係，Y隨X增加而減少

C：X,Y無一定關係

D：X,Y表面上看來有比例關係，但實際上是由2個異質族群所構成，各群內X,Y無關係存在，但從族群平均值來看則呈現X,Y間具有比例關係。**故必須先探求構成異質族群的原因，而由此可能會獲得難以預料的重要結果。**

兩個變數資料常用來探討的問題

相關(correlation)

- ✓ 相關是指兩個變數之間**直線關聯的強度與方向**
- ✓ 瞭解相關，通常有二種方式，一為**繪製資料散佈圖**，另為**計算相關係數 (correlation coefficient)**

預測(prediction)

- ✓ **當兩變數相關存在時，可進行簡單直線或曲線回歸分析** (simple linear/non-linear regression analysis)，通常由**一個解釋變數** (independent variable, 或稱自變數、獨立變數、預測變數, X)，來預測**一個反應變數** (dependent variable, 或稱隨變數、依變數、被預測變數, Y)
- ✓ 當欲利用**多個解釋變數** ($X_1, X_2, \dots$) 來預測一個反應變數 (Y) 時，則使用**複回歸分析** (multiple regression analysis)。

相關與回歸的使用場合差異性

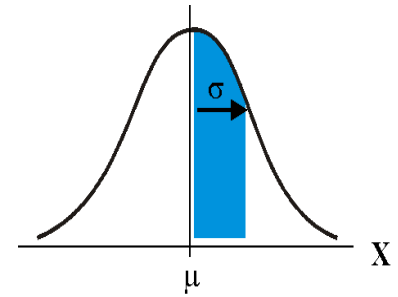
	試驗目的	數據性質
相關分析	探知兩變數X,Y (<u>無反應變數與解釋變數之分</u>) 間的 <u>相互關聯程度</u>	X,Y為 <u>不可自由變動的常態隨機變數</u> 例如所調查的兩數量性狀
回歸分析	描述反應變數Y受到解釋變數X的 <u>影響程度</u>	Y為常態隨機變數；而X <u>通常為可自由選擇或控制的固定值 (模式I回歸)</u> ，例如所設定之不同生長箱溫度或所施用之不同藥劑濃度

註1：若解釋變數X不是固定值，則使用模式II回歸

註2：模式I之回歸係數由最小平方法(least squared method. LS) 估得 (傳統方法，一般統計軟體回歸分析皆以此作為內設值)

模式II之回歸係數由最大概度法(maximum likelihood method, MLE) 估得 (某些統計軟體如SAS另提供可使用MLE的廣義線性模型)

變數的常態性檢定



- ✓ 常態分布是最常見的連續分布，是統計顯著性測驗最常見的理論基礎
- ✓ 呈鐘形的對稱曲線，平均、中位數和眾數都位於同一點上
- ✓ 自然界許多事件接近常態分布，例如：身高、體重、瞳孔距離、眼球屈折率等
- ✓ 當其他分布的觀測值個數 n 很大時，往往亦漸呈常態分布
- ✓ 若干變值本非常態，可進行資料轉換使其接近常態分布

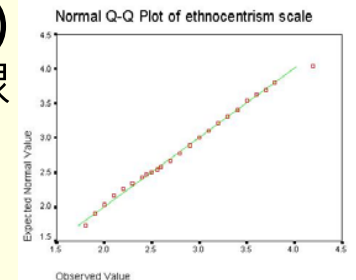
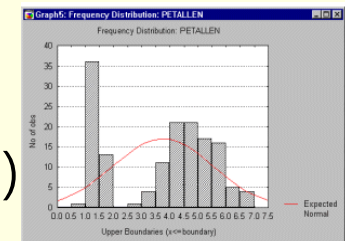
1. 配適度測驗:

- Shapiro-Wilk test
- Kolmogorov-Smirnov test
- Cramer-von Mises test
- Anderson-Darling test

上述統計量倘不顯著，代表常態。但樣本太小，不易不顯著；而樣本太多，則容易顯著，因此最好同時配合圖示方法來判斷常態性。(SAS提供上述4種或後3種統計量，視版本和指定選項)

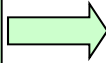
2. 圖示方法:

- Histogram(直方圖)
分布形狀要「看起來是常態」，但小樣本場合(<50)不太能有效看出資料真正的分布型態
- Probability plot(機率圖)
簡稱Q-Q plot，近於直線為常態，是最常用且有效的常態性檢測圖示工具，但有時會流於主觀判斷



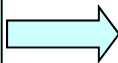
並非所有資料都一定呈常態分布

產量、乾物質重、株高、
生育日數、開花日數....



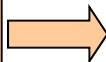
- 連續變數
- 測量(measure)資料形式
- 呈常態分布

防治率、危害率、落果
率、罹病度、雜草覆蓋
率、成功率、感染率...



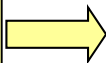
- 連續變數
- 百分比(%)資料形式
- 呈二項分布

開花數、收穫果數、
培植體數...



- 離散變數
- 個數(count)資料形式
- 樣本夠大時可趨於常態分布

細菌數、蟲數、被害果
數、雜草株數、罹病植
株數...



- 離散變數
- 個數(count)資料形式
- 呈倍數成長的非線性資料或 Poisson 分布

轉換方法的常見選擇(一)

- ❖ 呈倍數成長的非線性資料, 其變方和平均值間通常呈比例關係
: 例如罹病植株數或蟲數 (但組織培養培植體數通常不呈倍數成長, 故不屬此類) → 對數轉換

設原始資料為 Y , 採用 $\log(Y)$ 轉換;

若資料含有0時, 用 $\log(Y+1)$ 轉換

(用 $\log_{10}$ 或 $\log_e$ 均可, 但人工計算上前者較方便)

- ❖ 由小的計數值所組成時的Poisson分布資料, 其變方等於平均值
: 例如在一個培養皿之菌落數或區域內某一指定物種的蟲數、產蛋數或植株數 → 開方根轉換

設原始資料為 Y , 採用 $\sqrt{Y}$

當資料大都 <15 尤其等於0時, 用 $\sqrt{Y + \frac{1}{2}}$

當所有值都 ≤ 2 , 用 $\sqrt{Y} + \sqrt{Y + 1}$

轉換方法的常見選擇(二)

❖以比例值呈現之二項資料(不呈常態分布時)：例如種子發芽或單位試區內感染植株的比例，當所有數值都均勻落在各種可能比例值，或僅集中在0-20%與80-100%兩極端比值→角度轉換

設原始比例為Y,則採用 $\arcsin \sqrt{Y}$ (即 $\sin^{-1}\sqrt{Y}$)

若其中資料數值為0，則該值用 $\arcsin \sqrt{1/(4n)}$

若其中資料數值為1，則該值用 $\arcsin \sqrt{1 - 1/(4n)}$

(n為形成比例值的觀測個數,而轉換後的值可採用弧度或度表示)

(但當資料落在30-70%之間無須轉換；若資料為0~20%或80~100%之極端值而非同時存在兩邊端極值時，採開方根轉換 $\sqrt{Y}$ ，但80~100%要先以被100%減去後的值再轉換)

※注意:上述比例值轉換所代入公式的Y值必須是0~1間比例值而非百分比!! 如25%→Y=0.25

❖等級(序位)資料：常態計分 $= (Y - \mu) / \sigma$ (μ, σ 各為平均和標準差)

❖計量反應資料：

- 機率值($\text{probit} = 5 + (y - \mu) / \sigma$)
- 對數值($\text{logit} = \log(p / (1 - p))$, $p = \% \text{data}$)

活用Excel函數進行資料轉換

範例

轉換方法	Excel函數
Log(Y) log(Y+1)	LOG10(Y) 或 LN(Y) LOG10(Y+1) 或 LN(Y+1)
$\sqrt{Y}$ $\sqrt{Y + 0.5}$ $\sqrt{Y} + \sqrt{Y + 1}$	SQRT(Y) SQRT(Y+0.5) SQRT(Y)+ SQRT(Y+1)
arcsin $\sqrt{Y}$ arcsin $\sqrt{1/(4n)}$ arcsin $\sqrt{1-1/(4n)}$	ASIN(SQRT(Y)) ASIN(SQRT(1/(4*n))) ASIN(SQRT((1-1/(4*n)))) Y為0~1數值非百分比值， n為形成比例值的觀測個數

注意:arcsin角度變換也可改用DEGREES(ASIN(SQRT(Y)))函數



動動腦-1



~不是所有兩變數都可同時進行相關與回歸~

- 用「相關」或「回歸」？
- 當使用回歸時，何者是「反應變數(Y)」、「解釋變數(X)」？
 - 樹齡 *VS* 樹高
 - 體重 *VS* 血壓
 - 汽油價格 *VS* 搭捷運乘人數
 - 殺草劑濃度 *VS* 雜草死亡率
 - 盤固草產量 *VS* 氣象因子

2. 相關分析

2.1 簡單相關係數

Pearson product-moment correlation, 簡稱simple correlation

相關是指在同一樣本空間兩個變數X,Y之間**直線關聯的方向和強度**，X,Y無解釋變數與反應變數之分並都符合**常態隨機**，兩變數間的關聯強度以**相關係數(r)**表示

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

Diagram labels for the formula:

- mean of X (points to $\bar{x}$)
- mean of Y (points to $\bar{y}$)
- SD of X (points to S_x)
- SD of Y (points to S_y)

限制條件： $|r| \leq 1$

利用t法檢定**r是否顯著存在** (r是否顯著異於0)

$$|t| = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}$$

(n: 觀測數目, 雙尾檢定)

當t值顯著性機率值:

Prob(t) < 0.05 顯著 (*)

< 0.01 極顯著 ()**

SAS輸出範例：欲探討某種作物4個農藝性狀(x1, x2, x3, x4)兩兩之間的相關性，各性狀均逢機取出10株進行調查

Pearson Correlation Coefficients, N = 10 Prob > r under H0: Rho=0				
	x1	x2	x3	x4
x1	1.00000	0.97417 <.0001	0.98350 <.0001	0.97341 <.0001
x2	0.97417 <.0001	1.00000	0.99510 <.0001	0.93972 <.0001
x3	0.98350 <.0001	0.99510 <.0001	1.00000	0.96041 <.0001
x4	0.97341 <.0001	0.93972 <.0001	0.96041 <.0001	1.00000

x1與x2間的相關係數0.97417，其顯著性機率值Prob<0.0001，即可寫成

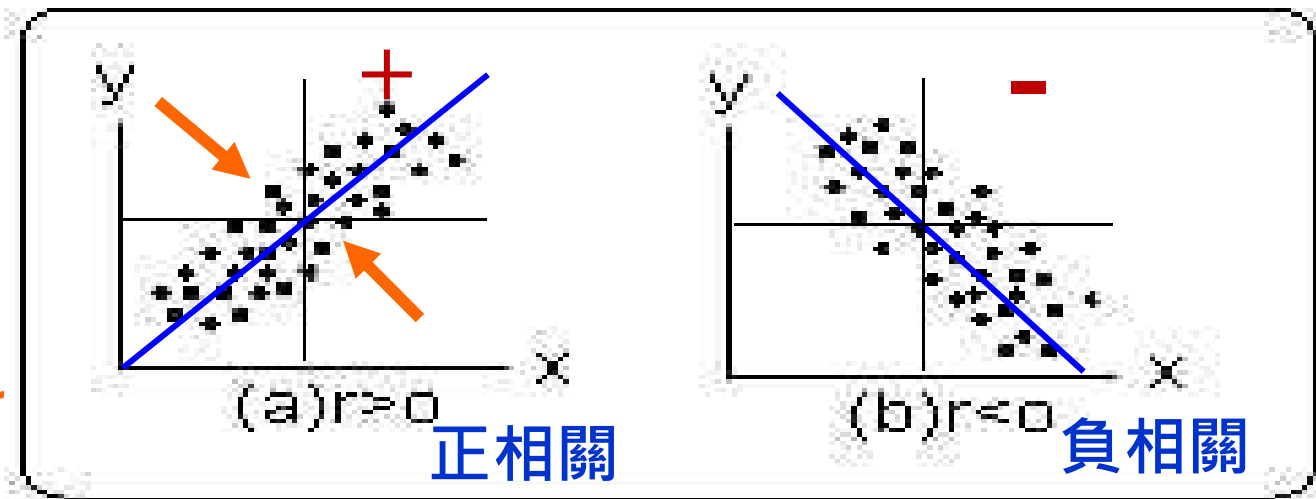
$r=0.97^{**}$

其他成對變數之分析結果類推

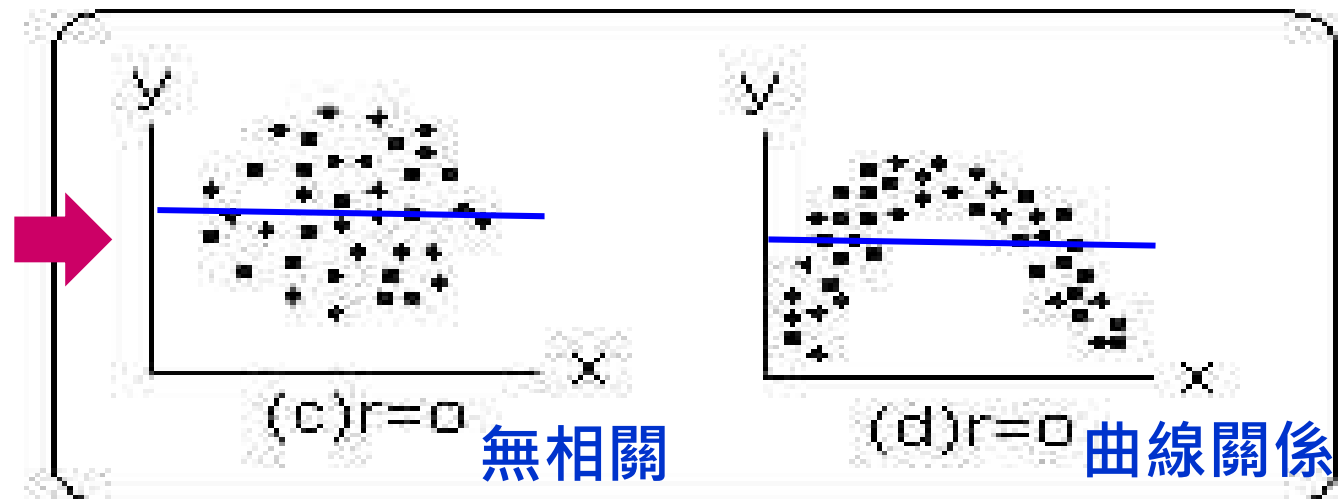
相關係數的範圍, $-1 \leq r \leq 1$

$|r|$ 值大小表示強度，值的 $+$, $-$ 表示方向

愈密集，
相關性增強，
 r 愈接近1



$r=0$
表示可能無
相關或其他
非線性關係



相關分析的結果解讀

【例1】自某校隨機抽取**46名**學生，檢定其大一與大四時體重間的關係，相關分析結果 $r = 0.9431$, $\text{Prob} < 0.0001$

→兩者間呈極顯著的正相關關係，相關係數為**0.94****

【例2】自某作物品種隨機抽取各**5片**葉片，檢定其葉綠素含量與葉片厚度間的關係，相關分析結果 $r = 0.8712$, $\text{Prob} = 0.104$

→兩者間的正相關係數雖達**0.87**，但未達顯著水準**5%**
(此是因為觀測數目**n**太小，故**r**雖大到**0.87**，但卻顯得無意義，相關性結果仍不足採信) → 宜增加樣本數

【例3】自某校隨機抽取**200名**學生，檢定其大一與大四時體重的關係，相關分析結果 $r = 0.2501$, $\text{Prob} = 0.002$

→兩者間呈極顯著的正相關關係，相關係數為**0.25****

(此結論顯得荒謬!!! **r**僅有**0.25**，應不足以認定為正相關，但卻因觀測數目**n**太大，故檢定結果呈極顯著關係)

$$|t| = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}$$

r 的 t 法顯著性檢定中，由於 t 值大小受到觀測個數 n 的影響，因此
當 n 太大時 t 值較大而易顯著；
當 n 太小時 t 值較小而不易顯著

前述【例 3】相關分析的正確作法



若樣本數太多，常導致 r 值不大但顯著；對此，宜先判斷 r 值是否真得已大到足以充分認定是正或負相關（**至少 r 絕對值 > 0.5**），再看其值是否已達統計上顯著水準。

千萬別僅直接以顯著性結果來對 r 下結論，卻忽略了 r 值大小！

2.2 相關的濫用及誤用



(1) 當心「無意義相關」

---- 以下皆是無意義相關，應避免濫用----

- 單株產量與公頃產量間呈正相關？
- 地上部與地下部乾重比例呈負相關？
- 調查性狀在不同取樣時間之間呈正相關？
- 牛胃和羊胃的反芻能力呈正相關？



(2)不是所有形式的資料都能 進行相關分析





(3)勿太倚賴相關的顯著性



相關係數不顯著，不表示兩變數間沒有關係

→表示兩變數間沒有顯著的「**直線**」關係，但無法證明其間沒有其他更複雜的非線性關係

不正確結論：(a) X, Y 互相獨立；(b) X, Y 之間無任何關係

正確結論： X, Y 間無(對稱增加)直線關係



受到樣本數極大的影響

→通常樣本數愈小， r 值愈大(只有兩點時 r 必定為1)，但若樣本數太多，常導致 r 值不大但顯著



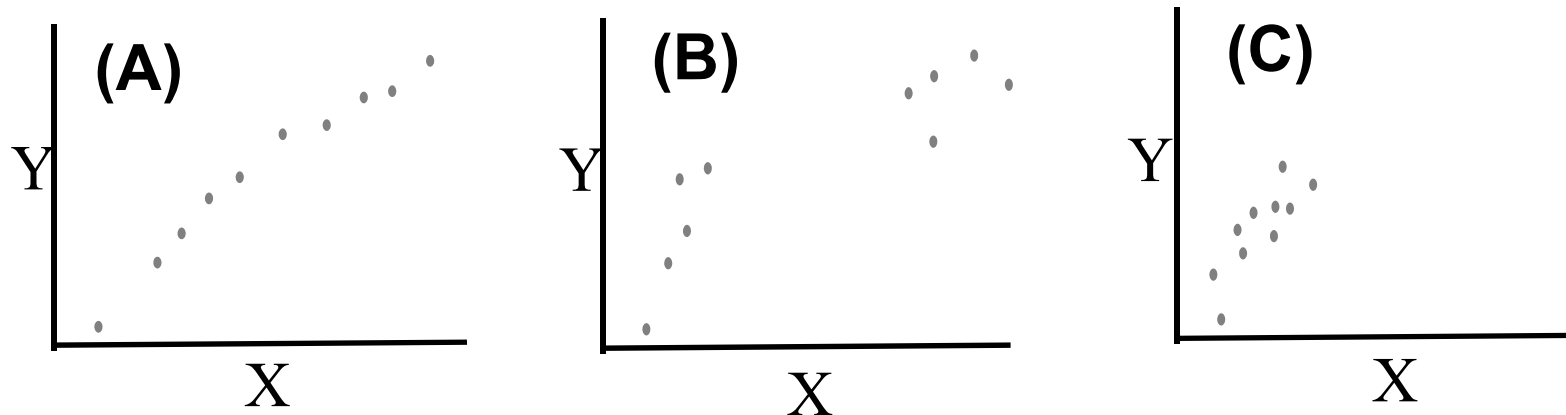
動動腦-2



相關分析的分析點數(觀測個數, n)
最好落在哪個範圍最適合?

難有定論，端視分析資料的變異性及範圍是否足以
看出兩變數間關係而定→先繪製散佈圖可約略拿捏

例如以下同樣都是10個分析點數，您覺得夠嗎？



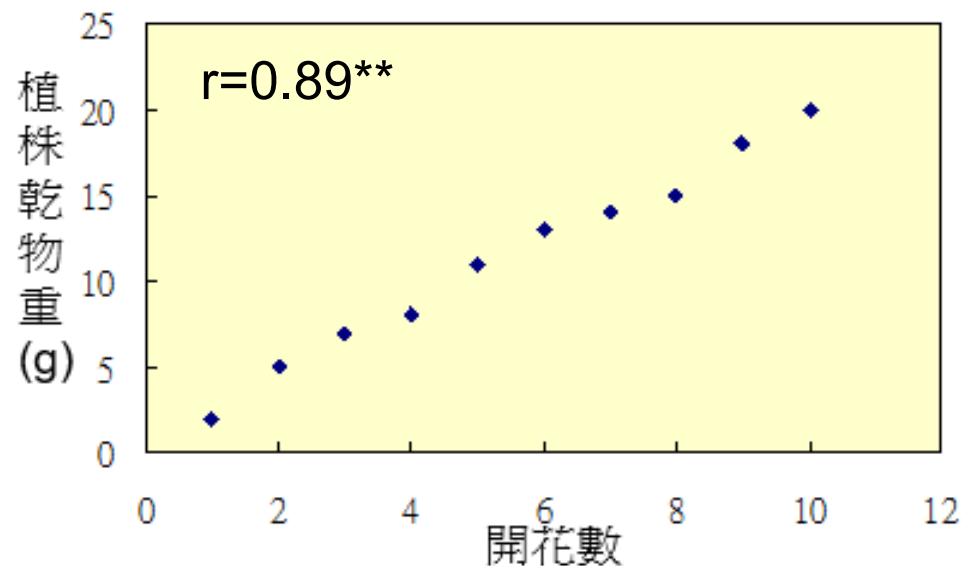
2.3 相關分析結果發表的正確呈現

僅有兩個變數之場合

可有以下兩種表達方式：

- 在內文中僅以文字描述，例如：植株乾物重與開花數間的相關係數(r)為 0.89^{**} ，呈現極顯著的正向關係，表示植株乾物質愈重，其開花數愈多
- 以散佈圖呈現，圖中標示 r 值及其顯著性星號(置於 r 值右側)，並於內文闡述分析結果所代表的意義

相關係數一般四捨五入取至小數點後兩位即可



多個變數兩兩間之場合

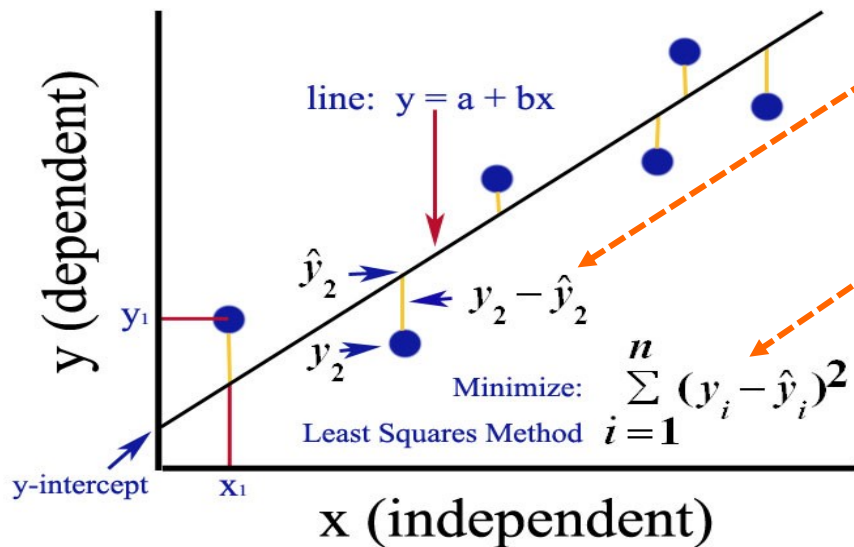
以**相關矩陣**形式呈現，顯著星號直接標示於相關係數右側，並於表下註明用以分析的觀測個數(n)。例如下表。
(上或下三角矩陣皆可，對角線都是1可略去)

	植株乾物重	開花數	結果數	果實重
植株乾物重				
開花數	0.86**			
結果數	0.67*	0.90**		
果實重	0.25	0.33	-0.71*	
n=10				

一般相關係數取自**小數點後兩位**即可，可讓表格看起來簡潔

3. 回歸分析

「回歸」一詞源於1885年英國優生學家高登(Sir F. Galton) 論文 “Regression towards mediocrity in hereditary stature”，他檢視兒童身高對應其父母身高，發現身高超過平均之父母的孩子身高通常也超過平均，但並沒有父母那麼高，他稱這種現象為「**朝平均數(中間值)回歸**」，意思是往回走。



■ 每個實際值(觀測點)與回歸估計值(觀測點對應到回歸線上的值)之間的差距，稱為**殘差(residual)** $y_i - \hat{y}_i = e_i$

■ **殘差平方和**代表無法以回歸估計式解釋的程度(偏差程度)；當所有觀測點都在線上時，殘差平方和=0

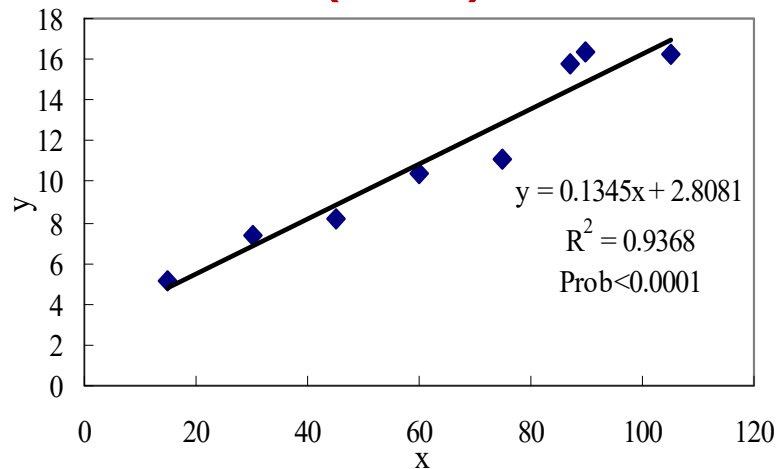
殘差平方和最小下估計出回歸係數，即最小平方方法

- **$R^2=1$ -殘差平方和/總平方和**，稱為**決定係數**(coefficient of determination) 或 **複相關係數**(multiple correlation coefficient)，值在0與1之間，表示總變異中利用**回歸式所能解釋之變異所佔比例**，**用來判斷回歸估計式的解釋能力**
- 當觀測個數n不太小時， R^2 愈接近1，表示對該模式愈滿意

回歸的型態

直線、非線性回歸：僅有一個解釋變數 X
複回歸：多個解釋變數 X_1 、 X_2 、 X_3

簡單(直線)回歸



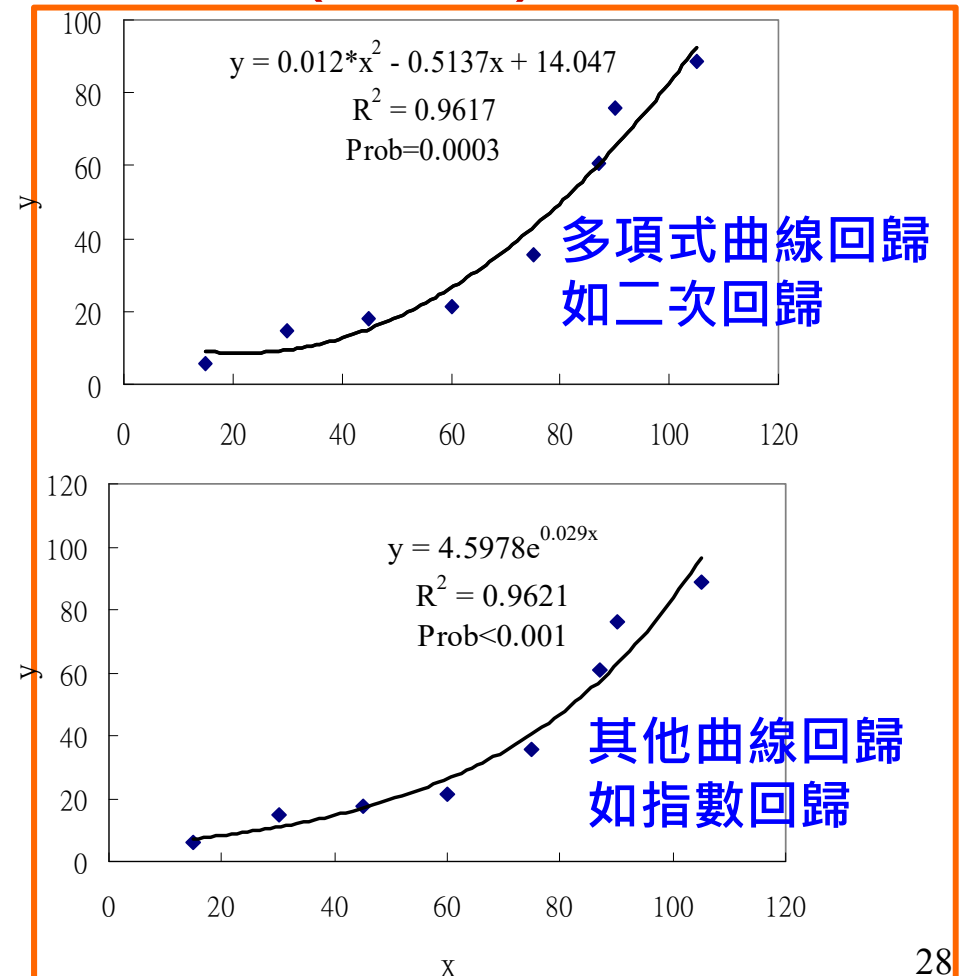
複回歸

$$Y = 10.2 + 0.4X_1 + 0.8X_2$$

$$R^2 = 0.81, \text{Prob} = 0.015$$

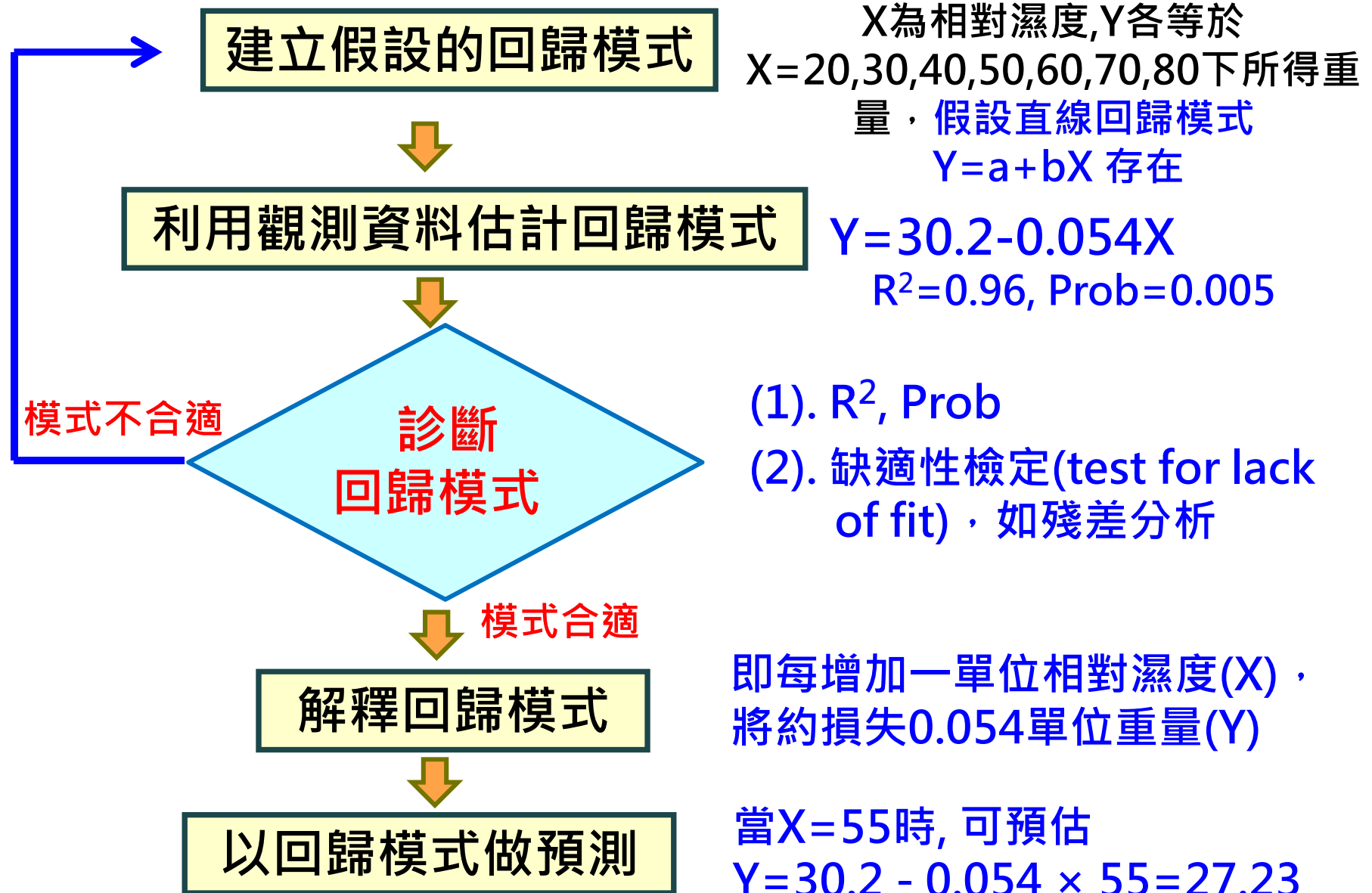
只有兩個解釋變數，尚可用3D立體圖繪製回歸圖，但含3個以上解釋變數則無法繪製回歸圖

曲線(非線性)回歸



回歸分析的步驟

範例



3.1 簡單(直線)回歸

測定單一解釋變數(X)對反應變數(Y)的線性影響程度
{ Y符合常態隨機，而假設X為固定值}

$$Y=a+bX$$

其中 a 為截距 (intercept)
b 為回歸係數 (regression coefficient)

回歸式估計結果的判斷步驟：

- (1) 模式的顯著性機率值(Prob)，顯著(<0.05)或極顯著(<0.01)，表示該模式存在機率很高；若不顯著(>0.05)，則無須再看以下步驟(2)和(3)
- (2) 決定係數 R^2 值愈接近於1，表示模式滿意度愈高
- (3) 當模式顯著且 R^2 值愈高時，可取截距與回歸係數的估計值寫入方程式中

【例4】欲探討水梨果園施用6種氮肥量(X)對梨果可溶性固形物含量(Y)之影響程度，並假設是直線關係

(SAS輸出結果)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.61157	3.61157	105.63	0.0005
Error	4	0.13676	0.03419		
Corrected Total	5	3.74833			

(1) Prob值

(2) R²值

Root MSE	0.18491	R-Square	0.9635
Dependent Mean	15.11667	Adj R-Sq	0.9544
Coeff Var	1.22320		

(3) 截距和回歸係數估計值

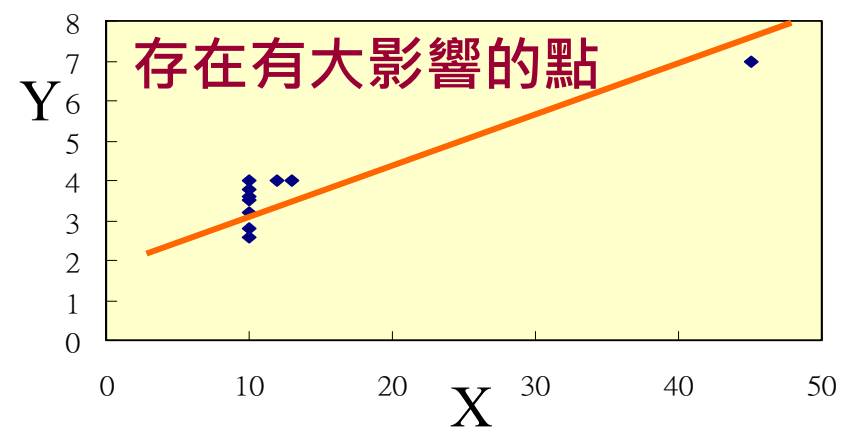
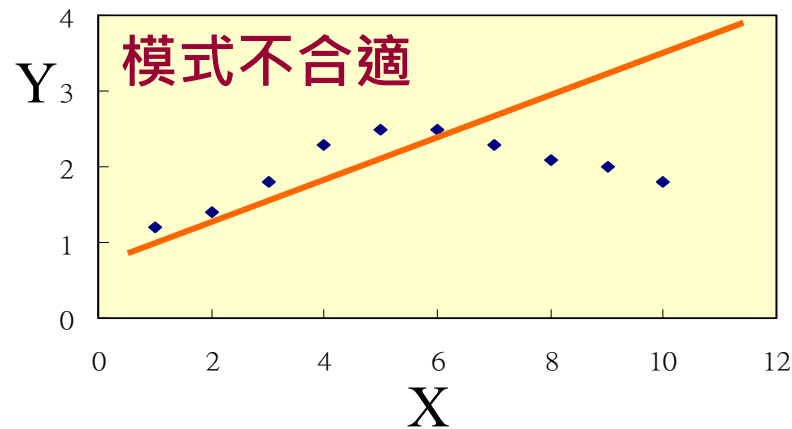
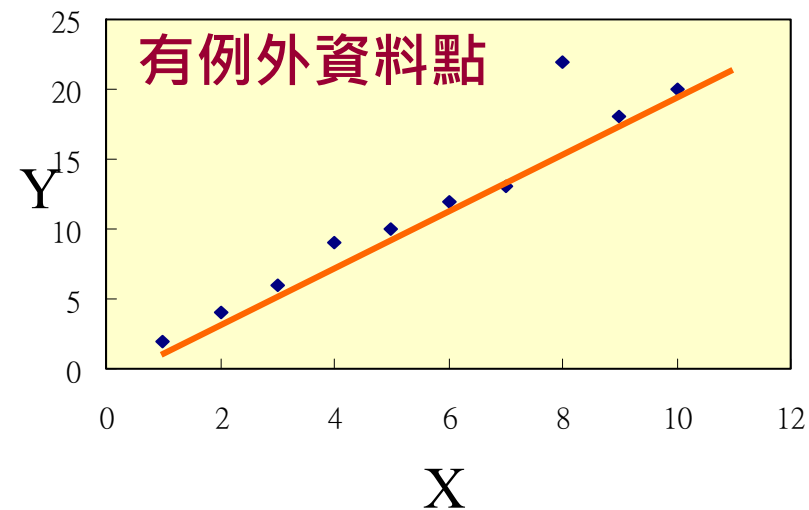
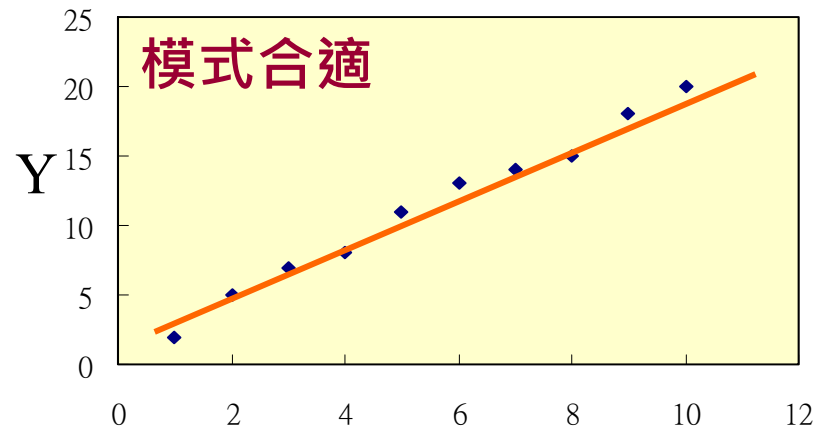
Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	Intercept	1	16.25238	0.13383	121.44	<.0001	0
施氮量(kg)	施氮量(kg)	1	-0.09086	0.00884	-10.28	0.0005	-0.98159

結論：回歸式為 $Y=16.2524-0.0909X$, $R^2=0.96$, Prob=0.0005
或寫成 $Y=16.2524-0.0909^{**}X$, $R^2=0.96$

(直線回歸模式中只有一個回歸係數，模式顯著即代表該回歸係數顯著)

但有時候會發生Prob和 R^2 都能接受 但模式其實不適合的情況

以下四種模式分別配合簡單回歸模式結果，從回歸式、 R^2 到各種平方和的差異都很小，可見由此等訊息並不能幫助我們進行回歸診斷，還必須利用其他一些專門的統計量和直接對殘差進行分析



回歸模式的統計前提

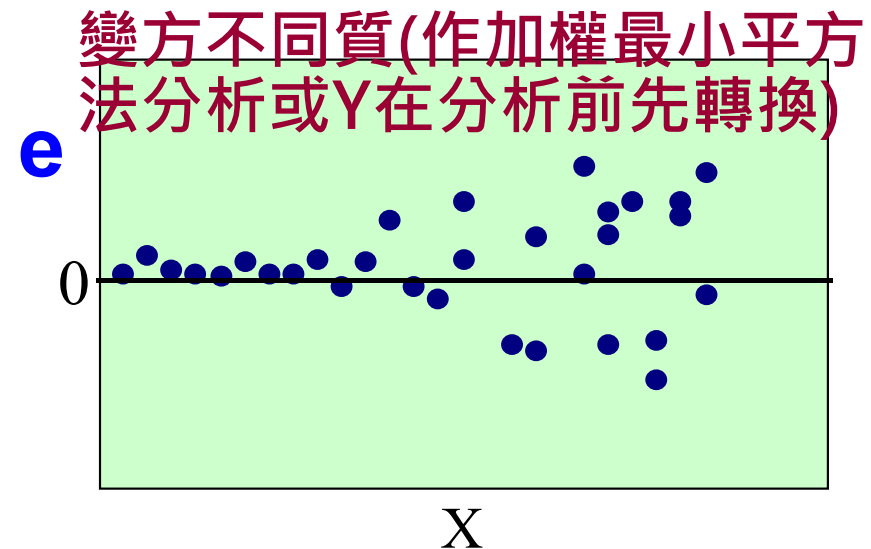
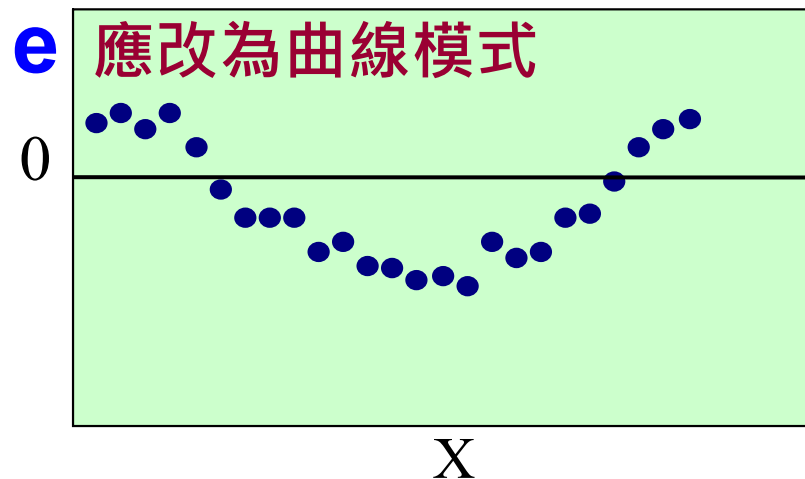
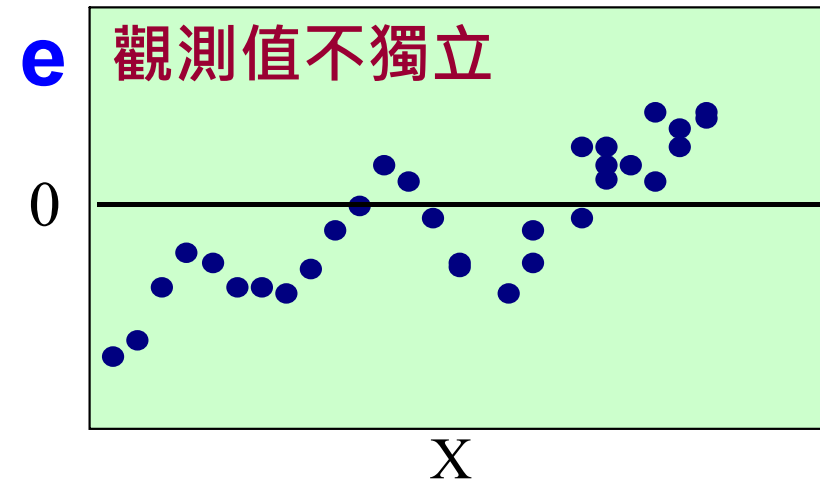
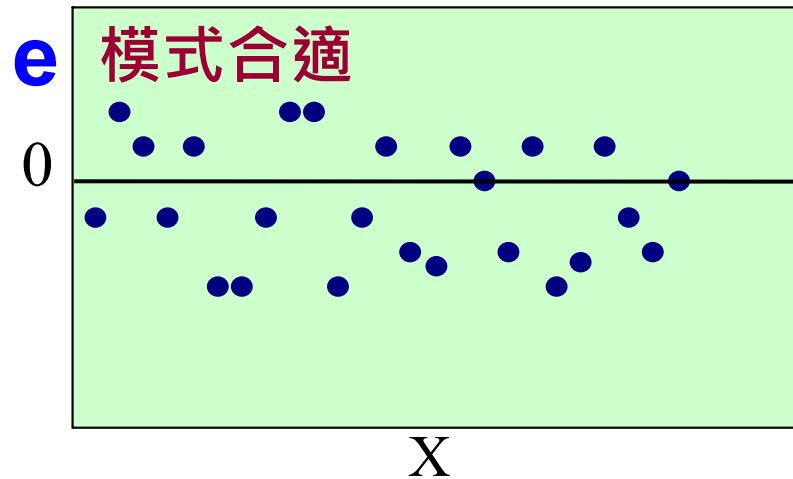
- ⊕ **獨立性**(independence) :
回歸模式的殘差項要互相獨立
- ⊕ **常態性**(normality) :
殘差項要服從常態分布
- ⊕ **線性關係**(linearity) :
解釋變數與反應變數的標準化殘差，都要呈線性關係
- ⊕ **變方同質**(homogeneity of variance) :
解釋變數間的殘差變方要相同



除常態性外，其它統計前提可用「殘差分析」進行檢驗

回歸診斷 - 殘差分析

利用繪製殘差圖 (殘差 e vs X 或 Y) 來檢驗獨立性、線性結構及變方同質



3.2 曲線(非線性)回歸

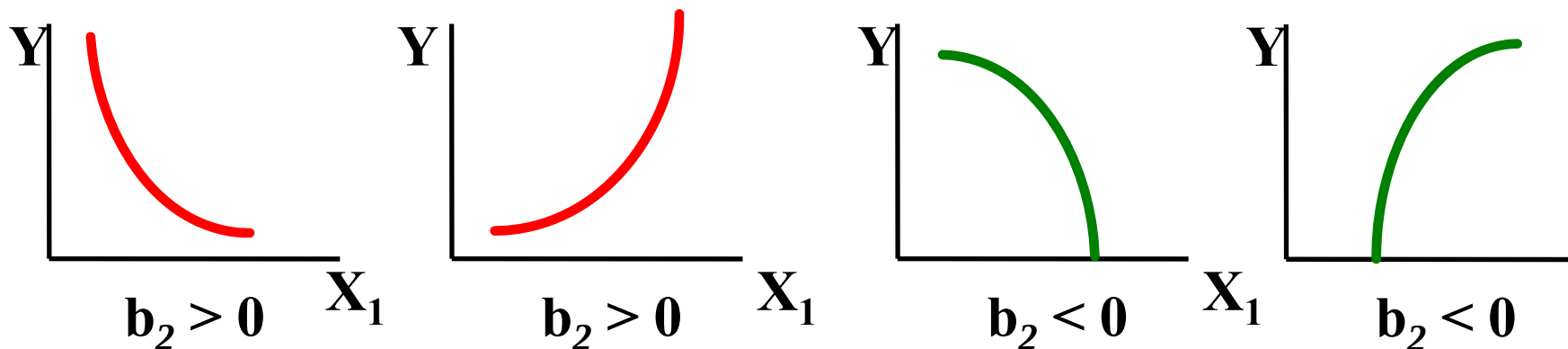
測定單一解釋變數(X)對反應變數(Y)的非線性影響程度
{ Y符合常態隨機，而假設X為固定值}

(1)多項式曲線(polynomial curve)

估計式 $y = a + b_1x + b_2x^2 + \dots + b_kx^k$

例如：二次回歸模式 $y = a + b_1x + b_2x^2$

(b_2 為二次項，其正負決定曲線開口向上或向下)



(2) 其他曲線回歸模式

❖ **實質線性**：經適當變數轉換，如倒數、對數、平方根變換後能改寫成線性模式之形式

例如：

(對數曲線) $y = ax_1^{b_1} x_2^{b_2} x_3^{b_3} \Rightarrow \ln y = \ln a + b_1 \ln x_1 + b_2 \ln x_2 + b_3 \ln x_3$

(指數曲線) $y = ab^x \Rightarrow \ln y = \ln a + x \ln b_1$

(Logistic生長曲線, S形曲線)

$$y = \frac{K}{1 + ae^{-bx}} \Rightarrow \ln\left(\frac{K - y}{y}\right) = \ln a - bx$$

(Logistic回歸模式, 適於Y為二項變數之場合(如罹病-1, 未罹病0))

$$y = \frac{e^{a+bx}}{1 + e^{a+bx}} \Rightarrow \ln\left(\frac{y}{1-y}\right) = a + bx$$

❖ 實質非線性

例如： $y = a + b_1 x^{-b_2 x}$

$$y = a + b_1 x + b_2 (b_3)^x$$

$$y = a x_1^{b_1} x_2^{b_2} x_3^{b_3}$$

- 針對多項式曲線回歸和實質線性曲線回歸，可先進行變數轉換後(例如二次回歸： X^2 ，指數或對數函數： LOG10 或 LN)，再利用統計套裝軟體以線性形式(複回歸或直線回歸)進行分析
- 但很多統計軟體未必提供實質非線性回歸分析

以二次回歸為例

測定單一解釋變數(X)對反應變數(Y)的二次曲線影響關係
{ Y符合常態隨機，而假設 X_i 為固定值}

$$Y=a+b_1X+b_2X^2 \longrightarrow Y=a+b_1X_1+b_2X_2$$

其中 a 為截距

b_1, b_2 為回歸係數

當統計軟體不直接提供曲線回歸時，
通常可利用 X^2 視為第二個新變數 X_2 ，
然後再以複回歸形式進行分析

回歸式估計結果的判斷步驟：

- (1) 模式的顯著性機率值(Prob)，顯著(<0.05)或極顯著(<0.01)，表示該模式存在機率很高；若不顯著(>0.05)，則無須再看以下步驟
- (2) 決定係數 R^2 值愈接近於1，表示模式滿意度愈高
- (3) 當模式顯著且 R^2 值愈高時，可取截距與回歸係數的估計值寫入方程式中
- (4) 並將回歸係數顯著性結果以星號標示於係數旁

範例：前述【例4】欲探討水梨果園施用6種氮肥量(X)對梨果可溶性固形物含量(Y)之影響程度，並假設是二次曲線關係

(SAS輸出結果)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.65455	1.82727	58.45	0.0040
Error	3	0.09379	0.03126		
Corrected Total	5	3.74833			

(1) Prob值

(2) R²值

Root MSE	0.17681	R-Square	0.9750
Dependent Mean	15.11667	Adj R-Sq	0.9583
Coeff Var	1.16964		

(3) 截距和回歸係數估計值

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	Intercept	1	16.13929	0.16025	100.71	<.0001	0
施氮量(kg)	施氮量(kg)	1	-0.05693	0.03015	-1.89	0.1554	-0.61504
施氮量-二次	施氮量-二次	1	-0.00136	0.00116	-1.17	0.3256	-0.38187

(4) 回歸係數顯著性
(截距顯著性不重要)

結論：二次回歸不成立

(雖模式Prob極顯著且R²也甚高，但兩個回歸係數並不顯著)

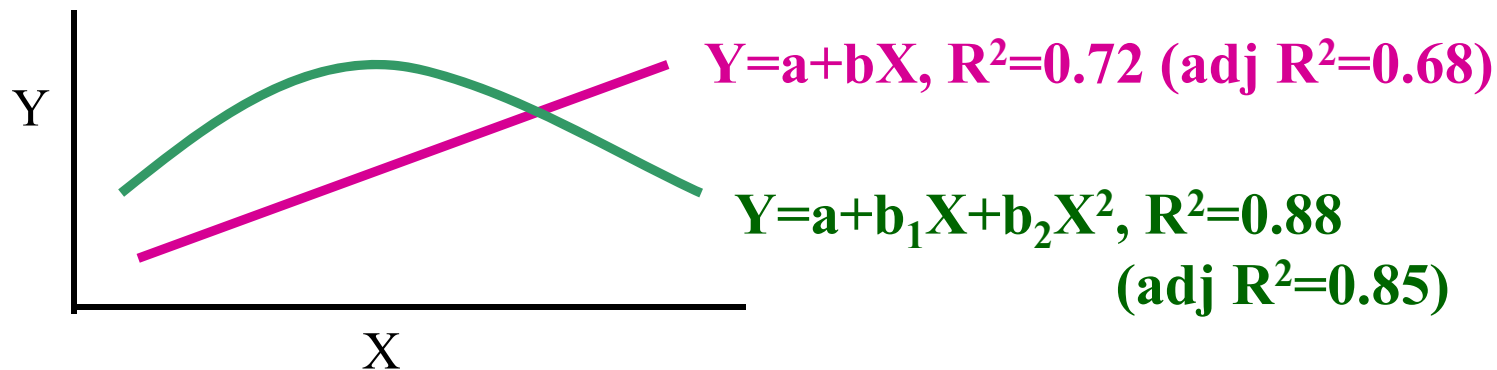
同一套資料集配合不同回歸模式的選擇準則

-- 矯正(adjusted) R^2

Root MSE	0.18491	R-Square	0.9635
Dependent Mean	15.11667	Adj R-Sq	0.9544
Coeff Var	1.22320		

$$R_{adj}^2 = 1 - \left[(1 - R^2) \frac{n-1}{n-k-1} \right]$$

- 當模式中的解釋變數愈多， R^2 值也會愈高，應採用矯正 R^2 來比較不同模式的滿意度
- R^2 是一種矯正值，可反映出解釋變數之數目(k)和樣本大小(n)，通常 $< R^2$
- 適於同一套資料在假設所有可能回歸模式之間的比較



Select which model?

同一資料集配合不同回歸式之結果，視**矯正R²**高低而取捨，但若R²增進有限，除非要求預測準確度極高，否則應考量選擇含有較少解釋變數之模式，避免模式因過於複雜而不利解釋

[例A]

$$Y=a+bX, \text{ adj } R^2=0.62$$

$$Y=a+b_1X+b_2X^2, \text{ adj } R^2=0.89 \quad \checkmark$$

$$Y=a+b_1X+b_2X^2+b_3X^3, \text{ adj } R^2=0.91$$

[例B]

$$Y=a+bX, \text{ adj } R^2=0.86 \quad \checkmark$$

$$Y=a+b_1X+b_2X^2, \text{ adj } R^2=0.89$$

$$Y=a+b_1X+b_2X^2+b_3X^3, \text{ adj } R^2=0.90$$

但仍須注意：

(1)分析點數不可太少

(2)由殘差分析結果，殘差分布須隨機分散

3.3 複回歸 (多元回歸)

測定多個解釋變數($X_1, \dots, X_k$)對反應變數(Y)的綜合影響
{ Y 符合常態隨機，而假設 X_i 為固定值}

$$Y = a + b_1 X_1 + \dots + b_k X_k$$

其中 a 為截距
 $b_1, \dots, b_k$ 為回歸係數

回歸式估計結果的判斷步驟：

- (1) 模式的顯著性機率值(Prob)，顯著(<0.05)或極顯著(<0.01)，表示該模式存在機率很高；若不顯著(>0.05)，則無須再看以下步驟
- (2) 決定係數 R^2 值越接近於1，表示模式滿意度越高
- (3) 當模式顯著且 R^2 值越高時，可取截距與回歸係數的估計值寫入方程式中
- (4) 並將回歸係數顯著性結果以星號標示於係數旁
- (5) 通常會考慮剔除VIF(variance inflation) > 10 或共線性診斷 >100 的變數

複回歸的共線性問題

- 許多研究人員在進行回歸分析時，常常對於解釋變數之間的相關性沒有審慎評估，就貿然將許多個解釋變項同時放到回歸方程式內
- 當複回歸模式中解釋變數之間有太高的相關時，由於在估計回歸係數的運算過程中會導致數理上所稱的奇異性(singularity)，就會產生一些「不合理」現象，造成**估計值的不穩定，或是回歸係數與相關係數正負符號不一致等問題**，這些稱為共線性(collinear 或 multicollinear)問題
- **共線性是一種程度的問題**(degree of matters)，而不是全有或全無(all or none)的狀態，應盡量消除它

共線性造成的困擾

【例5】收集某公司 12 位員工的年齡(X1)和年資(X2)對薪資(Y)的影響

所得複回歸式 $Y = 69.315 - 0.796X1 + 2.889X2$, $R^2=0.70$, $\text{Prob}=0.0045$
但兩兩變數相關係數 $r(Y,X1)=0.81^{**}$, $r(Y,X2)=0.83^{**}$, $r(X1,X2)=0.98^{**}$

X1與X2間的高度相關(共線性)，常造成以下問題：

(1) 回歸係數與相關係數符號不一致的困擾

上式中X1 回歸係數為負，表示年齡越大其薪資越少，這與 Y 和 X1之相關係數為正，看起來矛盾！

(2) 回歸係數檢定不顯著的困擾

上述整個模式雖極顯著，但兩回歸係數卻不顯著。解釋變數間有很高相關(如本例X1與X2)，常發現回歸係數檢定不顯著，此意味著該解釋變數可不放入模式內。事實上這是可理解的，因其中一個先進入模式中就無須再放入第二個變數。但初學者時常以為兩個解釋變數皆不顯著，便誤認它們對反應變數Y都沒有貢獻，其實只要放入其中一個解釋變數即可，不需要兩個都進入模式。

共線性檢驗方法

(1) 變異數膨脹因子(variance inflation factor, VIF)

(2) 共線性診斷(collinearity diagnostics)

當某解釋變數與其他解釋變數關係密切時，其 VIF 或共線性診斷之條件索引(condition index)的值也大，表示該解釋變數幾乎是其他解釋變數的線性組合，因此可考慮將它從模式中剔除。一般以 **VIF>10** 或 **共線性診斷之條件索引>100** 時，表示有共線性問題

Weight_log 這個解釋變數的 VIF 雖 <10 但共線性診斷之條件索引高達 448，建議從模式剔除

參數估計值							
變數	標籤	自由度	參數估計值	標準誤差	t 值	Pr > t	變異數膨脹
Intercept	Intercept	1	-4.07761	0.62034	-6.57	<.0001	0
Invoice		1	-0.00000724	0.00000166	-4.36	<.0001	3.30752
Cylinders		1	0.11465	0.00835	13.73	<.0001	3.06662
Horsepower		1	0.00161	0.00024256	6.64	<.0001	4.25340
Wheelbase	Wheelbase (IN)	1	-0.00359	0.00183	-1.96	0.0511	6.16558
MPG_Highway	MPG (Highway)	1	-0.00283	0.00182	-1.56	0.1205	3.28522
Length	Length (IN)	1	0.00430	0.00091700	4.69	<.0001	5.08111
Weight_log		1	0.49403	0.08668	5.70	<.0001	8.25953

共線性診斷											
			變異的比例								
數目	特徵值	條件索引	Intercept	Invoice	Cylinders	Horsepower	Wheelbase	MPG_Highway	Length	Weight_log	
1	7.80881	1.00000	0.00000145	0.00036191	0.00026551	0.00024334	0.00001436	0.00019255	0.00001926	0.00000113	
2	0.13702	7.54927	0.00002659	0.03336	0.01124	0.02157	0.00004655	0.04024	0.00007373	0.00001103	
3	0.02934	16.31402	0.00000347	0.43108	0.13590	0.00054771	0.00124	0.04766	0.00216	0.00000903	
4	0.01164	25.90010	0.00007690	0.13084	0.66311	0.50767	0.00164	0.02929	0.00160	0.00006024	
5	0.01014	27.75301	0.00047081	0.13087	0.11675	0.44766	0.00813	0.34973	0.00811	0.00052278	
6	0.00239	57.11399	0.01207	0.08051	0.04672	0.01615	0.05268	0.10316	0.14520	0.00575	
7	0.00061722	112.47895	0.00137	0.00411	0.00771	0.00000901	0.79586	0.02213	0.82051	0.00011870	
8	0.00003890	448.05712	0.98598	0.18887	0.01832	0.00615	0.14038	0.40760	0.02231	0.99352	

【例6】研究身高大致相同的13位男人血管收縮壓(Y)與受到身高(X1)、腰圍(X2)、體重(X3)和年齡(X4)的影響程度

(SAS輸出結果)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1455.88878	363.97220	94.29	<.0001
Error	8	30.88045	3.86006		
Corrected Total	12	1486.76923			

(1) Prob值

(2) R²值

Root MSE	1.96470	R-Square	0.9792
Dependent Mean	130.30769	Adj R-Sq	0.9688
Coeff Var	1.50774		

(3) 截距和回歸係數估計值

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate	Variance Inflation
Intercept	Intercept	1	88.64612	29.84946	2.97	0.0179	0	0
身高(cm)	身高 (cm)	1	-0.17578	0.13065	-1.35	0.2154	-0.09964	2.11232
腰圍(CM)	腰圍 (CM)	1	-0.24351	0.13294	-1.83	0.1044	-0.24647	6.97350
體重(Kg)X1	體重 (Kg) X1	1	1.15366	0.08783	13.13	<.0001	1.36419	4.15482
年齡(歲)X2	年齡 (歲) X2	1	0.36721	0.07551	4.86	0.0013	0.44639	3.24512

(4) 回歸係數顯著性

(5) VIF值

結論：回歸式為 $Y = 88.65 - 0.18X1 - 0.24X2 + 1.15^{**}X3 + 0.37^{**}X4$
 $R^2 = 0.98$, Prob < 0.0001

(5)共線性診斷之條件索引

Variable	Collinearity Diagnostics							
	Number	Eigenvalue	Condition Index	Proportion of Variation				
				Intercept	身高(cm)	腰圍(CM)	體重(Kg)X1	年齡(歲)X2
Intercept	1	4.85582	1.00000	0.00001411	0.00002606	0.00010598	0.00034918	0.00116
身高(cm)	2	0.13198	6.06560	9.300107E-7	0.00000571	0.00237	0.01335	0.15165
腰圍(CM)	3	0.00955	22.55338	0.00478	0.01991	0.00085563	0.37952	0.39371
體重(Kg)X1	4	0.00244	44.64628	0.00208	0.03971	0.67255	0.58222	0.21169
年齡(歲)X2	5	0.00021536	150.15598	0.99312	0.94034	0.32411	0.02456	0.24180

除年齡(X4)外其餘變數皆<100

表示年齡(X4)這個解釋變數的VIF雖<10但共線性診斷之條件索引大於150，存在共線性問題，建議從模式剔除

	身高(X1)	腰圍(X2)	體重(X3)	年齡(X4)	血壓(Y)
身高(X1)					
腰圍(X2)	-0.68**				
體重(X3)	-0.66**	0.86**			
年齡(X4)	0.43	-0.81**	-0.70**		
血壓(Y)	-0.64**	0.64**	0.90**	-0.36	

年齡(X4)與腰圍(X2)、體重(X3)間存在極顯著負相關，但與血壓(Y)無顯著相關

前例倘一開始若不考慮共線性問題，同時納入四個解釋變數所估得的回歸式為

$$Y = 88.65 - 0.18X_1 - 0.24X_2 + 1.15^{**}X_3 + 0.37^{**}X_4$$

$$R^2 = 0.98, \text{Prob} < 0.0001$$

其中X1(身高)、X2(腰圍)兩個解釋變數的回歸係數不顯著，意味著這兩個變數對Y的貢獻度不高



注意！不可以將回歸係數不顯著的解釋變數從原回歸式中直接移除。必須將資料集剔除不顯著變數資料後，再重新分析，以獲得新回歸式

進而不考慮這兩個解釋變數，重新進行Y(血壓)與X3(體重)、X4(年齡)之複回歸分析，結果如下：

$$Y = 39.99 + 1.09^{**}X_3 + 0.46^{**}X_4, \quad R^2 = 0.98, \text{Prob} < 0.0001$$

前例若一開始就將存在共線性的年齡(X4)不納入，三個解釋變數所估得的回歸式為

$$Y = 169.16 - 0.36X_1 - 0.64^{**}X_2 + 1.12^{**}X_3, R^2 = 0.92, \text{Prob} < 0.0001$$

其中X1(身高) 回歸係數不顯著，意味著該變數對Y的貢獻度不高



進而不考慮這個解釋變數，重新進行Y(血壓)與X2(腰圍)、X3(體重)之複回歸分析，結果如下：

$$Y = 97.59 - 0.56^{*}X_2 + 1.18^{**}X_3, R^2 = 0.90, \text{Prob} < 0.0001$$

共線性移除後，X2(腰圍)的貢獻重要性就凸顯出來了

該複回歸式也不存在共線性問題：(兩解釋變數的VIF < 10且共線性診斷之條件索引 < 100)

參數估計值						
變數	自由度	參數估計值	標準誤差	t 值	Pr > t	變異數膨脹
Intercept	1	97.58628	8.61286	11.33	<.0001	0
腰圍(CM)	1	-0.55975	0.20086	-2.79	0.0192	3.95926
體重(Kg)	1	1.17746	0.17193	6.85	<.0001	3.95926

共線性診斷					
數目	特徵值	條件索引	變異的比例		
			Intercept	腰圍(CM)	體重(Kg)
1	2.97890	1.00000	0.00179	0.00050393	0.00099748
2	0.01818	12.80141	0.51344	0.00378	0.19873
3	0.00292	31.91729	0.48477	0.99571	0.80027

$$Y = 97.59 - 0.56 \times X_2 + 1.18 \times X_3, \quad R^2 = 0.90, \text{ Prob} < 0.0001$$

結果闡釋：

X_2 (腰圍)與 X_3 (體重)都對 Y (血壓)有影響，在體重固定情形下，每增加腰圍1單位，血壓減少0.56單位；在腰圍固定情形下，每增加1單位體重，血壓增加1.18單位

預測範例：

當 $X_2 = 85\text{cm}$ 且 $X_3 = 60\text{kg}$ 時，可預估

$$\begin{aligned} Y &= 97.59 - 0.56 \times 85 + 1.18 \times 60 \\ &= 120.79 \text{ mm/Hg} \end{aligned}$$

只能內插預測！即 X_2 與 X_3 用以預測的數據都必須落在該兩變數原資料範圍內

複回歸分析的最佳模式選擇

~~決定模式中應納入哪些解釋變數的方法~~

常見方法如下：

✚ 所有可能回歸模式法

將所有可能的回歸模式皆考慮，再依一些準則 (如 R^2 、矯正 R^2 、 $C(P)$ 、預測平方和PRESS等統計量) 來選擇變數

✚ 順向選擇法

選定一個標準，開始式中沒有解釋變數(常數項除外)，按解釋變數對Y的貢獻度由大到小依序挑選進入式中，直到沒有變數可被引入為止

✚ 反向剔除法

選定一個標準，開始將所有變數均放入式中，按解釋變數對Y的貢獻由小到大依序剔除變數，直到沒有變數可被剔除為止

✚ 逐步回歸法

結合順向選擇法與反向剔除法兩種方式的優點

3.4 逐步回歸

Stepwise regression

逐步回歸不是某種回歸型態，而是複回歸分析中的一種最佳模式選擇方法---適合解釋變數甚多的場合

✿ 分析原理：

- (1)選定一個標準 【通常在SAS中解釋變數的進入標準(F機率值)為0.15，剔除標準則為0.15】
- (2)開始方程式中沒有解釋變數(常數項除外)，按解釋變數對Y的貢獻度由大到小依序挑選進入式中 【順向選擇】
- (3)在每一步驟已被納入的解釋變數，再按解釋變數對Y的貢獻度由小到大依序將變數剔除 【反向剔除】
- (4)直到方程式外的變數均達不到入選標準，沒有解釋變數可被納入式中為止

✿ 回歸式估計結果的判斷步驟：如同複回歸分析之五個步驟

如前述【例6】血壓資料，當處理共線性問題將年齡變數(X4)先從模式中剔除後，再配合用逐步回歸法來看選定的最佳模型，其結果如下：

變異數分析					
來源	自由度	平方和	平均值平方	F 值	Pr > F
模型	2	1331.55083	665.77541	42.89	<.0001
誤差	10	155.21840	15.52184		
已校正的總計	12	1486.76923			

(1) Prob值

(2) R²值

根 MSE	3.93978	R 平方	0.8956
應變平均值	130.30769	調整 R 平方	0.8747
變異係數	3.02344		

(3) 截距和回歸係數估計值

參數估計值						
變數	自由度	參數估計值	標準誤差	t 值	Pr > t	變異數膨脹
Intercept	1	97.58628	8.61286	11.33	<.0001	0
腰圍(CM)	1	-0.55975	0.20086	-2.79	0.0192	3.95926
體重(Kg)	1	1.17746	0.17193	6.85	<.0001	3.95926

(4) 回歸係數顯著性

(5) VIF值

共線性診斷					
變異的比例			條件索引		
數目	特徵值	條件索引	Intercept	腰圍(CM)	體重(Kg)
1	2.97890	1.00000	0.00179	0.00050393	0.00099748
2	0.01818	12.80141	0.51344	0.00378	0.19873
3	0.00292	31.91729	0.48477	0.99571	0.80027

(5) 共線性診斷之條件索引

$$Y = 97.59 - 0.56 * X_2 + 1.18 * X_3, \\ R^2 = 0.90, \text{Prob} < 0.0001$$

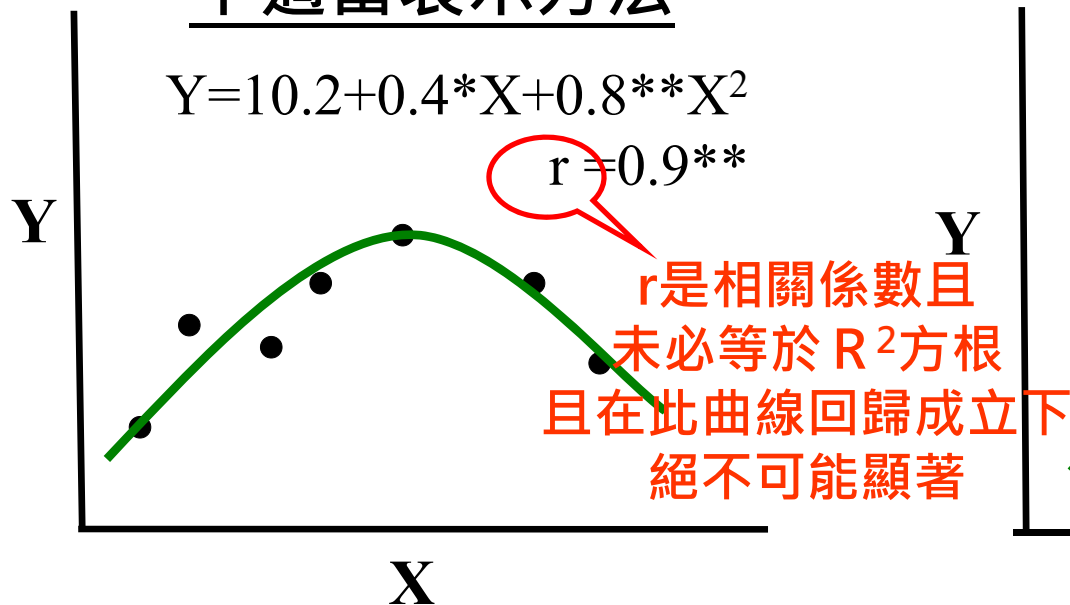


3.5 回歸的濫用及誤用

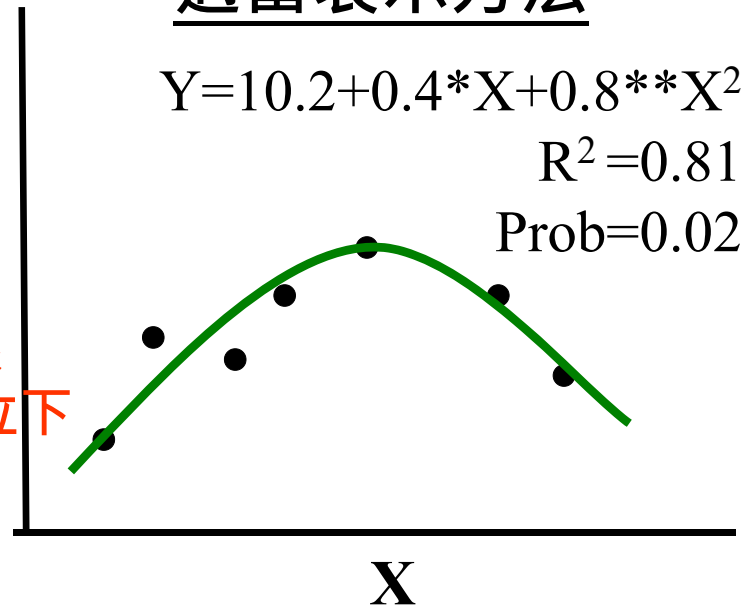
(1) 當心回歸分析之決定係數(R^2)的意義

- R^2 用來檢測回歸配合的適合度： R^2 愈接近 1，表示回歸方程式愈有效
- 直線回歸時 R^2 恰為簡單相關係數 r 的平方值，但複回歸的 R^2 的開方根並不等於 r

不適當表示方法

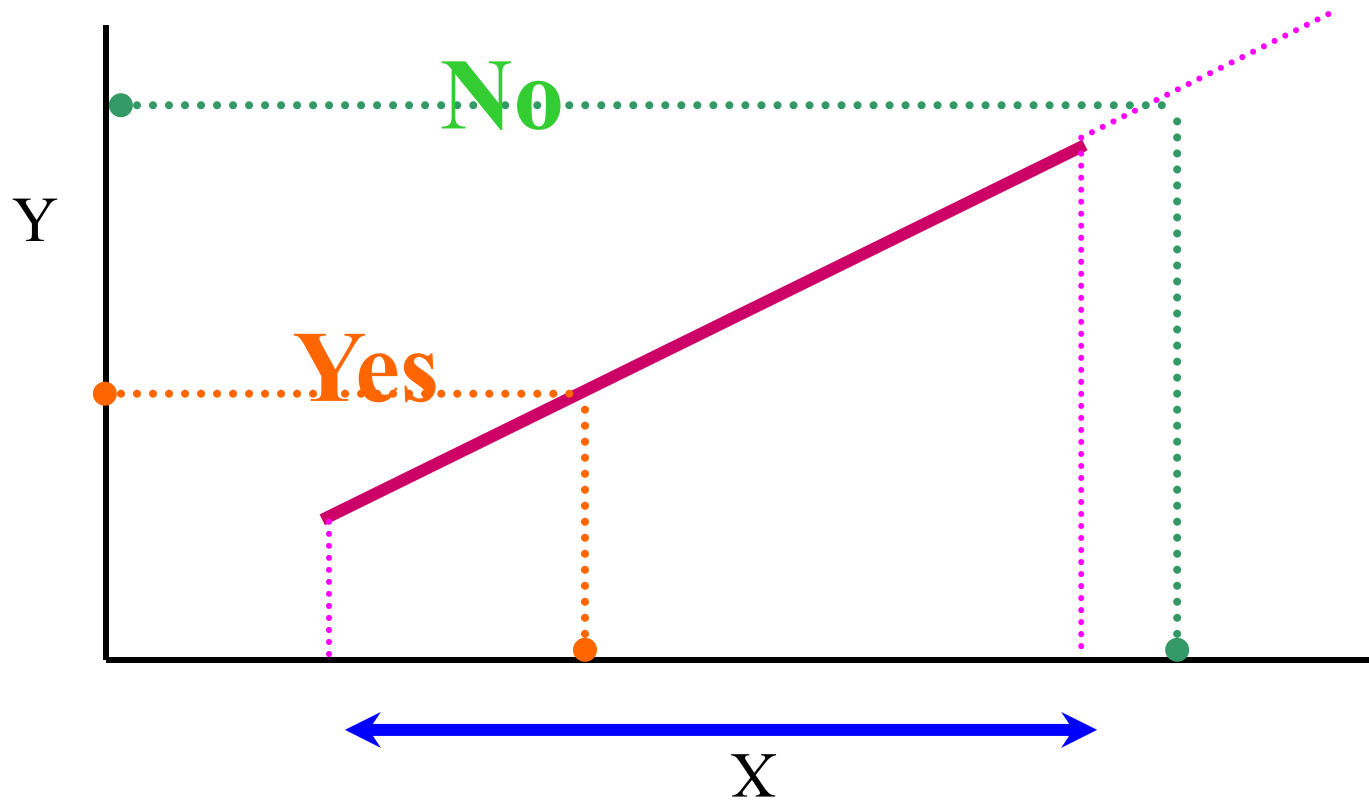


適當表示方法





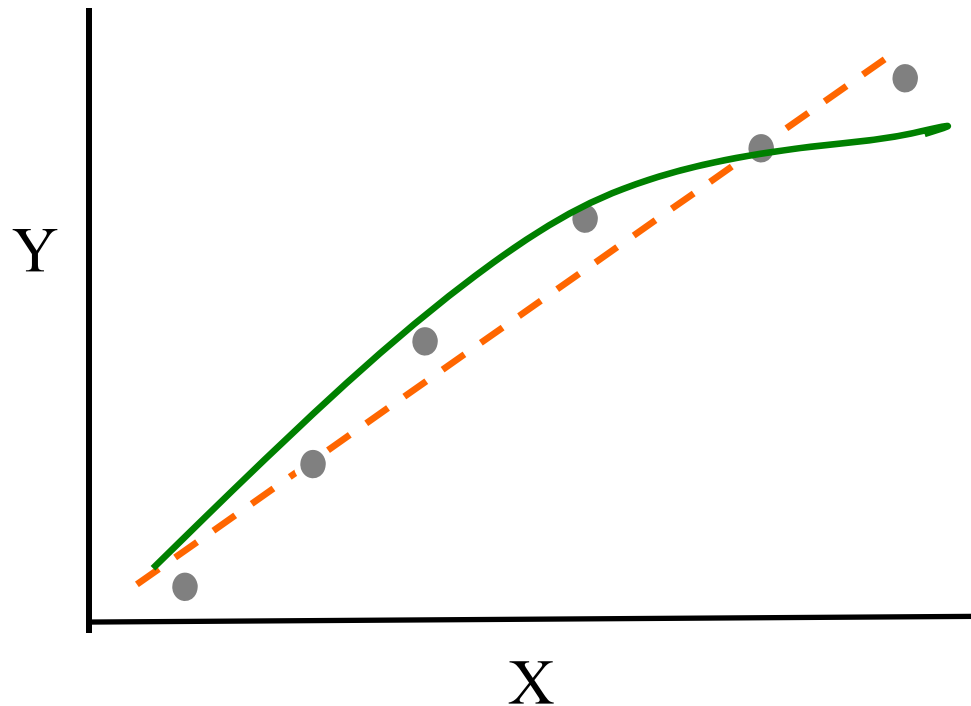
(2) 回歸方程式僅限於內插預測





(3)注意分析點數太少的迷思

當分析點數太少時，任何回歸配合都會顯得很滿意（無論直線？曲線？ R^2 都很高!!），但其實此時的可信度都不高，故不宜貿然取最高 R^2 的回歸關係作為結論





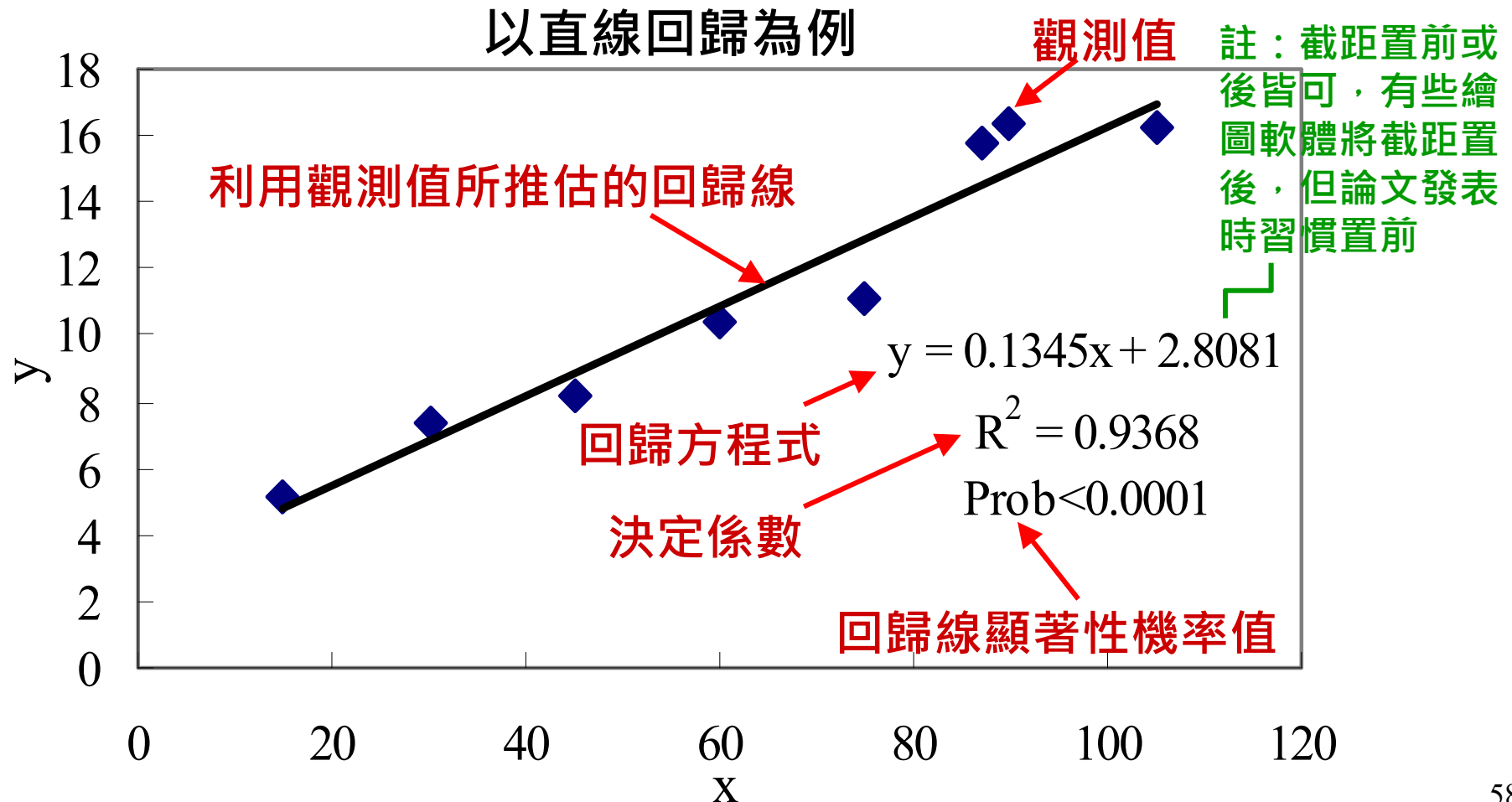
動動腦-3



回歸分析最少需要幾個
分析點數(觀測個數, n)才可信?

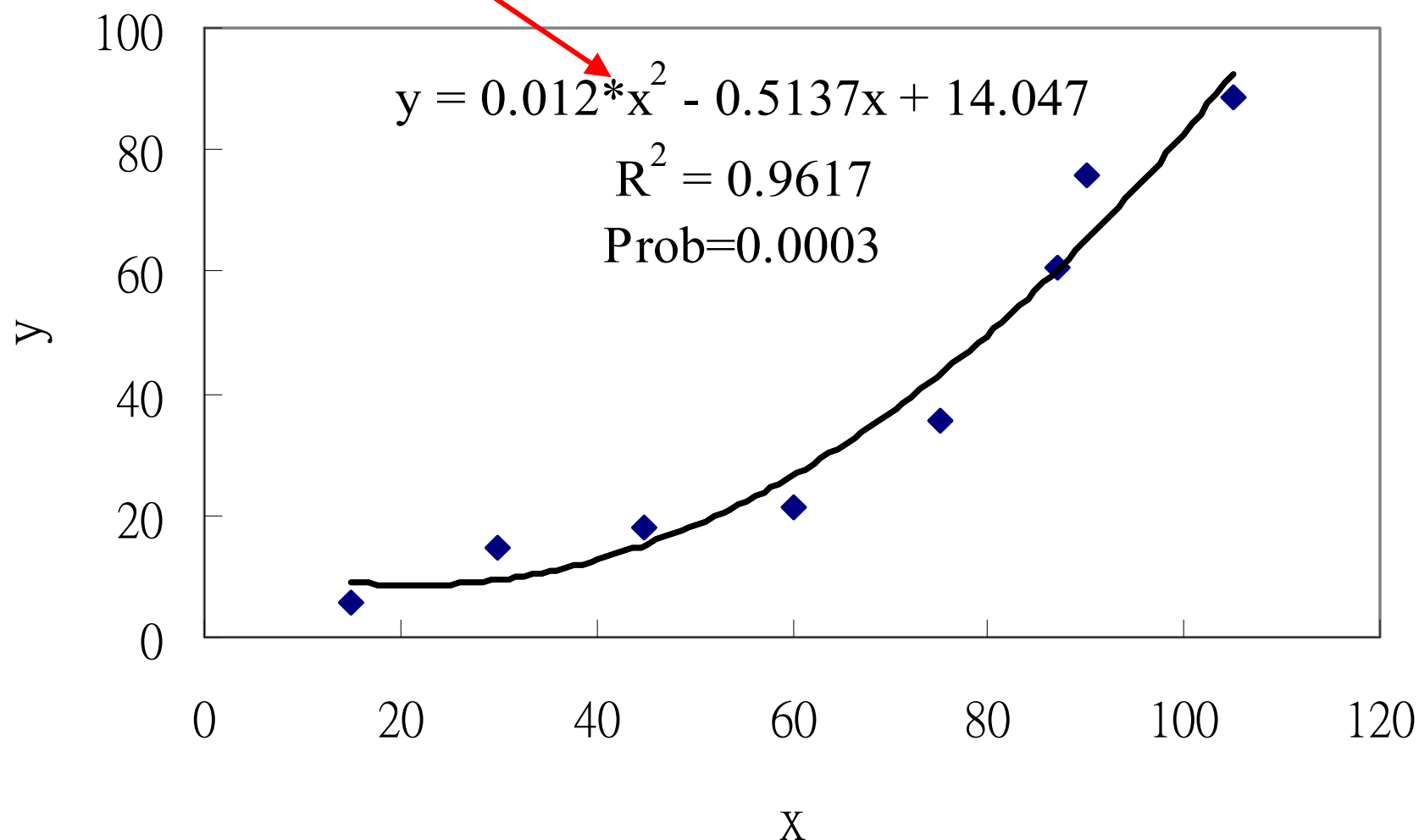
3.6 回歸分析結果發表的正確呈現

- ✓ SCI期刊論文通常要求高品質的回歸線圖
- ✓ 回歸線圖內應包含：觀測值點、回歸線、回歸式、決定係數及顯著性結果



多項式回歸或複回歸的方程式中各係數顯著性 應標示清楚

顯著性標示 以二次曲線回歸為例



Thanks for Your Attention



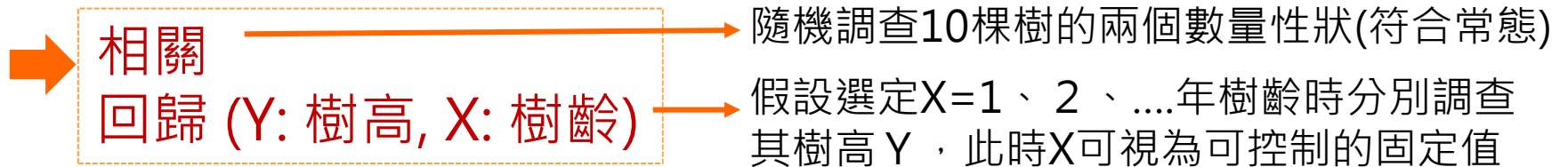


動動腦-1 【解答】



- 用「相關」或「回歸」？
- 當使用回歸時，何者是「反應變數(Y)」或「解釋變數(X)」？

- 樹齡 *vs* 樹高



- 體重 *vs* 血壓

➡ 相關
回歸 (Y: 血壓, X: 體重)

- 汽油價格 *vs* 搭捷運乘人數

➡ 相關
回歸 (Y: 搭乘人數, X: 油價)

- 殺草劑濃度 *vs* 雜草死亡率

➡ 回歸 (Y: 死亡率, X: 濃度)

殺草劑濃度X為固定值，不會是常態隨機變數

- 盤固草產量 *vs* 氣象因子

➡ 回歸 (Y: 產量, X: 氣象因子)

氣象因子X (如溫度、日照時數等)為固定值，不會是常態隨機變數



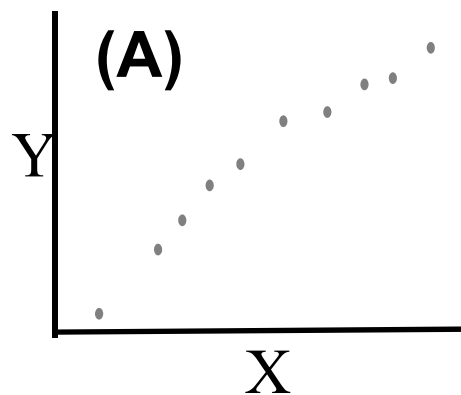
動動腦-2 【解答】



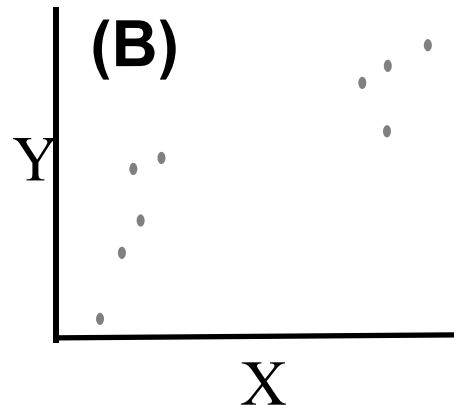
相關分析的分析點數(觀測個數, n)
最好落在哪個範圍最適合?

難有定論，端視分析資料的變異性及範圍是否足以看出兩變數間關係而定→先繪製散佈圖可約略拿捏

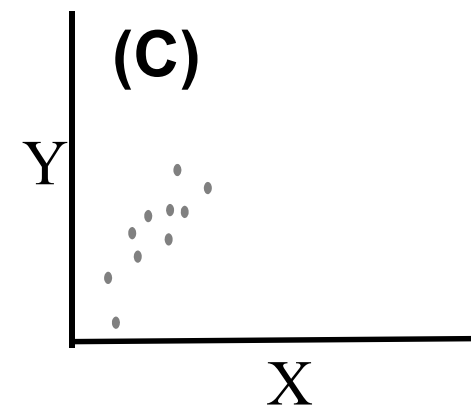
例如以下同樣都是10個分析點數，您覺得夠嗎？



應足以看出趨勢



X變異範圍或許夠廣但集中於兩端，且Y在同樣X值下變異較大，建議增加分析點數，尤其補充X範圍中間數值



建議擴大X變異範圍



動動腦-3 【解答】



回歸分析最少需要幾個分析點數(觀測個數, n)
才可信?

難有定論，端視分析資料的變異性及範圍

- 實驗室要取得標準曲線(檢量線)，因都是標準樣本，通常實驗室認證SOP只要求5個樣本即可。
- 但若是其他試驗場合，在直線回歸模式假設下最好建議至少8個點，這與變方分析(ANOVA)中殘差自由度 $n-2$ 最好不小於6有關(以前試驗設計講習課程曾講過最少重複的計算問題)；若是更複雜的曲線回歸或複回歸，分析點數就要更多。
- 但這並非絕對，要看回歸估計式是要用來探討趨勢而已，還是要用來做精確預測，後者所需分析點數應更多。
- 不管如何，當 n 較少時，即便估得回歸式，在結果闡釋上就應該更保守。