Linear Regression Analysis (1)

杜裕康

國立台灣大學公共衛生學院
流行病學與預防醫學研究所

yukangtu@ntu.edu.tw

學習目標

上完之後，學生應該能夠：

- 瞭解什麼是相關係數(correlation), 簡單線性迴歸(simple regression), 複迴歸(multiple regression) 以及它們的計算方法
- 瞭解相關性不等同於因果關係
- 瞭解什麼是“dummy variables” (虛擬變量)
- 使用複迴歸調整干擾因子(confounders)
- 如何解釋這些分析方法的結果

探討兩個變項之間的關係：

- 一個連續變項 vs 一個類別變項
  - 身高 vs 性別
  - 收入 vs 職業
  - **當類別變項含有兩個類別：t-test**
  - **當類別變項超過兩個類別：ANOVA**
- 一個連續變項 vs 一個連續變項：**相關或簡單線性迴歸分析**
  - 體重 vs 身高
  - 收入 vs 年齡
- 一個連續變項 vs 多個連續變項：**複迴歸分析**

Covariance (共變異數)

變異數：每個觀察值與平均值的距離

- Let Var(*x*) be the variance of *x*:

$$Var(x) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

N-1，因自由度

  - $\bar{X}$ is the mean of x

- Let Cov(*x*, *y*) be the covariance of x and y

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

- Cov(*x*, *x*) is the Var(*x*)

在探討XY，兩個變項觀測值的變化和關係
EX：身高與體重的關係
兩個值要相乘會得正除以N－1(自由度)
變異大可大可小不夠客觀，需一個標準化（以標準差來校正）後的數值來比較（抵消）

變異大可大可小不夠客觀，需一個標準化（以標準差來校正）後的數值來比較（抵消）

Correlation (相關係數)

- 又稱為 Pearson correlation coefficient，是用英國統計學家 Karl Pearson 來命名。
- Correlation coefficient 通常用英文字母 *r* 來代表：

$$r = Cov\left(\frac{x-\bar{X}}{S_x}, \frac{y-\bar{Y}}{S_y}\right) = \frac{Cov(x,y)}{S_x S_y} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{[\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2]}}$$

  - S_x: standard deviation of x; S_y: standard deviation of y

Karl Pearson (1857 – 1936)

現代統計學之父
因他而從數學中獨立而出

Francis Galton (1822 – 1911)

- Galton 是達爾文的表弟，由於受到達爾文的影響，他熱衷於研究遺傳學和優生學，尤其是智力的演化，但是當時沒有一套完善測量智力的方法。於是他想到一個比較容易測量的人類特徵「身高」，高爾登發現「非常高的父母所生的孩子，往往會比父母矮些，而非常矮的父母所生的孩子，則往往比父母高」，他把這個現象稱作「regression toward mediocrity」。他也找到一個數學方法去測量這樣的關係，並且定義為相關係數。

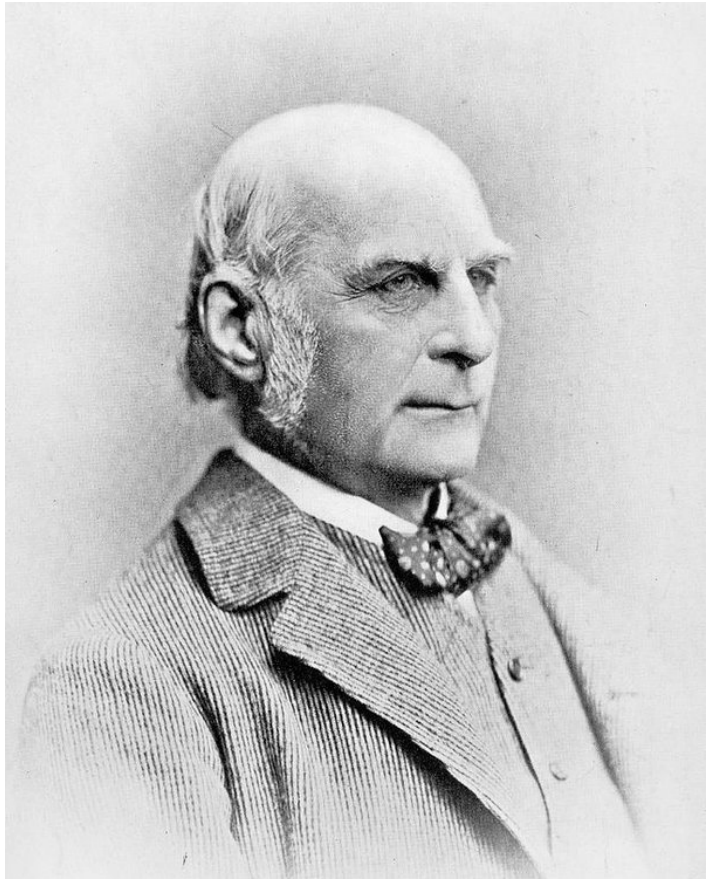
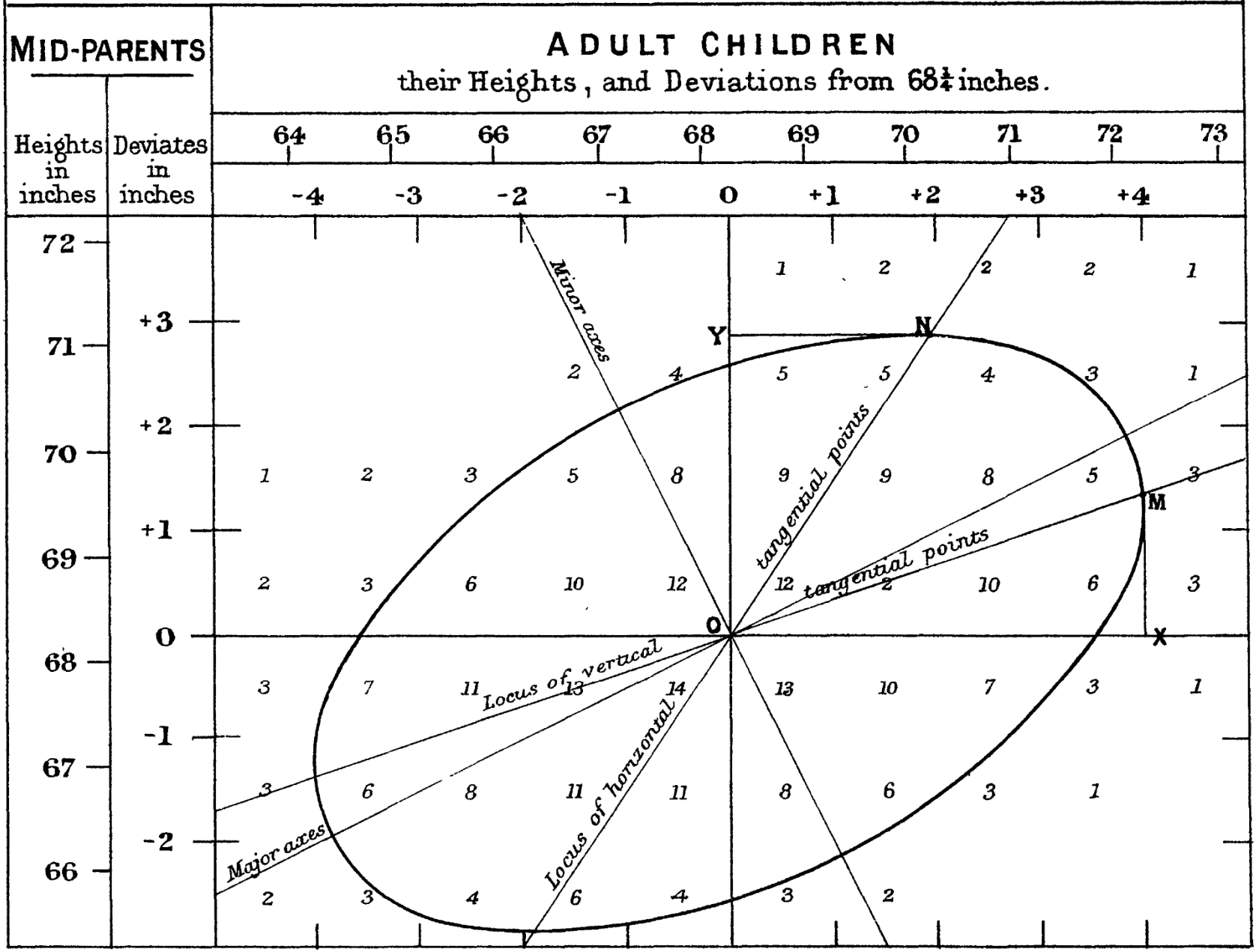


Table 1 One of Galton' s correlation tables (from Francis Galton, Family likeness in stature, *Proceedings of the Royal Society of London* 1886; **40**: 42–73). Galton's 1885 crosstabulation of 928 'adult children' born of 205 mid-parents, by their height and their mid-parent's height

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	40	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

DIAGRAM BASED ON TABLE I.

(all female heights are multiplied by 1·08)



Correlation (相關係數)

- *r* 值的範圍是從 -1 到 1
  - *r* = 1, perfect positive relation
  - *r* = 0, no relation **BH,BW兩者無相關**
  - *r* = -1, perfect negative relation

Proof:

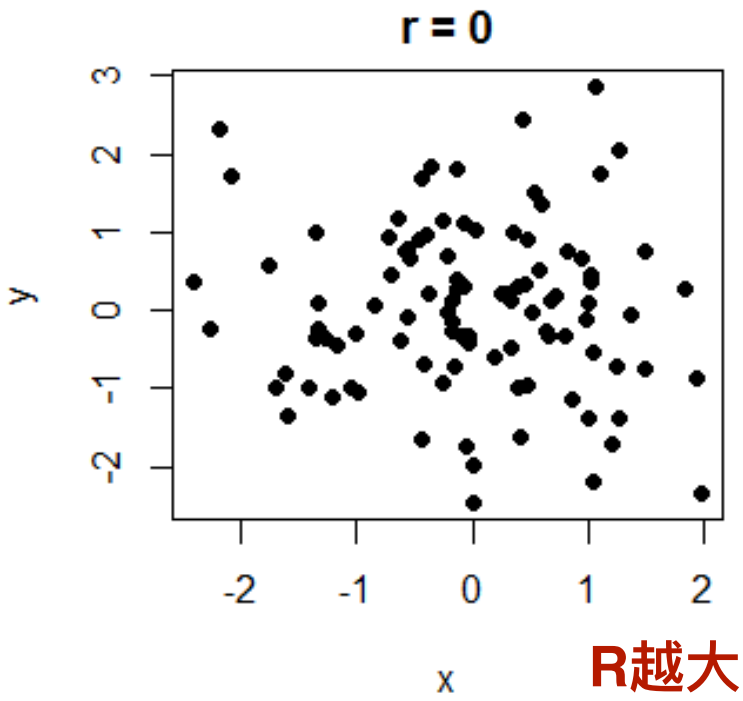
Let $Z_x = \frac{(x-\overline{X})}{S_x}$, and $Z_y = \frac{(y-\overline{Y})}{S_y}$; so $r_{x,y} = Cov(Z_x, Z_y)$

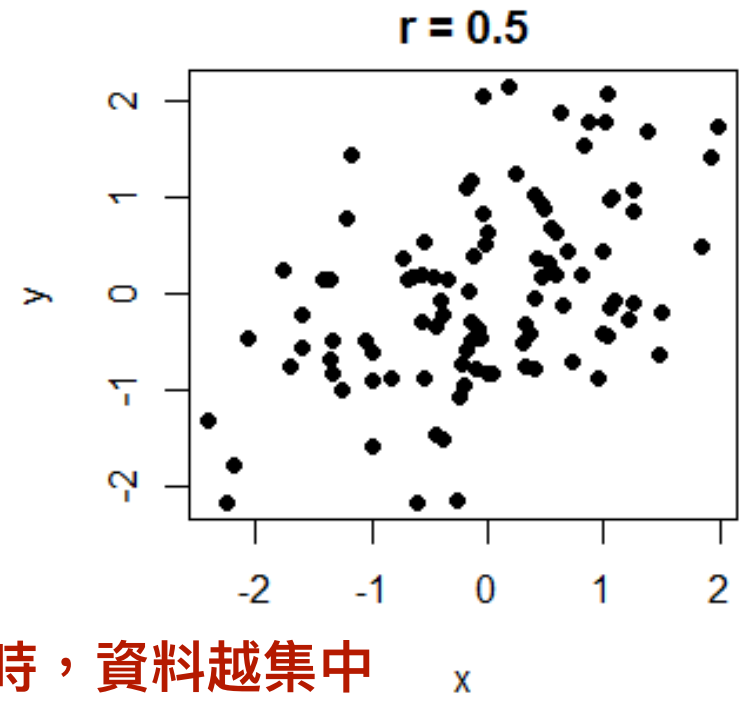
$Var(Z_x \pm Z_y) = Var(Z_x) + Var(Z_y) \pm 2Cov(Z_x, Z_y) = 2 \pm 2r_{x,y}$,

since $Var(Z_x) = 1, Var(Z_y) = 1$.

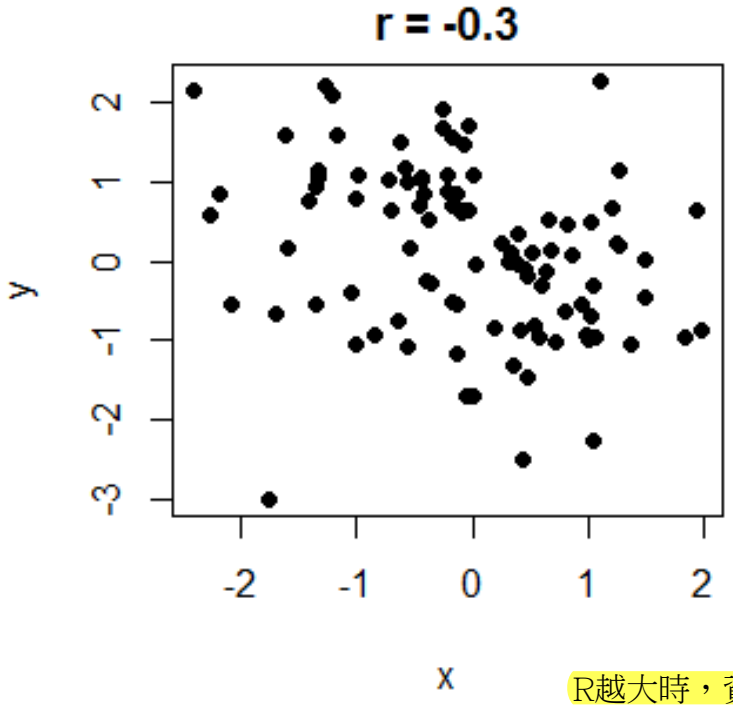
$Var(Z_x \pm Z_y) \geqq 0$, so $2 \pm 2r_{x,y} \geqq 0$

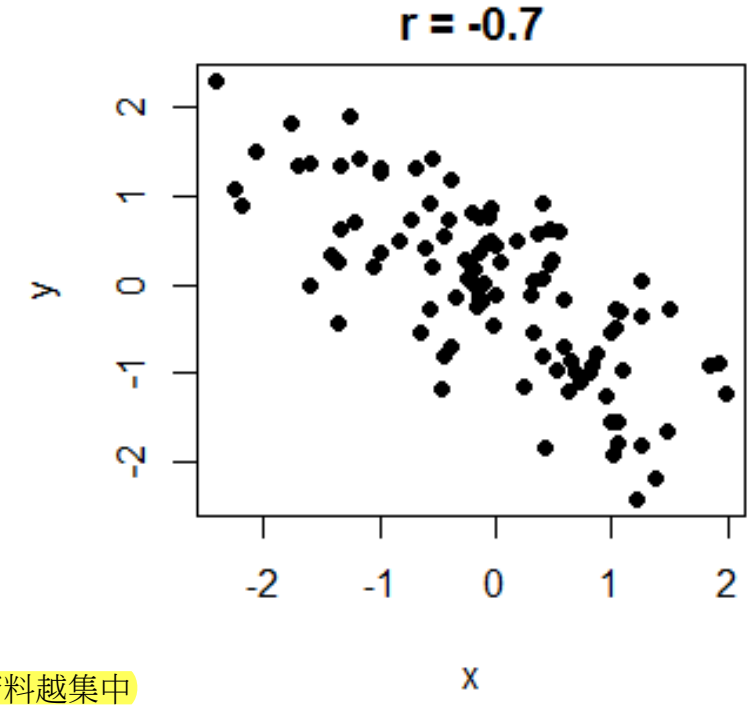
Therefore, $-1 \leqq r_{x,y} \leqq 1$





R越大時，資料越集中

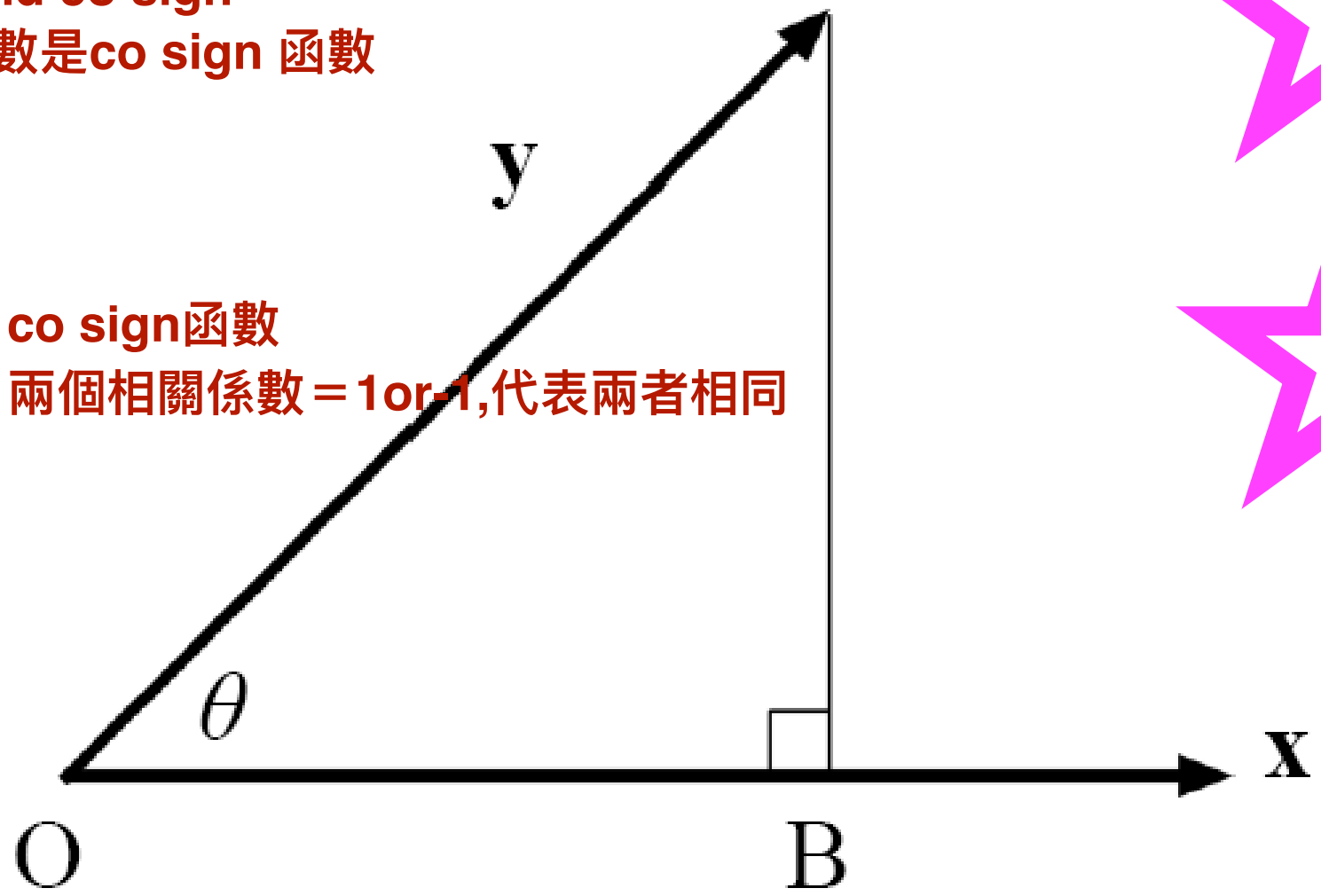




R越大時，資料越集中

Vector Geometry of Correlation

範圍在－1~1,有包含的2個三角函數，正旋與負
sign and co sign
相關係數是co sign 函數



co sign函數
兩個相關係數＝1or-1,代表兩者相同

- *r* 的信賴區間(confidence interval)需利用Fisher's *z* transformation才能得到：

$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) \qquad r = \frac{\exp(2z_r)-1}{\exp(2z_r)+1}$$

轉換成負無限大～無限大計算，再轉回R

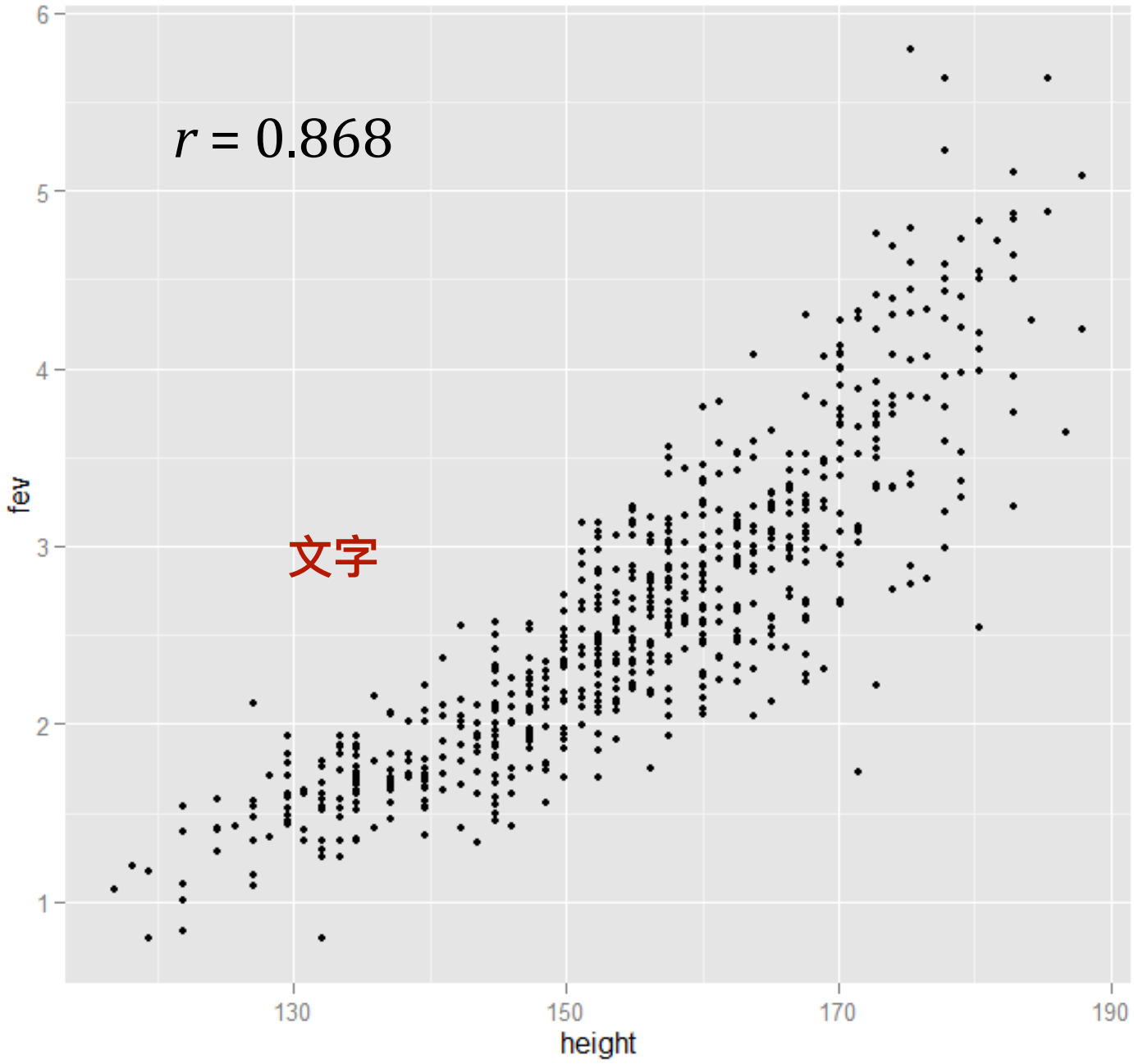
- The standard error of the transformed correlation z_r is approximately $\frac{1}{\sqrt{n-3}}$, so a 95% confidence interval for z_r is:

$$95\%\ CI = z_r - 1.96/\sqrt{n-3} \text{ to } z_r + 1.96/\sqrt{n-3}$$

Example

FEV	Height
1.708	144.78
1.724	171.45
1.72	138.43
1.558	134.62
1.895	144.78
2.336	154.94
1.919	147.32
1.415	142.24
1.987	148.59
1.942	152.4
…	…

FEV: Forced Expiratory Volume



- First transform *r* to z_r using Fisher's transformation:

$$Z_{0.868} = \frac{1}{2}\log_e\left(\frac{1+0.868}{1-0.868}\right) = 1.325$$

$$95\%\ CI = z_{0.868} - 1.96/\sqrt{654-3} \text{ to } z_r + 1.96/\sqrt{654-3}$$
$$= 1.248 \text{ to } 1.402$$

- Now we transform z_r back to *r*:

$$95\%\ CI\ for\ r = 0.848 \text{ to } 0.886$$

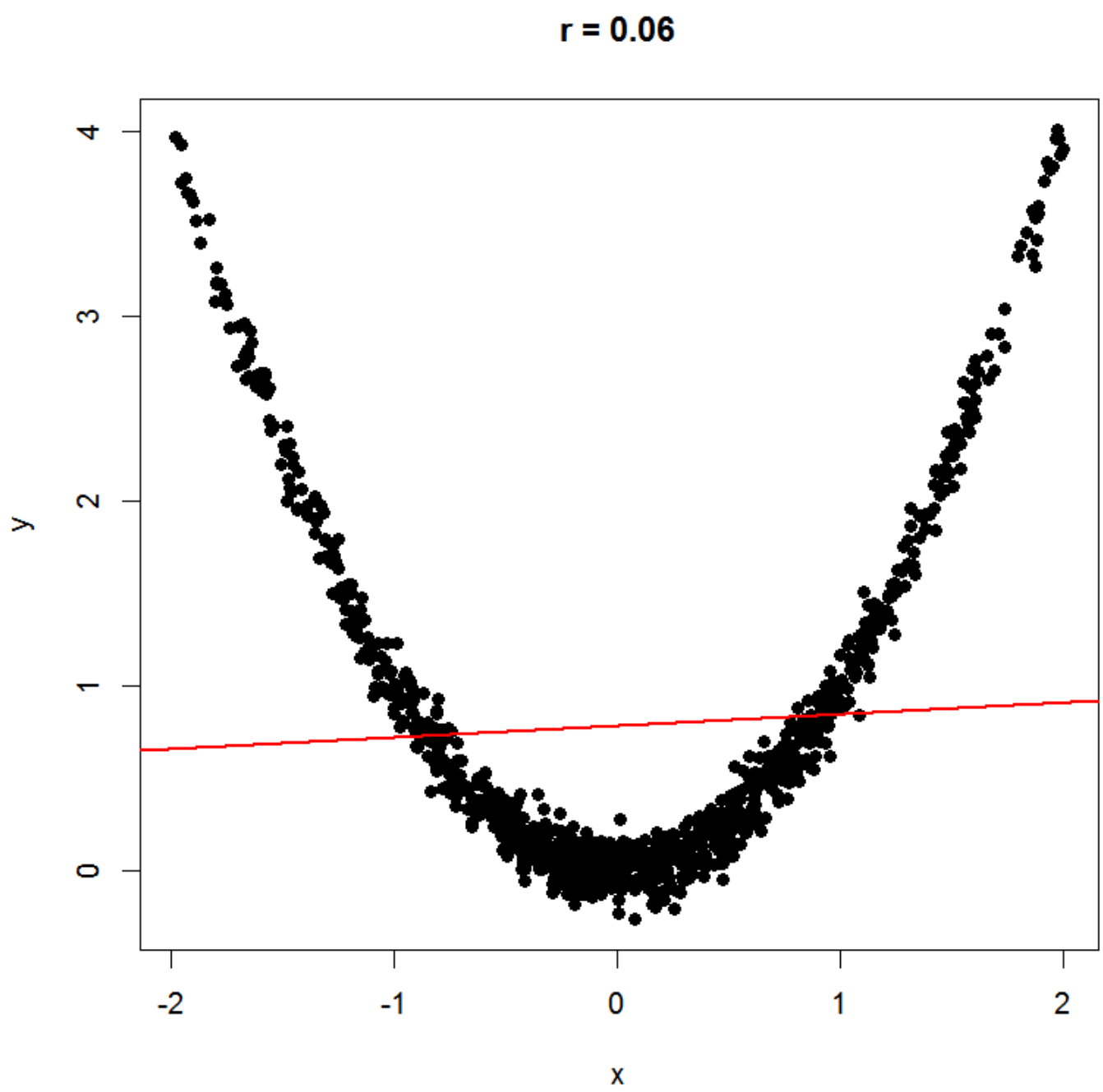
0.886已經很靠近1,為不對稱的

Warnings

相關系數再高也不代表兩者有因果關係

- **相關性不等同於因果關係！**
  - 我們所觀察到的兩個現象或變數的相關性， 有可能是受到背後的一個共同的潛在因素的影響
  - 例如：有人發現一場火災造成的財產損失和趕來現場的救火隊員人數成極強的正相關，但這不表示火災造成的財產損失是救火隊員造成的！
- **相關係數是用來測量兩個變數的線性(linear)關係**
  - 兩個現象或變數可能有很強的因果關係，卻沒有相關性(correlation)，也就是它們的相關係數可以為零。

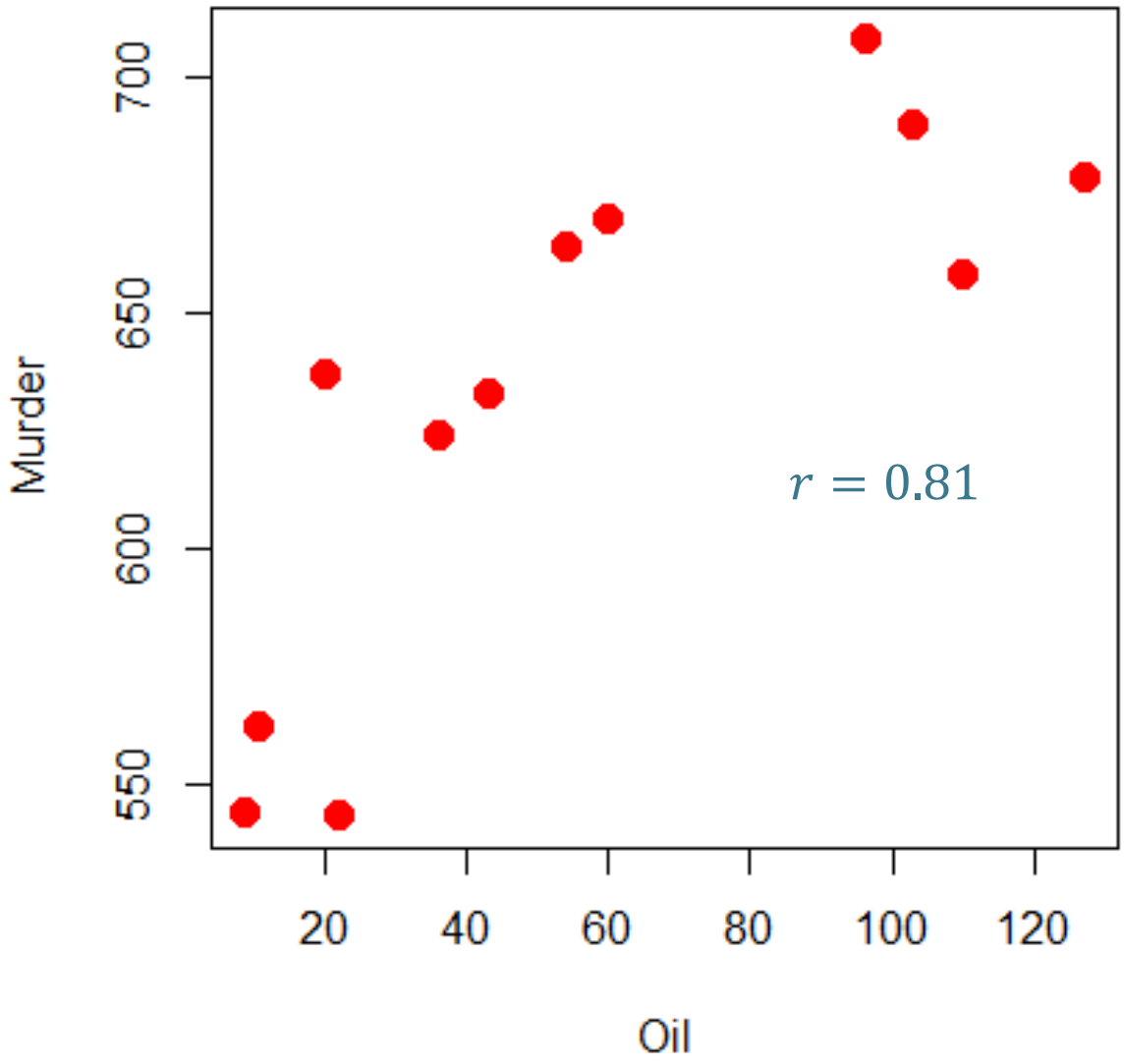
兩個相關係數等於0時，不代表兩者沒有關係


r = 0.06
y
x
4
3
2
1
0
-2
-1
0
1
2

Spurious Correlation

US crude oil imports from Norway correlates with Murders by hanging, strangulation and suffocation between 1999 and 2010

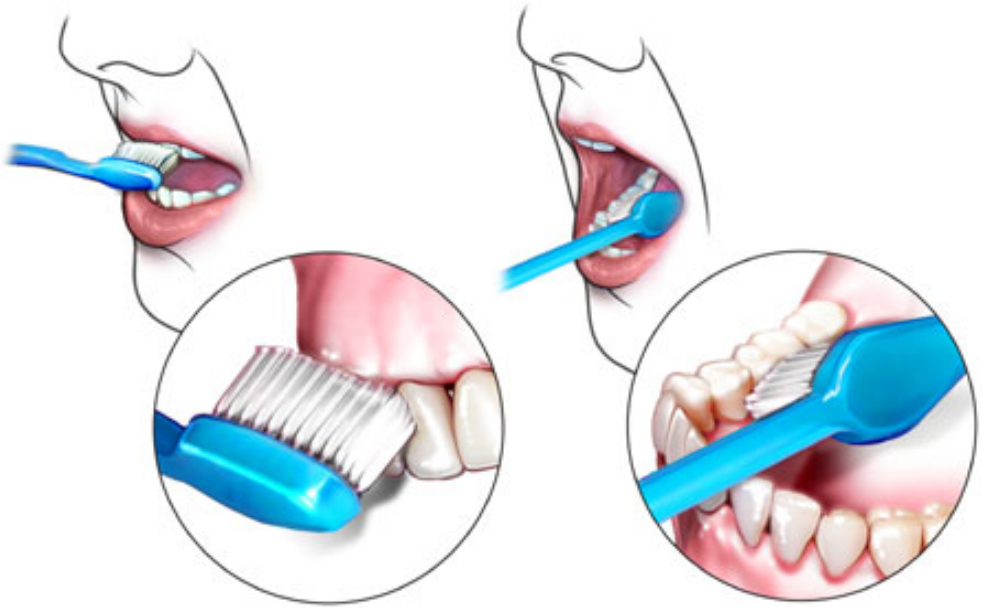
Oil (10^6 $barrels$)	Murder
96	708
110	658
103	690
127	679
60	670
54	664
43	633
36	624
20	637
11	562
22	543
9	544



Ordinary Least Squares Regression

假設我們測量了一群來看牙科門診病人的牙菌斑指數和刷牙的技巧，我們想知道：

- **刷牙的技巧是否會影響牙菌斑指數**(plaque index，一種測量牙齒有沒有刷乾淨的指標)?
- 牙菌斑指數的範圍從 0 (牙齒表面非常乾淨)到 100%(牙齒表面都是牙菌斑) 。 刷牙技巧指數的範圍也是從 0 (完全不會刷牙)到 100%(專家級) 。那我們要**如何預測一個刷牙技巧為 20% 的病人，她的牙菌斑指數大概會是多高**？

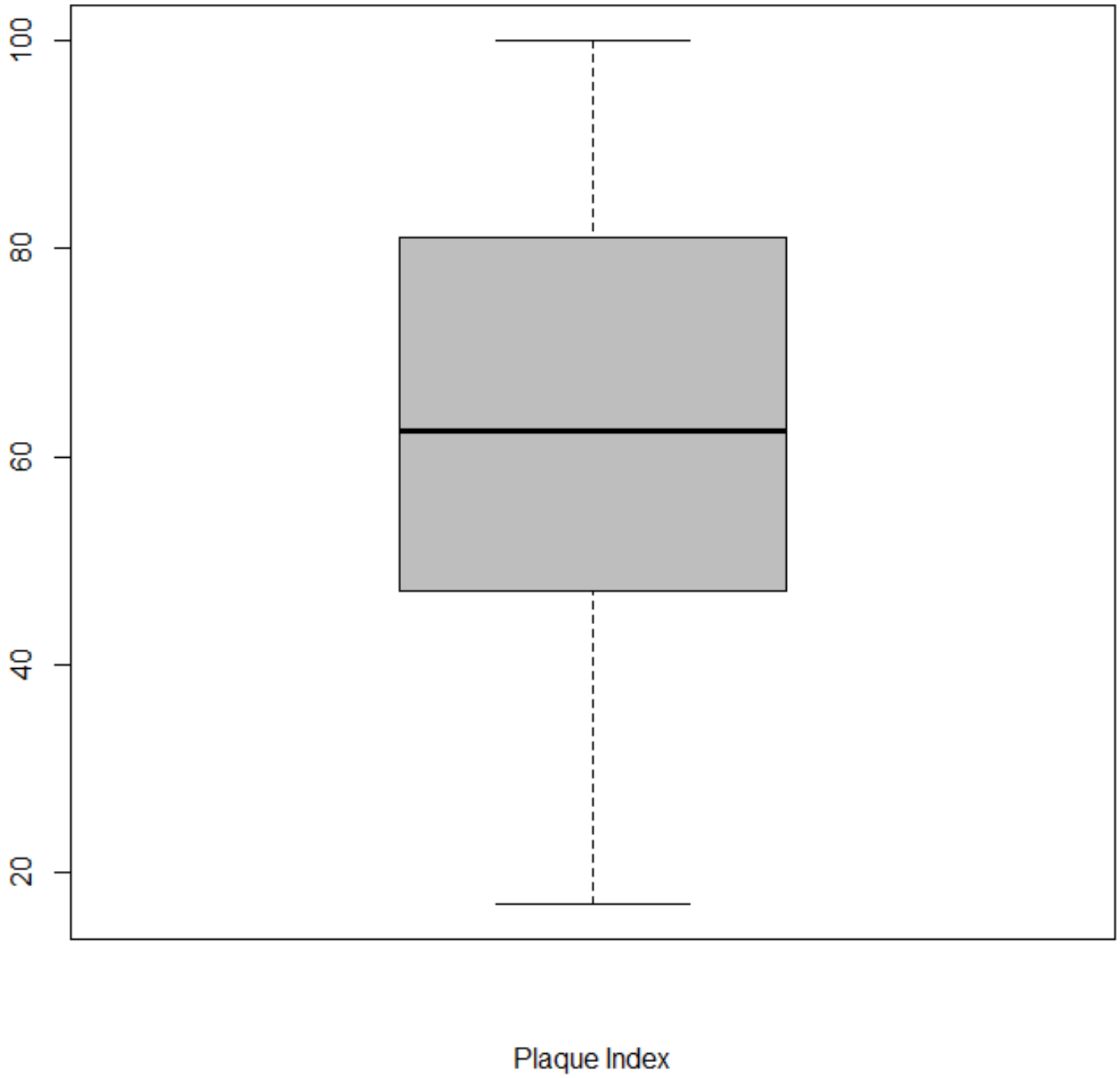


id	bqualit	pindex	recentb	howrecen
1	0	100	yes	45
2	10	81	yes	5
3	10	97	yes	39
4	12	98	no	180
5	14	50	yes	4
6	17	48	yes	37
7	22	62	yes	80
8	23	63	yes	51
9	28	83	no	195
10	28	60	yes	90
11	28	47	yes	10
12	29	65	yes	50
13	35	55	yes	45
14	36	70	no	190
15	39	82	yes	65
16	41	78	yes	105
17	41	85	yes	25
18	51	67	yes	55
19	51	72	no	180
20	57	26	yes	75
21	60	47	no	240
22	62	81	no	135
23	67	82	no	210
24	75	53	no	190
25	79	31	yes	60
26	80	61	no	240
27	87	24	yes	60
28	89	43	no	150
29	91	30	yes	45
30	92	17	yes	75

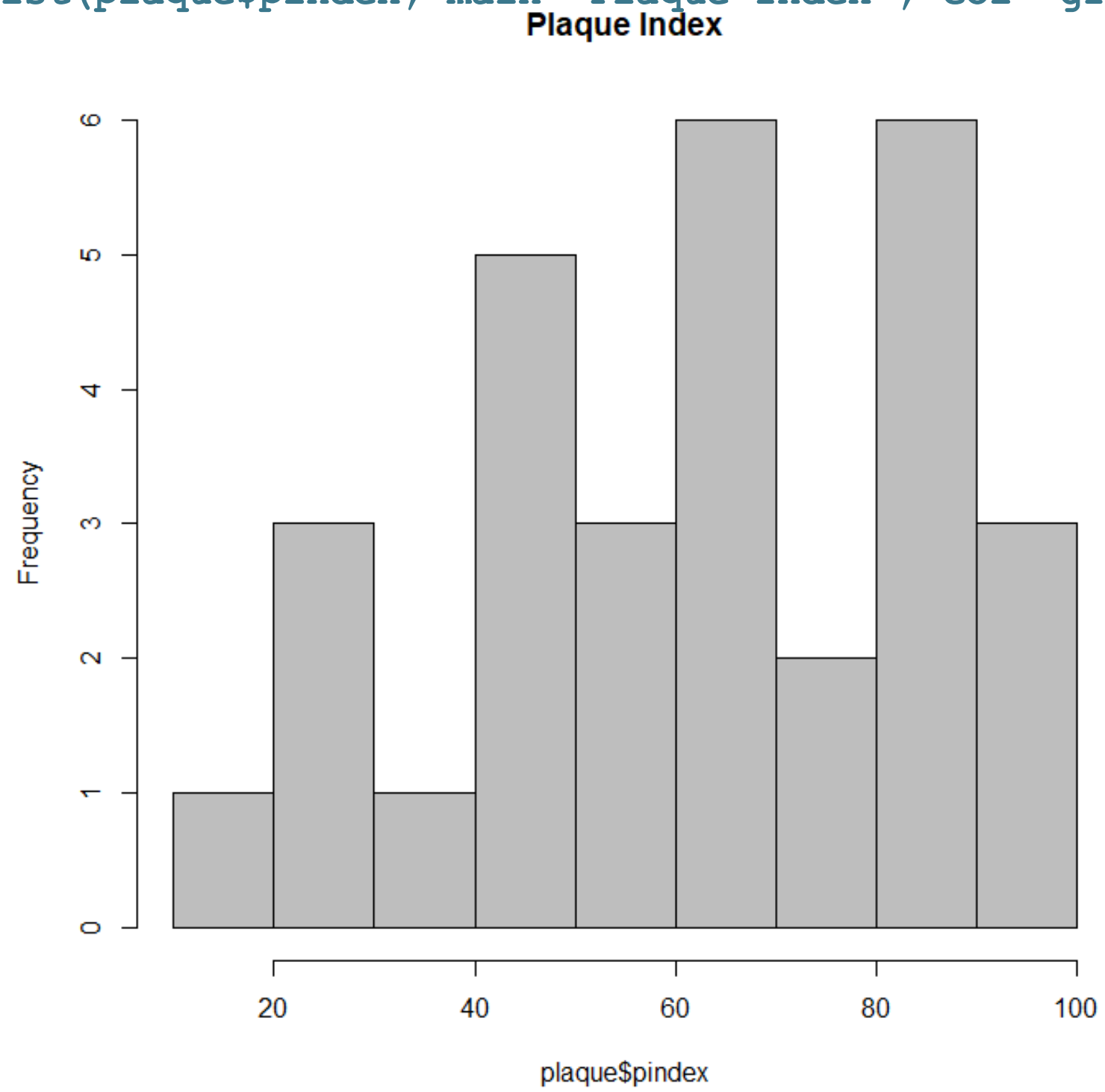
Summary Statistics

```
> summary(plaque)
       id              bqualit          pindex        recentb     howrecen
 Min.   : 1.00   Min.   : 0.00   Min.   : 17.00   no :10   Min.   :  4.0
 1st Qu.: 8.25   1st Qu.:24.25   1st Qu.: 47.25   yes:20   1st Qu.: 45.0
 Median :15.50   Median :40.00   Median : 62.50            Median : 70.0
 Mean   :15.50   Mean   :45.13   Mean   : 61.93            Mean   : 97.7
 3rd Qu.:22.75   3rd Qu.:65.75   3rd Qu.: 81.00            3rd Qu.:172.5
 Max.   :30.00   Max.   :92.00   Max.   :100.00            Max.   :240.0
```

```
boxplot(plaque$pindex, xlab="Plaque Index", col="grey")
```



```
hist(plaque$pindex, main="Plaque Index", col="grey")
```



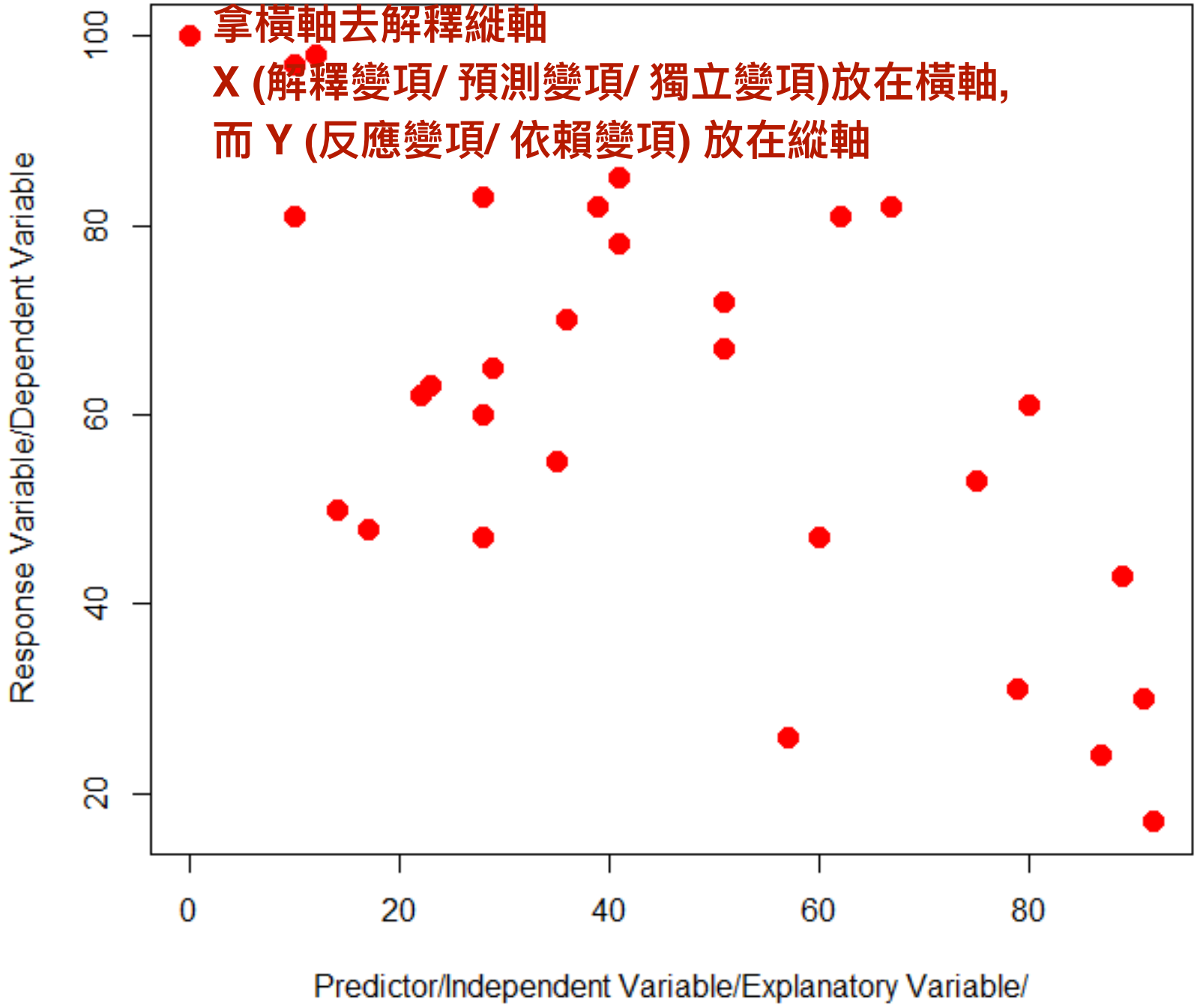
- 回答這些問題要用迴歸分析(regression analysis)
- 相關係數只回答兩個變數是否有關聯(association)
- 但是我們通常還想要**描述**兩個變數間的數量關係，也想用其中一個變數的變化來**預測**另一個變數的變化
- 如果有理由相信其中一個變數的變化會**導致**另一個變數的變化，也用迴歸分析

左上到右下＝負向變項

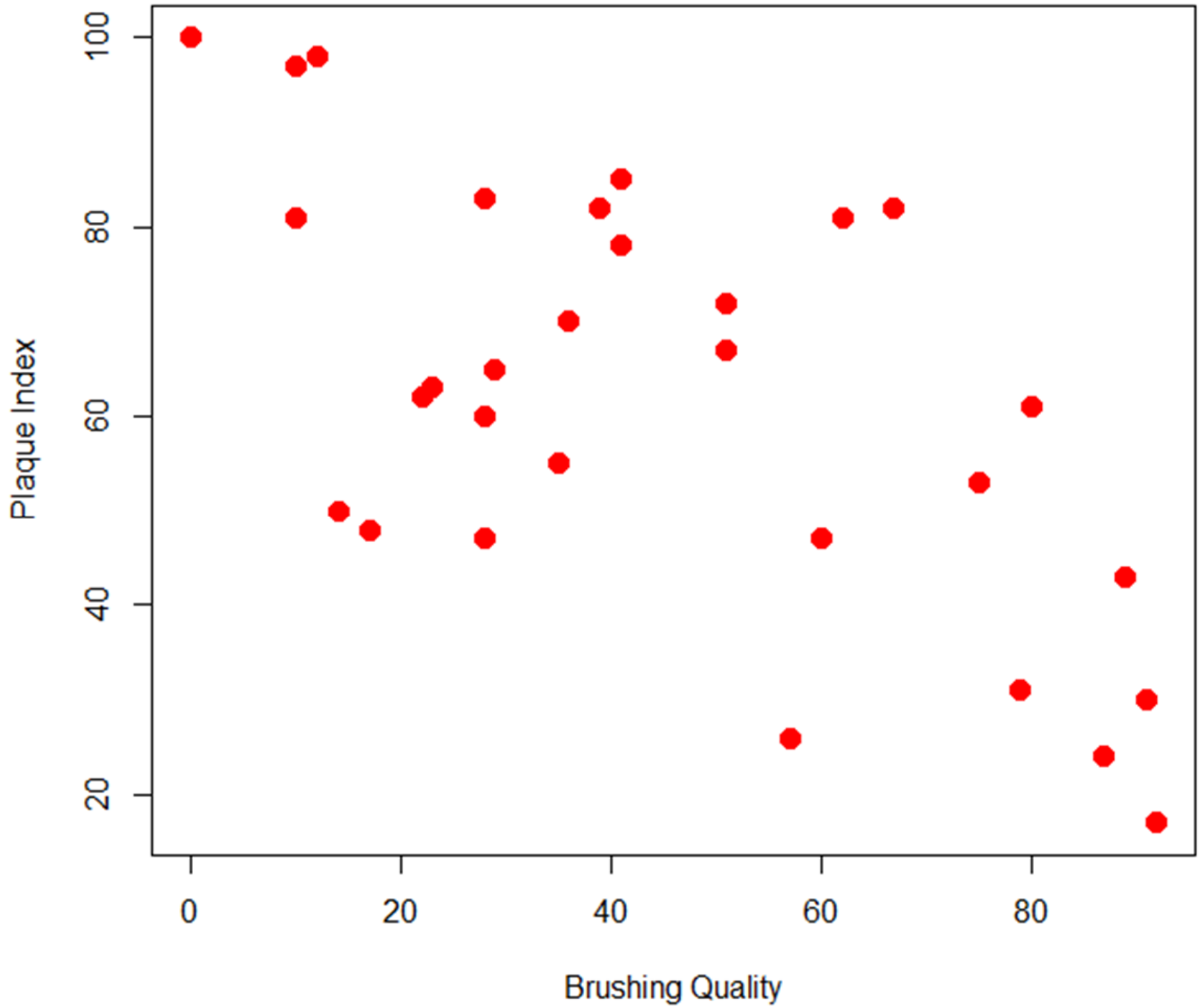
拿橫軸去解釋縱軸

X (解釋變項/ 預測變項/ 獨立變項)放在橫軸,

而 Y (反應變項/ 依賴變項) 放在縱軸

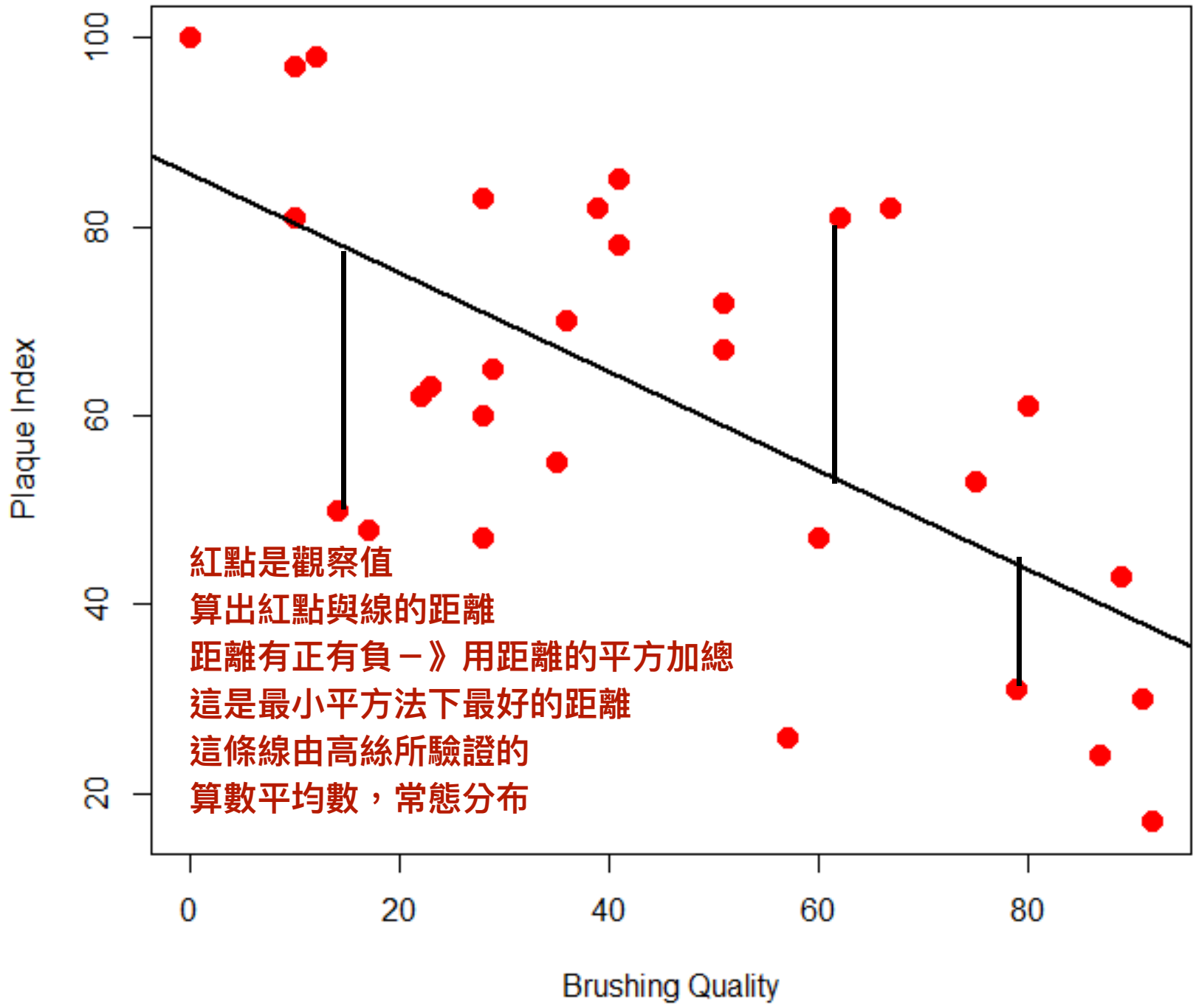


- X (解釋變項 / 預測變項 / 獨立變項) 放在橫軸，而 Y (反應變項 / 依賴變項) 放在縱軸
- We try to fit the “best” straight line 我們將試圖找出一條最好的直線來描述 X 和 Y 的關係
- 如果X 和 Y 的關係真的是線性的，那這條最好的直線將提供利用 X 值來預測 Y 值最佳的結果
- 大家可以自己用筆和尺試看看


Plaque Index
100
80
60
40
20
0
20
40
60
80
Brushing Quality

標準的作法是用**最小平方法** (least squares)

這個方法是找出一條最好的直線讓 Y 的觀察值和它在這條線的**垂直距離的平方和**為最小。


Plaque Index
100
80
60
40
20
0
20
40
60
80
Brushing Quality
紅點是觀察值
算出紅點與線的距離
距離有正有負－》用距離的平方加總
這是最小平方法下最好的距離
這條線由高絲所驗證的
算數平均數，常態分布

One can get an equation for the regression line (best fitting line)

$Y = a + b*X$

a = intercept (截距)

b = slope (斜率)

30個點的組合加上線性的限制，只有一個結果
它是唯一的，但不會通過所有的點

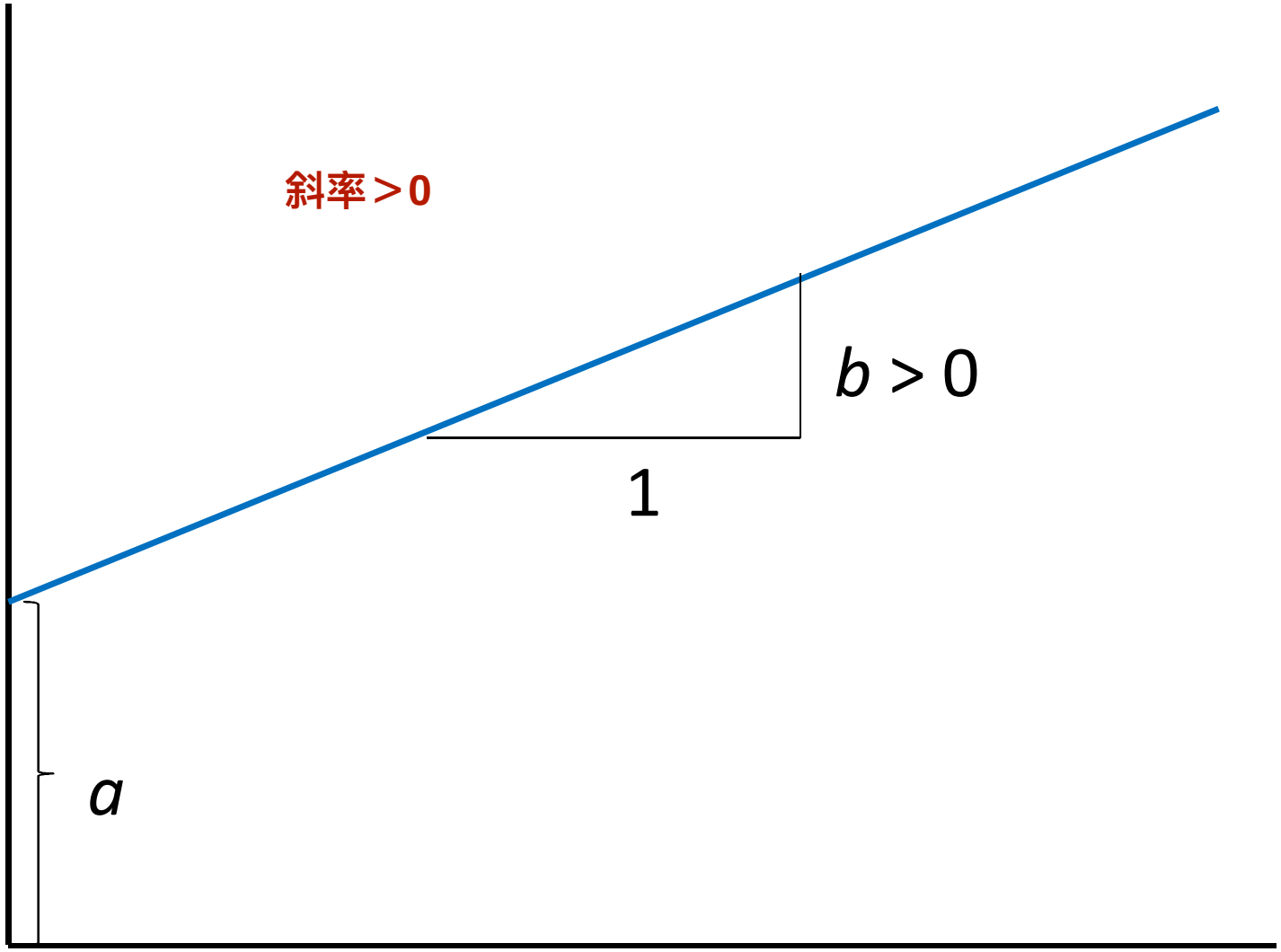
Example:

plaque index = 85.5 - 0.52 x brushing quality

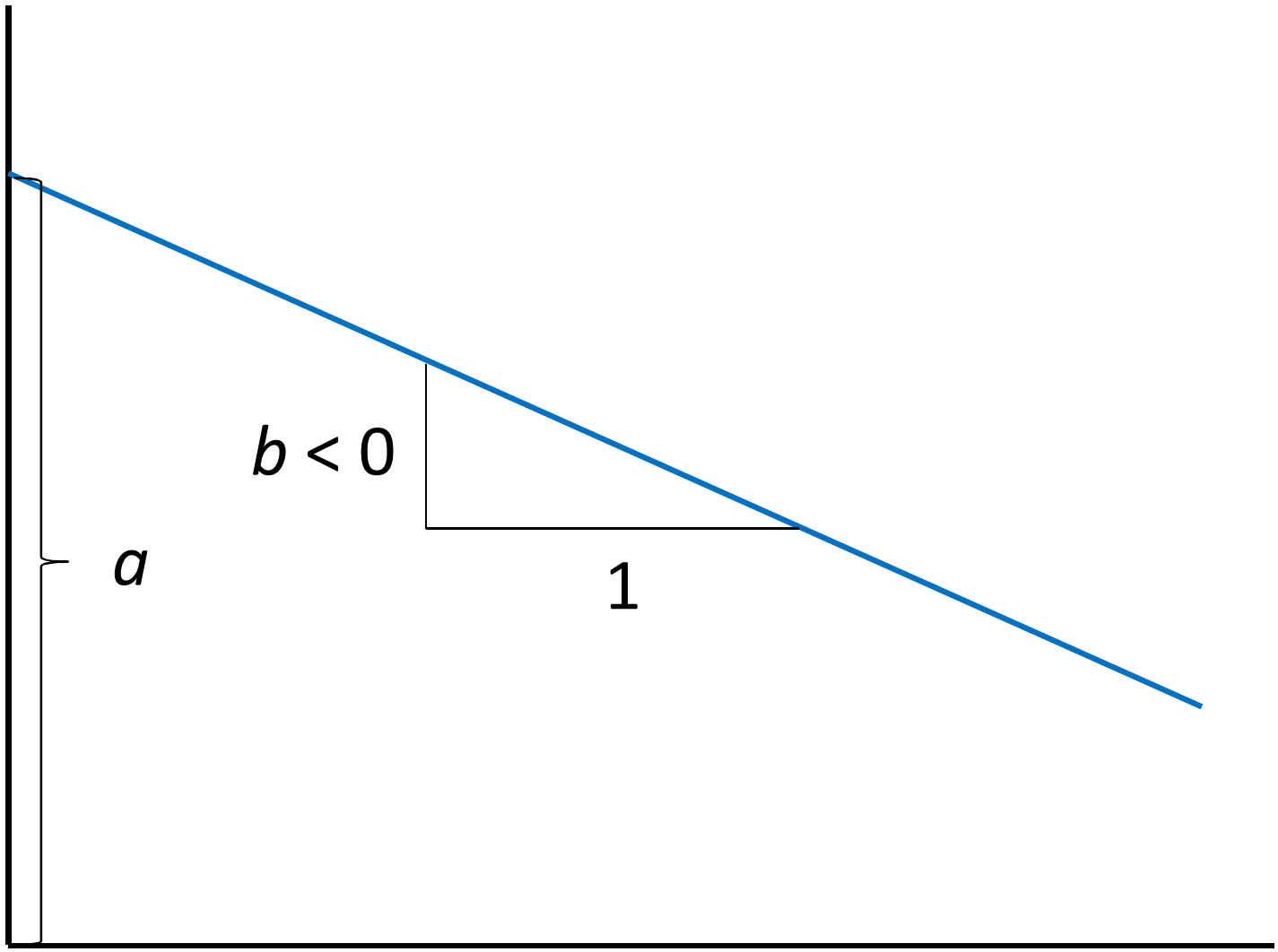
here, a = 85.5 (intercept)

b = -0.52 (slope)

Linear Regression (線性迴歸)



Linear Regression (線性迴歸)



R Output

~代表迴歸

報表的判讀

```
> lm1<-lm(pindex~bqualit,data=scores)
> summary(lm1)

Call:
lm(formula = pindex ~ bqualit, data = scores)

Residuals:
    Min      1Q  Median      3Q     Max
-29.725 -12.185   1.987  14.316  31.507

Coefficients:
    截距       Estimate Std. Error t value Pr(>|t|)
(Intercept)   85.5464     6.2927  13.594 7.44e-14 ***
bqualit       -0.5232     0.1194  -4.382  0.00015 ***
---slope (斜率)
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.8 on 28 degrees of freedom
Multiple R-squared:  0.4068,    Adjusted R-squared:  0.3856
F-statistic:  19.2 on 1 and 28 DF,  p-value: 0.0001499
```

So the equation for the best fitting straight line (regression line) is:

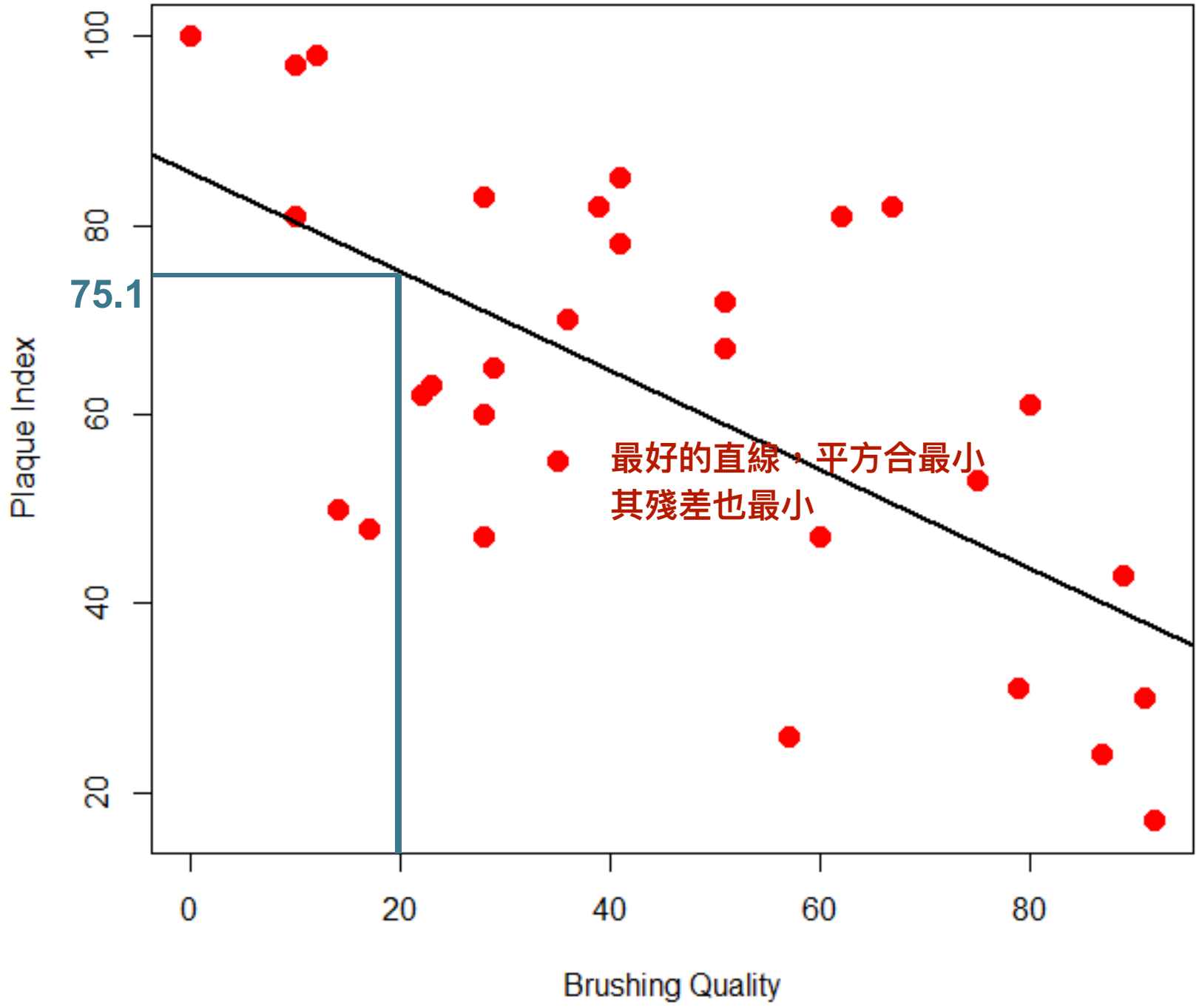
plaque index = 85.5 - 0.52 x brushing quality

注意：雖然這一條“最好的直線”是統計上顯著的，但說不定某種曲線可能更接近這兩個變數之間的關係。

一個刷牙技巧為 20% 的病人，她的牙菌斑指數大概會是多高？

兩者劃出來的線為非直線，代表兩者無關X

85.5 – 0.52*20 = 75.1


Plaque Index
100
80
60
40
20
75.1
0
20
40
60
80
Brushing Quality
最好的直線，平方合最小
其殘差也最小

Ordinary Least Squares Regression

- For $y = a + bx + e$, OLS regression tries to minimize the error sum of squares:

$$\sum_{i=1}^{n} (Y_i - (a + bX_i))^2$$

- To minimize, we set the partial derivatives equal to zero:

$$2\sum(Y_i - a - bX_i)(-1) = 0$$

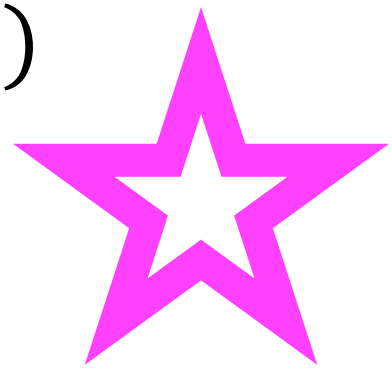
$$2\sum(Y_i - a - bX_i)(-X_i) = 0$$

Estimation of Parameters

- Solve for *a* and *b*, we get:

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{Cov(x,y)}{Var(x)}$$



co balance越大，斜率越大，越偏離0
但co balance 會受單位影響，係數也會跟著膨脹
改變X的單位由公斤變公克，用原本係數除以1000

Relation between Correlation and Regression

- Remember

$$r = Cov\left(\frac{x-\bar{X}}{S_x}, \frac{y-\bar{Y}}{S_y}\right) = \frac{Cov(x,y)}{S_x S_y}$$ S的標準差和Y的標準差

- So $r = b\frac{S_x}{S_y}$

- Then $b = r\frac{S_y}{S_x}$

斜率

如果X和Y的標準差相同，R等於斜率

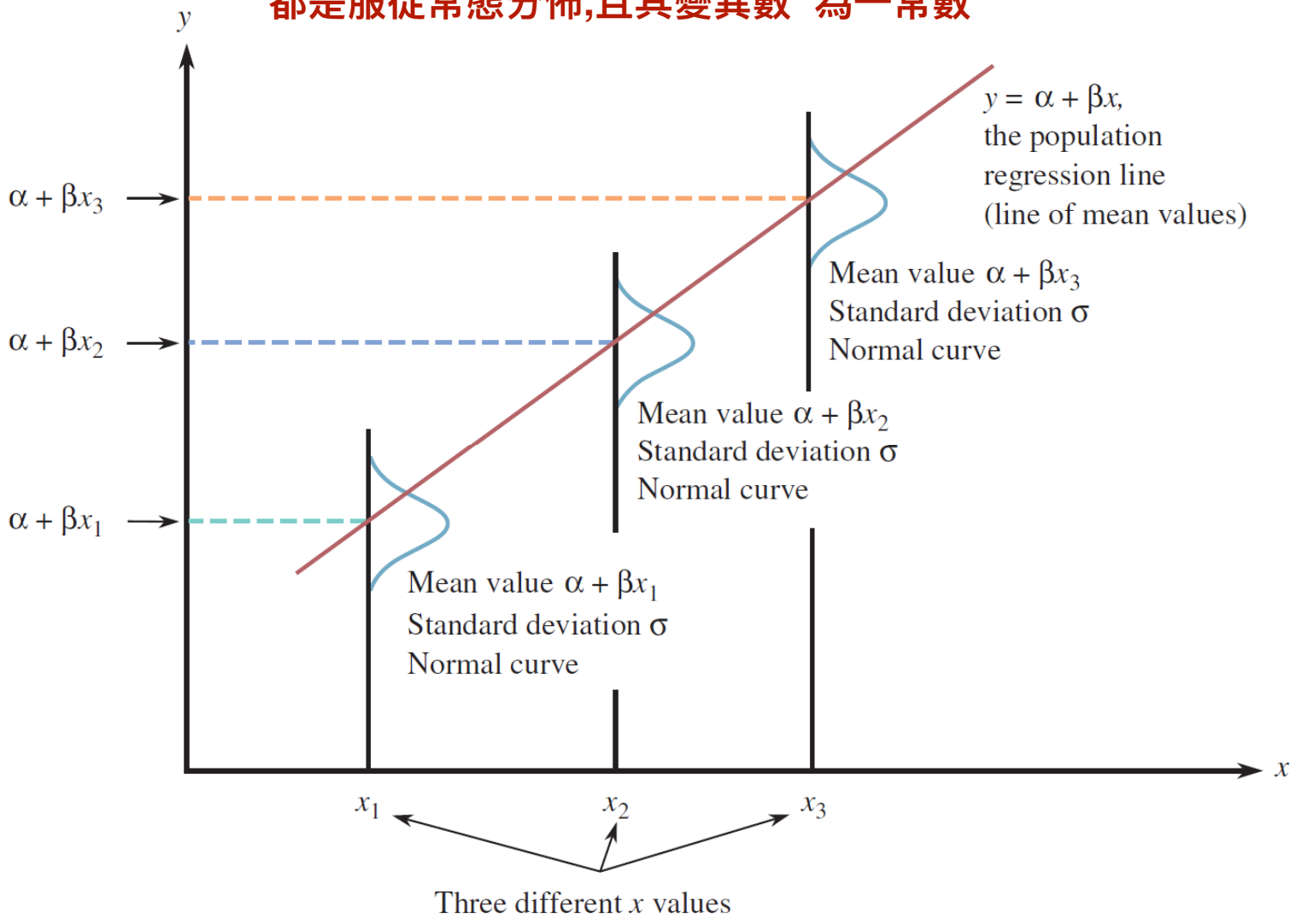
Assumptions for Linear Regression

- 解釋變數 X 和反應變數 Y 之間的關係是直線的(linear)
- 樣本的觀測值皆是獨立的(independent) 。
- 反應變數 Y 在解釋變數 X 的任何一個數值之下，都是服從常態分佈，且其變異數σ^2為一常數。
- X 的觀測值是正確的不含測量誤差(measurement error) 。 **注意： 我們對 X 的分佈並不做任何假設**。

文字

變異數σ^2為常數

反應變數 Y 在解釋變數 X 的任何一個數值之下,
都是服從常態分佈,且其變異數 為一常數



Residuals（殘差）

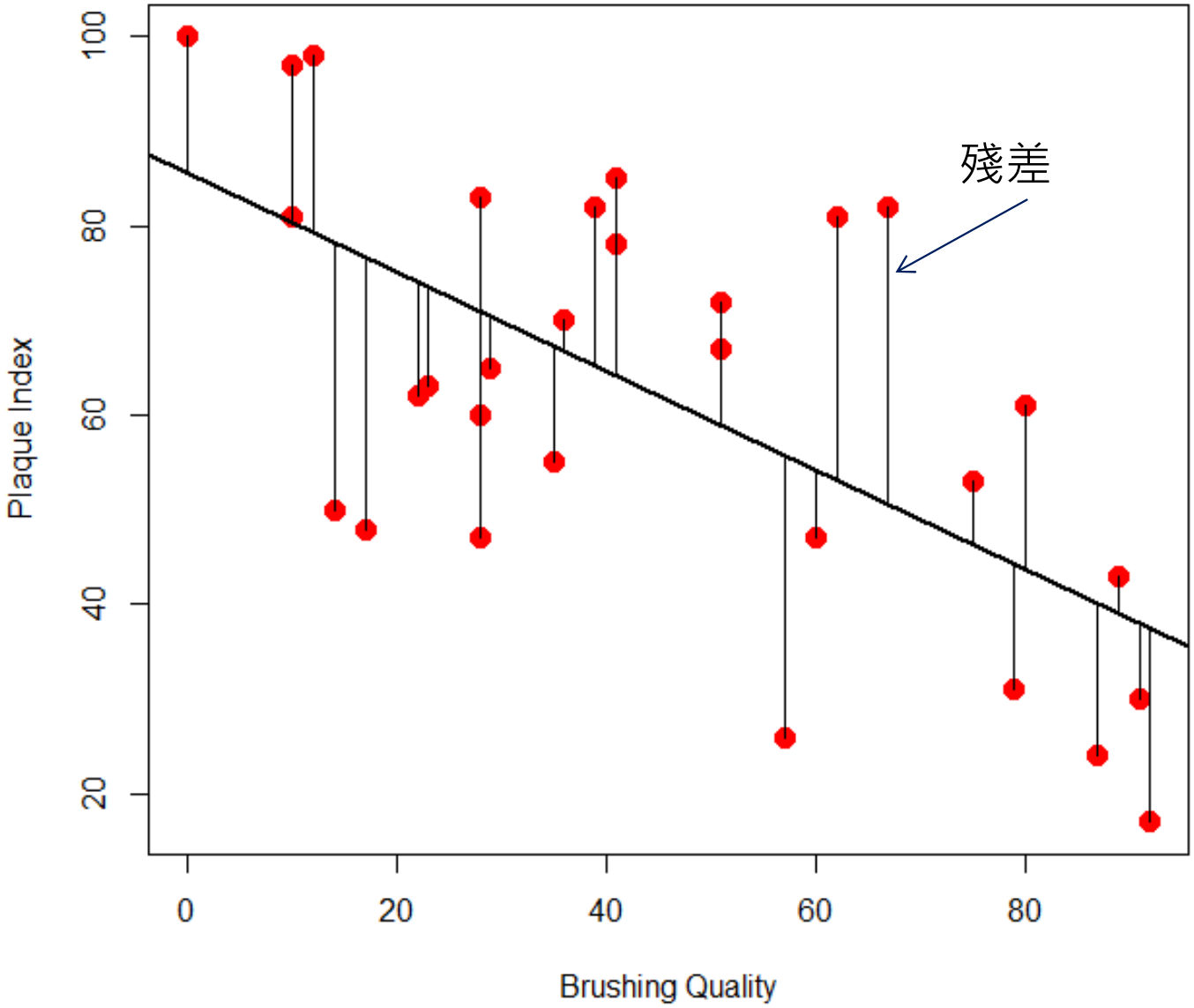
殘差 = 觀測值 (observed values) − 估計值 (fitted values)

殘差平方和 (Residual sum of squares) 定義為：

$$SS_{res} = \sum[(Y_i - (a + bX_i)]^2$$

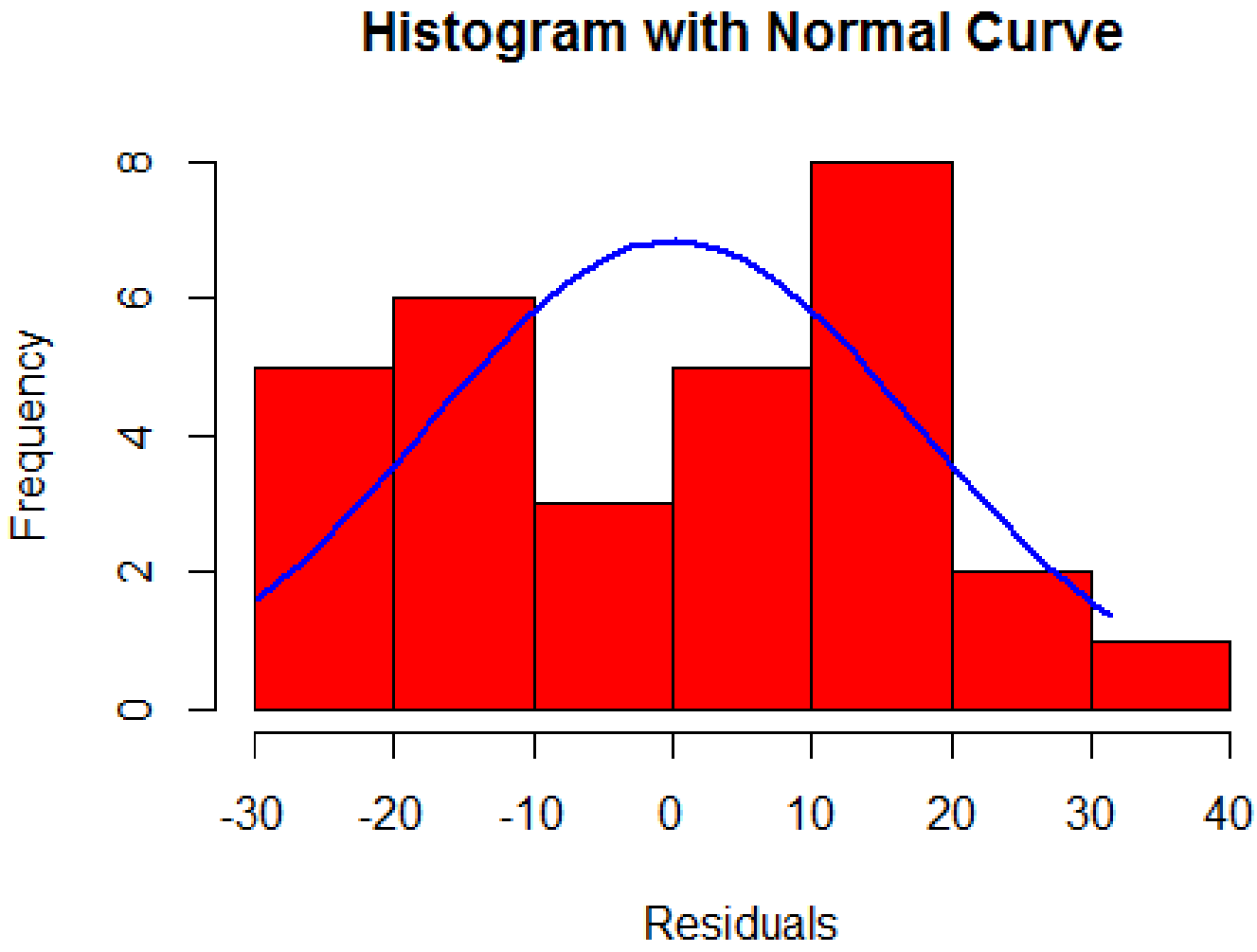
我們可藉由檢查殘差圖 (residual plots) 來檢查迴歸分析模型的一些假設是否成立， 和是否需要考慮在模型中加入其它的變數。

Residuals（殘差）



殘差的直方圖

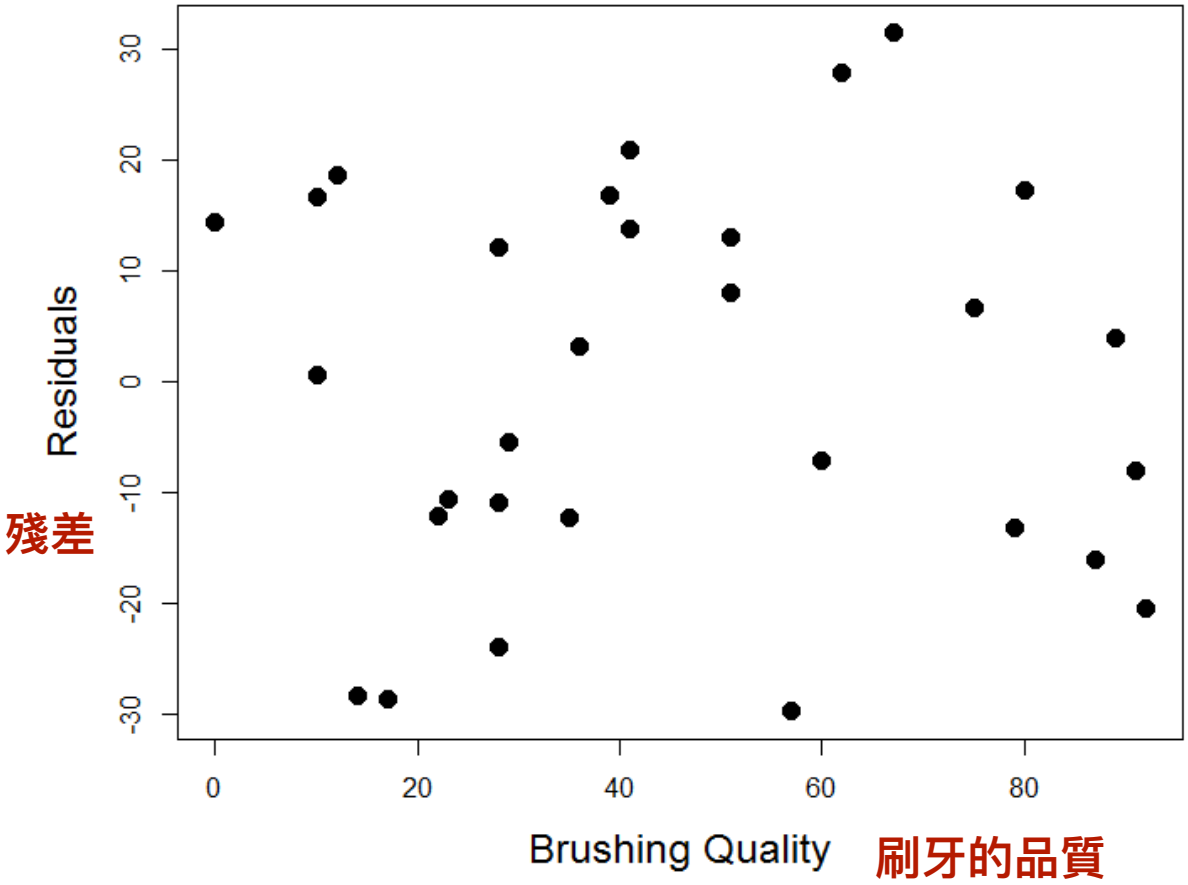
可用來檢查殘差的分佈是否接近常態



殘差對解釋變數殘差圖

可用來檢查

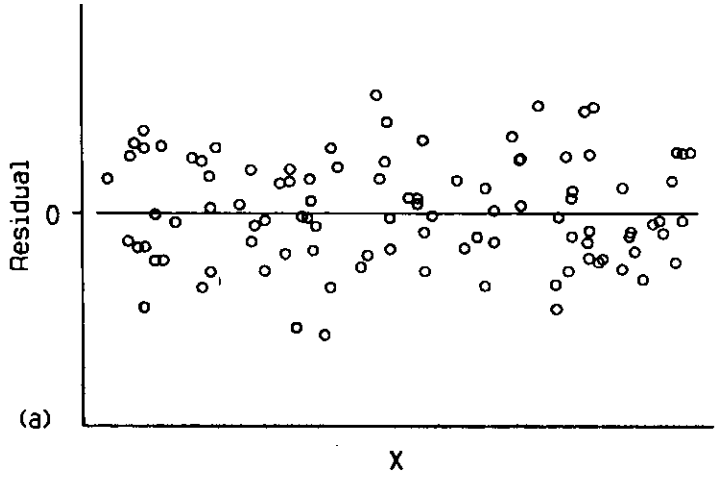
- X 和 Y 之間的關係是否是直線
- Y 在 X 的任何一個數值之下，其變異數σ^2是否為一常數。



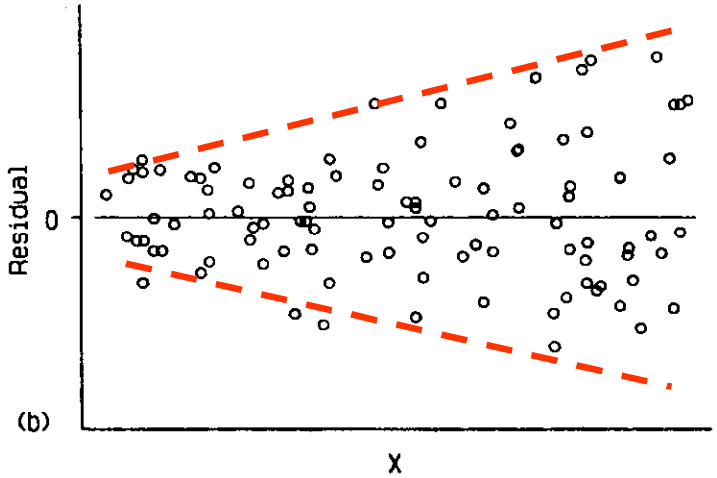
理論上：

- 殘差的分佈在 X 的任何一值皆相同
- 殘差和 X 之間無任何明顯關係

看起來像一把米散落，和Y有相關的X已在線上，剩餘的殘差則和X無相關。

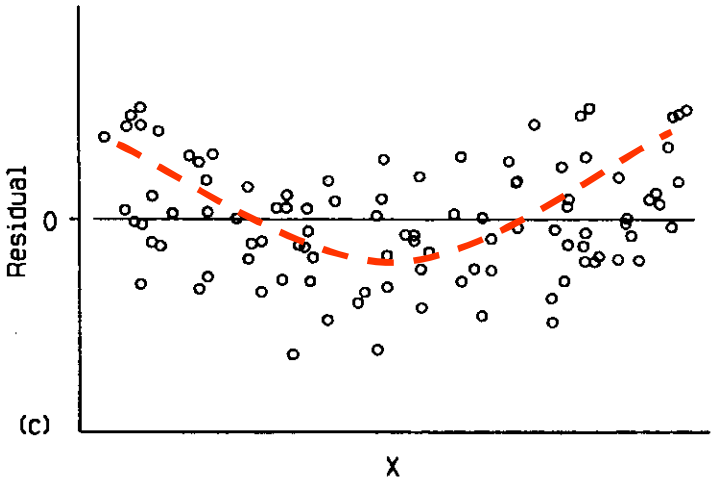


殘差和 X 之間無任 何明顯關係



殘差和 X 之間無任 何明顯關係
但殘差分布隨X越大而越大
身高越高的人其肺活量的變化越大
暗示～～～

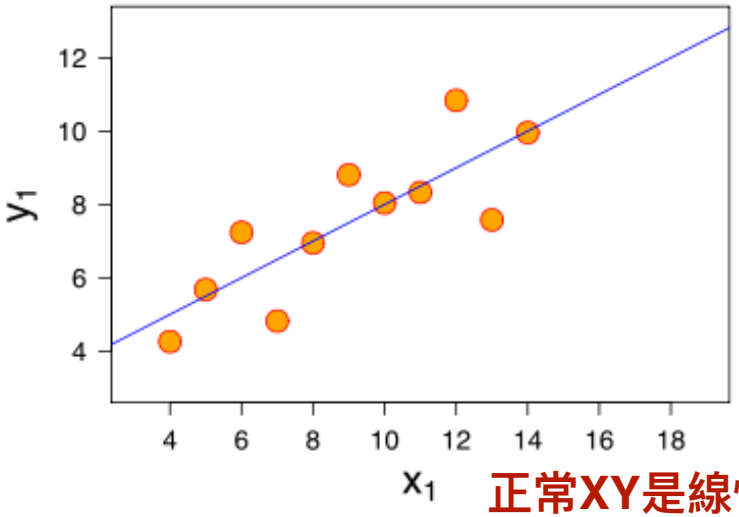
文字



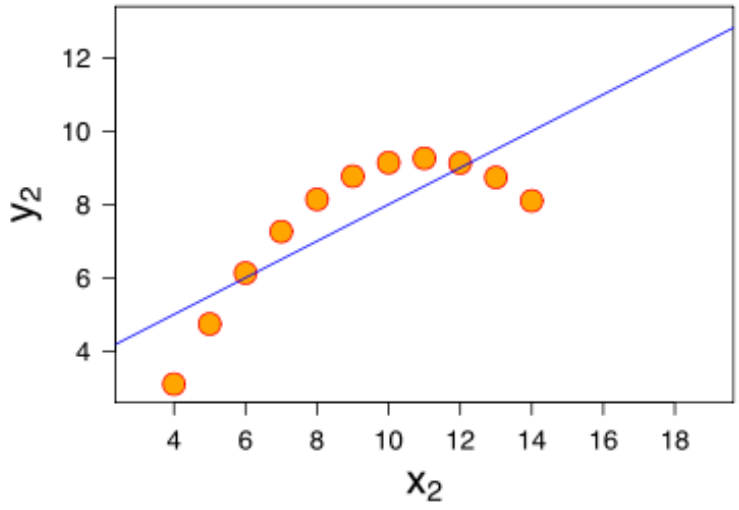
殘差和 X 之間呈U shape關係
代表兩者（XY）非線性關係

Some Examples for Cautions:

4張圖的結果都是 $y = 3.00 + 0.500x$

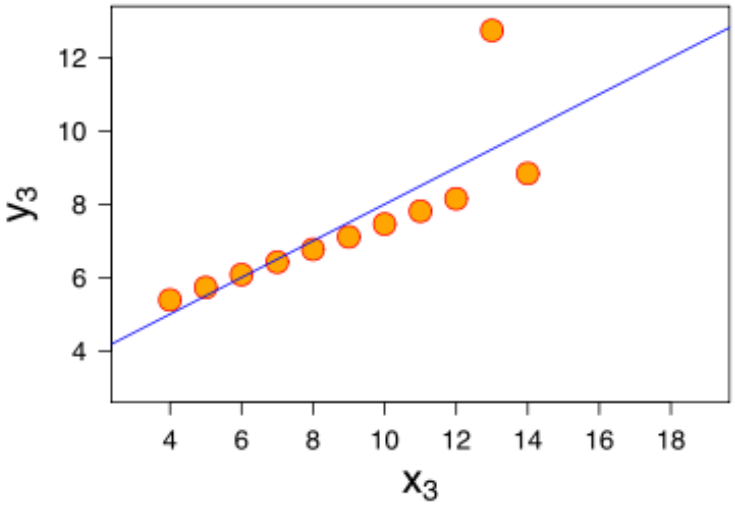


正常XY是線性

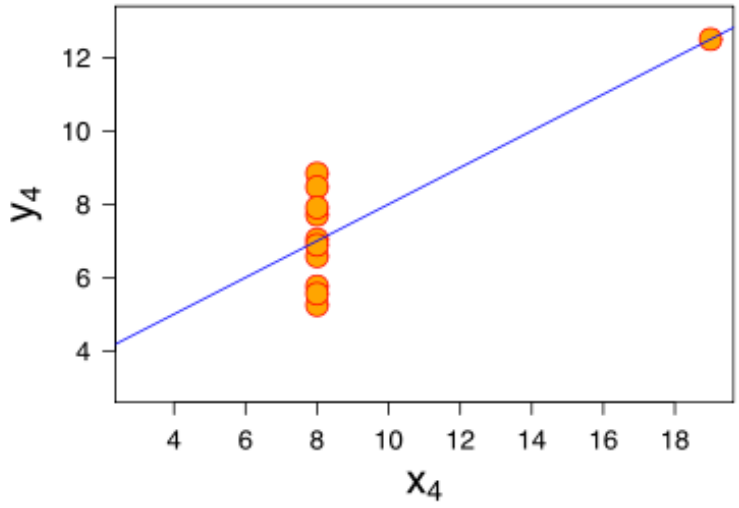


文字

XY非線性，但10過後Y變化了

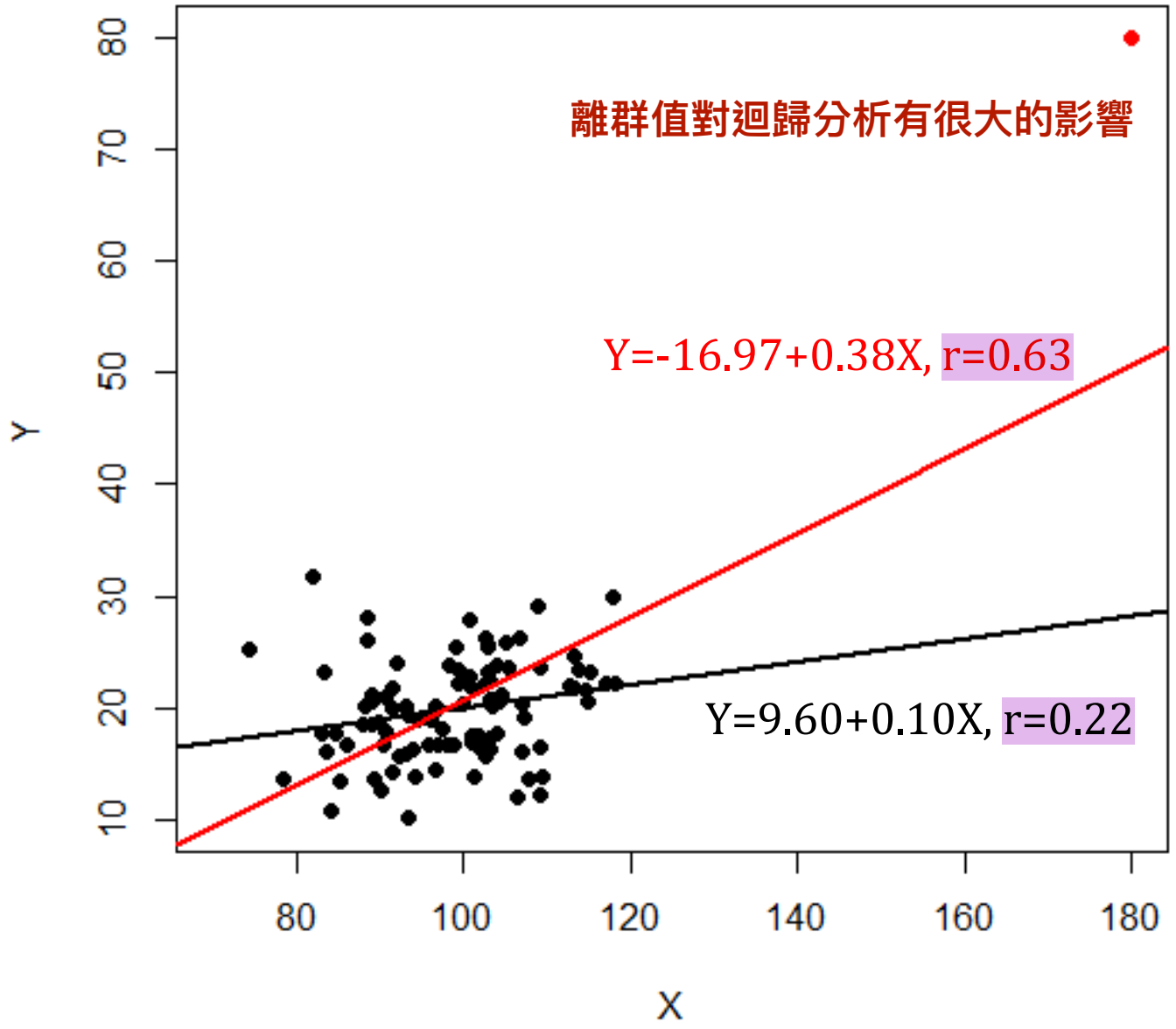


有離群值，或資料KEY錯



X和Y來自不同層次，兩者無相關

Some Examples for Cautions:



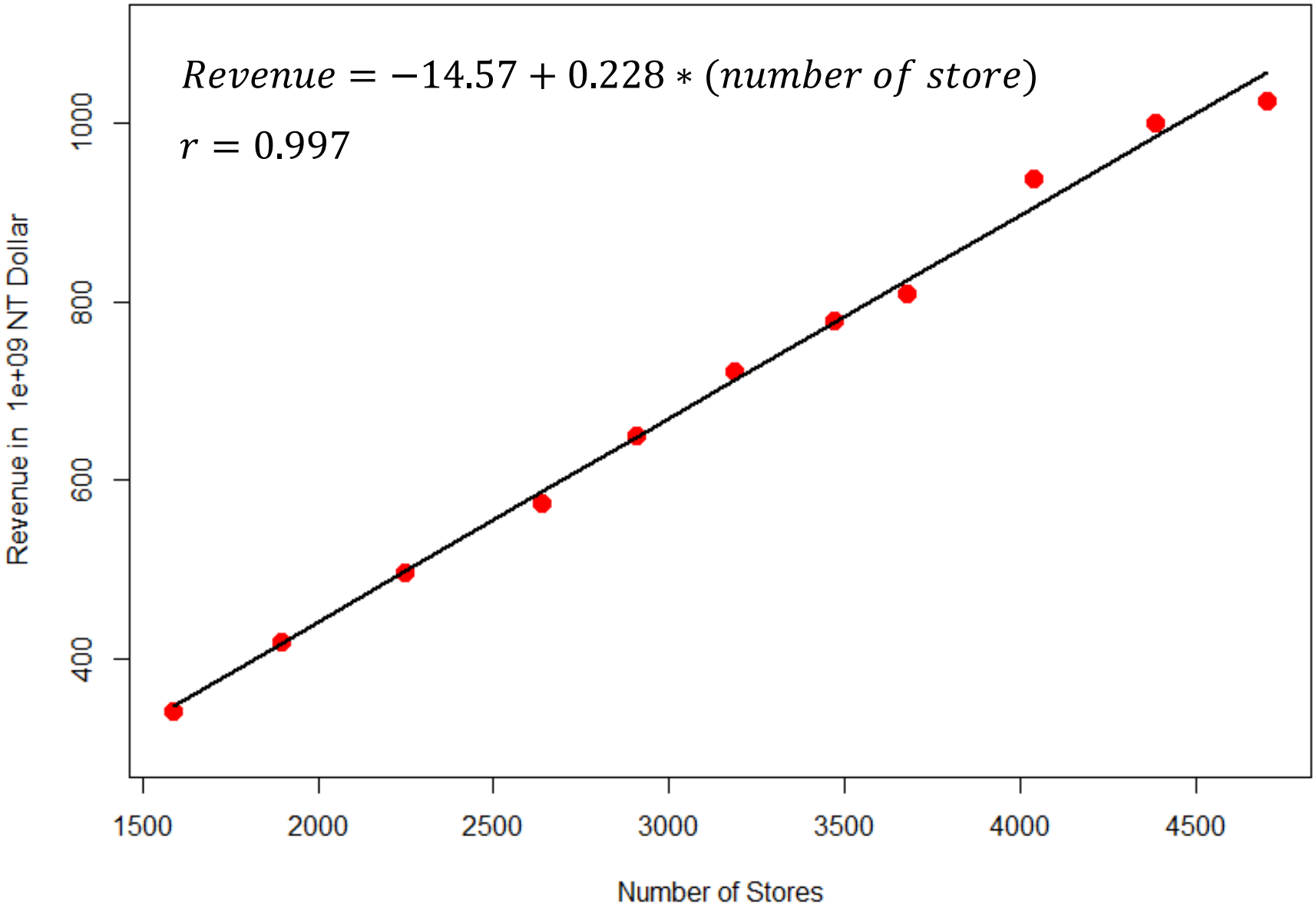
迴歸回析用現有資料來預測未來

Extrapolation: 門市數目與營業收入

統一超商門市數目的成長速度跟統一超商的營業收入是否有關係？(取自台大數學系陳宏教授統計與生活第一版125-127頁)

年份(西元)	門市數目(間)	營業收入(億元)
1997	1588	341.9
1998	1896	419.5
1999	2248	497.2
2000	2638	574.8
2001	2908	649.9
2002	3187	721.9
2003	3470	778.6
2004	3680	809.4
2005	4037	936.7
2006	4385	999.8
2007	4705	1023.6

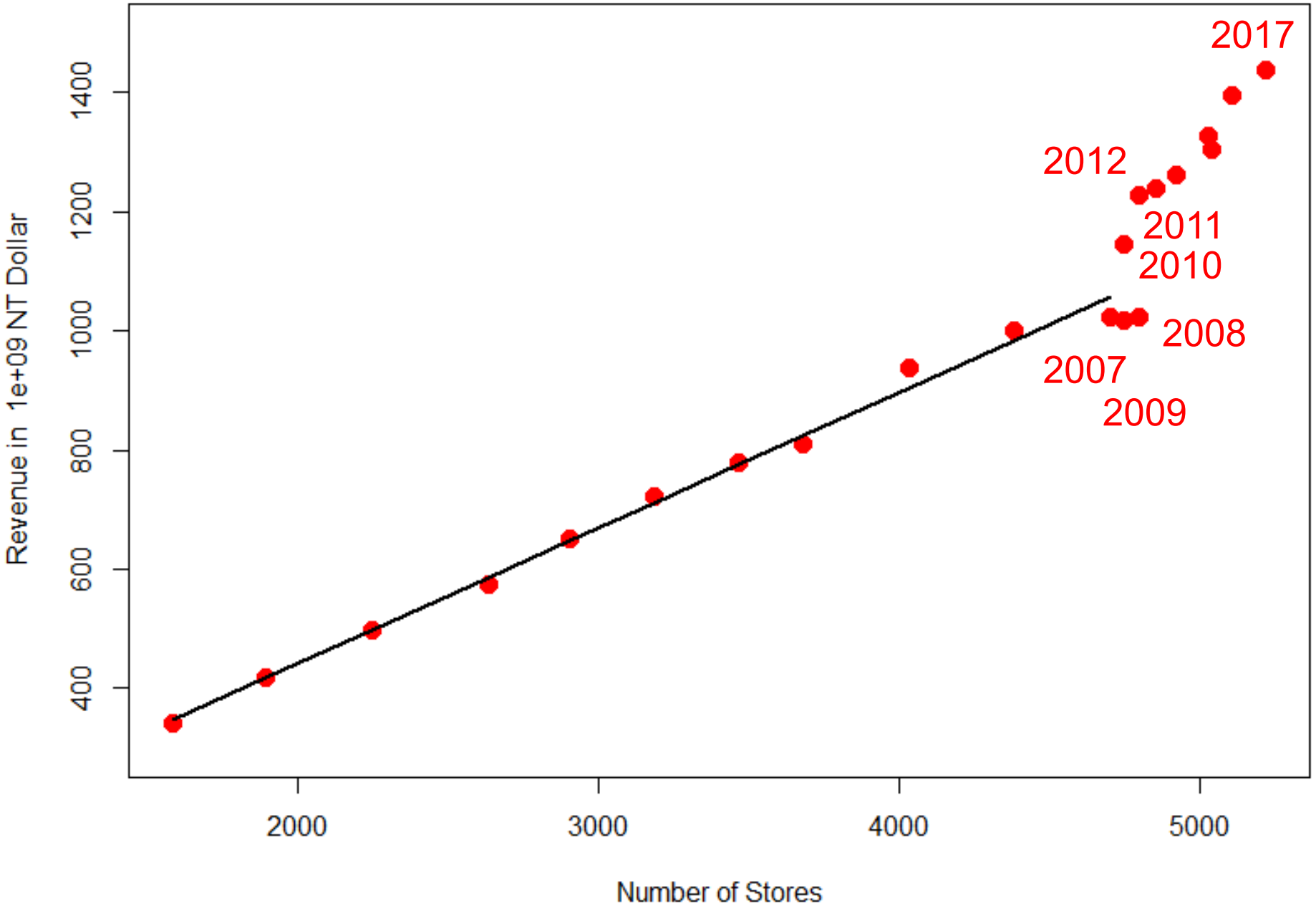
Extrapolation : 門市數目與營業收入



Extrapolation: 門市數目與營業收入

- 就上頁的資料顯示，統一超商門市數目和營業收入有極強的正相關的趨勢。
- 也就是說，統一超商門市數目的增加，營業收入也就隨著增加。
- **這是不是表示如果統一超商門市一直開下去，營業收入會永遠隨之增加？**

門市數目與營業收入: 2008 – 2017



Extrapolation