TAINAN UNIVERSITY OF TECHNOLOGY

線性迴歸與相關分析

Statistics, Autumn 2009 , C. J. Chang

什麼是相關分析

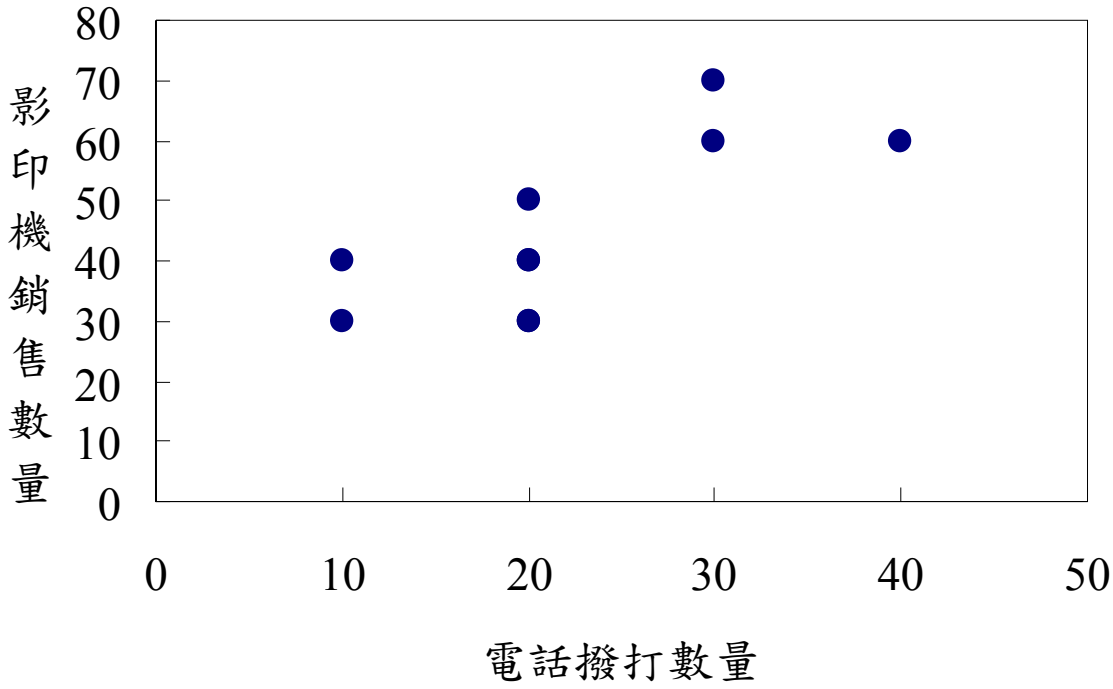
- 相關分析是一種測量兩個變數間關係強弱的方法
- Ex.

業務員	電話撥打數量	影印機銷售業績
1	20	30
2	40	60
3	20	40
4	30	60
5	10	30
6	10	40
7	20	40
8	20	50
9	20	30
10	30	70

左邊的表格似乎顯示撥越多電話的業務員，其銷售業績也較好，像這樣的關聯性，就是所謂的相關性。但僅止於此的描述，並不夠精確，因此我們必須建立統統計測量值，用更精確的方式描繪兩個變數間的關係，這樣的統計技巧稱為相關分析。

散佈圖(scatter diagram)

- 獨立變數(自變數)
  提供進行估計基礎的變數，會在*X*上取值
- 相依變數(應變數)
  要進行估計或是預測的變數，值對應在*Y*上
- 散佈圖
  散佈圖一般用於探討兩個變數之間的關係，水平軸(*X*軸)是一個變數，而垂直軸(*Y*軸)是另一個變數



散佈圖的繪製步驟

- 畫出*X*軸與*Y*軸(一般而言，變量*X*軸在橫軸，變量*Y*在縱軸)
- 依據資料標示出數軸的值域範圍
- 對應每對數值(自變數為*x*值，應變數為*y*值)，並在圖中將相對應的點標示出來

相關係數(coefficient of correlation)

- 相關係數由Pearson所創建，其用於描述兩組數值變數(區間尺度或比例尺度)間關係的強度，一般用*r*來表示。相關係數的範圍在-1與1之間
- 正負符號代表變數間關係的方向。正稱為正相關，表示兩變數的關係是正向的(自變數大(小)應變數就大(小))；負則稱為負相關，意指兩變數的關係是反向的(自變數大(小)應變數就小(大))
- 數值大小則為相關性的強弱。當數值越接近1，表示相關性越高；若數值越接近0，則相關性越低
- Ex.

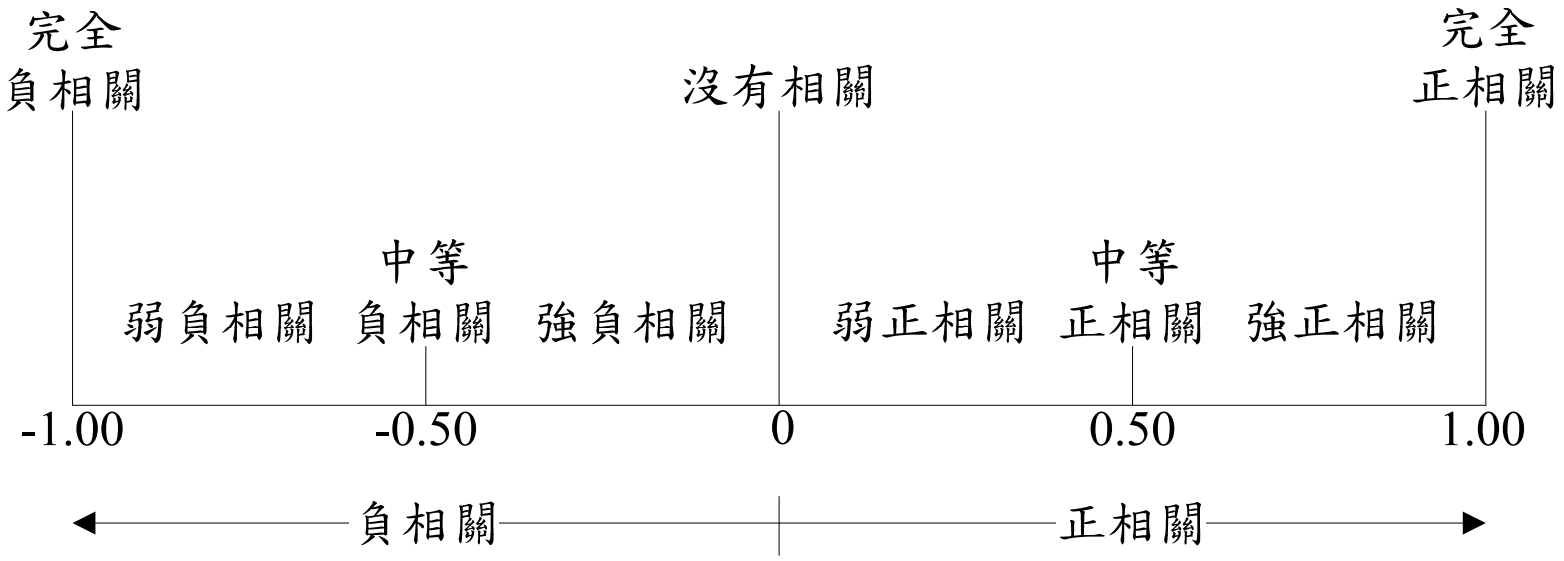
  r=+1稱為完全正相關

  r=-1稱為完全負相關

  r=0則表示兩變數沒有任何相關

相關係數的強度與正負號方向

- 相關係數的特徵
  - 母體的相關係數用 ρ (rho)表示，樣本相關係數則用r表示
  - 相關係數能呈現變數線性關係的方向與強度
  - 相關係數的值介於-1與+1之間，包含-1與+1
  - 相關係數接近0表示變數之間的相關性小
  - 相關係數接近+1表示變數之間有很強的正相關
  - 相關係數接近-1表示變數之間有很強的負相關



相關係數的計算

- 相關係數

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}$$

- 計算步驟
  - 依配對列出兩組資料
  - 計算各組資料的算術平均數
  - 計算兩組資料各別與其算術平均數的差
  - 依配對計算兩兩組資料差值的乘積並加總
  - 各別計算兩組資料的標準差
  - 將上述計算所得資料代入公式求得相關係數

相關係數(Ex.)

- 假設x與y的數據如下，試計算相關係數

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
20	30	22	45	-2	-15	4	225	30
40	60	22	45	18	15	324	225	270
20	40	22	45	-2	-5	4	25	10
30	60	22	45	8	15	64	225	120
10	30	22	45	-12	-15	144	225	180
10	40	22	45	-12	-5	144	25	60
20	40	22	45	-2	-5	4	25	10
20	50	22	45	-2	5	4	25	-10
20	30	22	45	-2	-15	4	225	30
30	70	22	45	8	25	64	625	200
總和 220	450	220	450	0	0	760	1850	900

$$\text{相關係數 } r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{900}{\sqrt{760}\sqrt{1850}}\approx 0.759$$

相關係數的延伸意義

- 相關係數僅能說明兩變數之間關係或關聯的強度與方向，但不能以此說明兩變數有因果關係(一個變數引起另一個變數的變化)
- Ex.
  某國小進行數學能力測驗赫然發現身高與分數之間呈現正相關，這樣的結果顯然與常理不合。(實際上的原因，是學力測驗未按年級分組進行，而高年級同學的身高較低年級同學高，其考試分數亦比低年級同學高；因此影響測驗分數的因素是年級，而非身高)
- Ex.
  - 教堂數量增加與犯罪人數
  - 教授薪水與精神病患人數
  - 花生消費量與阿斯匹靈消費量

假相關或間接相關

判定係數(coefficient of determination)

- 判定係數是相關係數的平方，其表示一個變數的變數可由另一個變數解釋的百分比(兩變數共享的意特徵越多，其共享的變異性也就越高，他們也就會越相關)
- Ex.

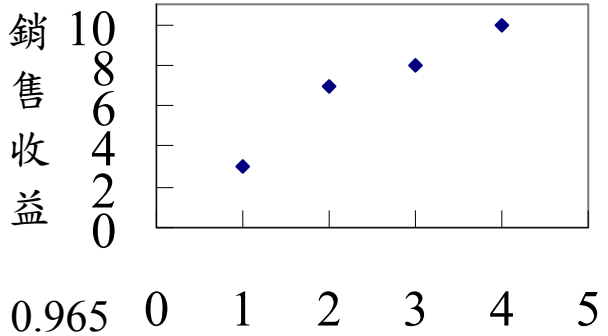
  相關係數如果是0.7，則判定係數為$0.7^2=0.49$，這說明我們有49%的變異可以被解釋。當然反過來看，其也代表51%的變異不能被解釋，雖然0.7已屬於強相關，但仍存在著我們無法解釋的原因導致變數之間存在意變化差異

相關係數(Ex.)

- 某公司為增加銷售業績，因此投入了廣告經費行銷產品，下表為最近四個月廣告支出與銷售收異的資料(以百萬美元為單位)。a.何者為自變數？何者為應變數？ b.請畫出散佈圖 c.請計算相關係數 d.解釋相關係數的強度 e.計算判定係數並解釋其意義

a. 廣告支出為自變數
銷售收異為應變數

b. 如右圖



月份	廣告支出	銷售收益
7	2	7
8	1	3
9	3	8
10	4	10

c.

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{11}{\sqrt{5}\sqrt{26}}\approx 0.965$$

d. 廣告支出與銷售收益間有強正相關

e. $r^2=0.965^2=0.931$判定係數為0.931，表銷售收益的變異中，有92%可以由廣告支出的變異解釋

	廣告支出							
x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
2	7	2.5	7	-0.5	0	0.25	0	0
1	3	2.5	7	-1.5	-4	2.25	16	6
3	8	2.5	7	0.5	1	0.25	1	0.5
4	10	2.5	7	1.5	3	2.25	9	4.5
10	28	10	28	0	0	5	26	11

相關係數的t檢定

- 當相關係數是大於或小於0，我們可以確認變數間真得有關係嗎？有沒有可能其僅是機率所造成呢(正好抽出有關聯的樣本)？為了確認這個問題，我們必須進行假設檢定
- 相關係數的t檢定統計量

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}},\ \text{自由度為}n-2$$

相關係數的樣本統計量我們用羅馬(英文)字母r表示，而母體參數則使用希臘字母ρ

- 相關係數t檢定的統計假設

統計假設	決策準則
雙尾檢定：$H_0:\rho=0, H_1:\rho\neq 0$	若$t<-t_{\alpha/2}(n-2)$ 或 $t>t_{\alpha/2}(n-2)$，拒絕H_0
左尾檢定：$H_0:\rho\geqq 0, H_1:\rho<0$	若$t<-t_{\alpha}(n-2)$，拒絕H_0
右尾檢定：$H_0:\rho\leqq 0, H_1:\rho>0$	若$t>t_{\alpha}(n-2)$，拒絕H_0

相關係數的t檢定(Ex.)

- 根據25個人口超過50000的鄉鎮，其縣長選舉的樣本資料指出，得票率與競選經費之相關係數為0.43。試問在顯著水準0.05下，檢定這兩個變數是否呈正相關

建立統計假設 $\begin{cases} H_0 : \rho \le 0 \\ H_1 : \rho > 0 \end{cases}$

使用單尾檢定 $\rightarrow$ 臨界值 $t_{0.05}(23) = 1.714$

樣本統計量 $t = \dfrac{0.43\sqrt{25-2}}{\sqrt{1-0.43^2}} \approx 2.2841 > 1.714$

$\therefore$ 拒絕虛無假設，即得票率與選舉經費為正相關

相關係數的t檢定(Ex.)

- 航空公司協會想了解班機乘客數量與飛行成本之間的關係。隨機挑選15班飛機做為樣本，其乘客數量與飛行成本之間的關係為0.667。試問在顯著水準0.01，是否可說這兩個變數呈現正相關

建立統計假設 $\begin{cases} H_0 : \rho \le 0 \\ H_1 : \rho > 0 \end{cases}$

使用單尾檢定 $\rightarrow$ 臨界值 $t_{0.01}(13) = 2.65$

樣本統計量 $t = \dfrac{0.667\sqrt{15-2}}{\sqrt{1-0.667^2}} \approx 3.2278 > 2.65$

$\therefore$ 拒絕虛無假設，即乘客數量與飛行成本呈現正相關

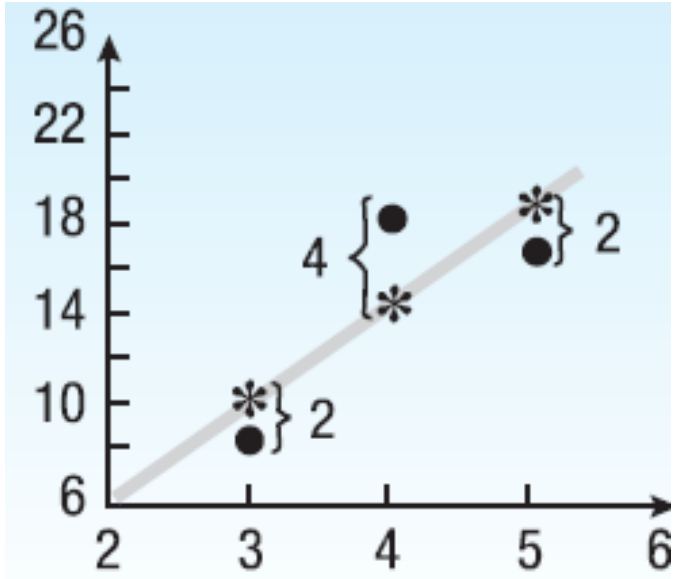
迴歸分析(regression analysis)

- 迴歸分析

  利用自變數*X*所挑選出的值，估計應變數*Y*值，這樣建構方程式的方法稱為迴歸分析，其中用來定義兩變數間線性關係的方程式稱為迴歸方程式

- 最小平方法(least squares principle)

  利用實際*Y*值與預測*Y*值之間垂直距的平方和最小化，來求取迴歸方程式的方法



最小平方法(least squares principle)

令線性迴歸方程式的一般形式為 $\hat{Y} = a + bX$

$$\begin{aligned}\sum (Y-\hat{Y})^2 &= \sum (Y-(a+bX))^2 \\ &= \sum Y^2 + \sum (a+bX)^2 - 2\sum Y(a+bX) \\ &= \sum Y^2 + \sum a^2 + \sum b^2X^2 + 2\sum abX - 2\sum aY - 2\sum bXY \\ &= \sum Y^2 + na^2 + b^2\sum X^2 + 2ab\sum X - 2a\sum Y - 2b\sum XY\end{aligned}$$

對a微分並令其為0可得 $\rightarrow 2na + 2b\sum X - 2\sum Y = 0$

$$\rightarrow na + nb\bar{X} - n\bar{Y} = 0$$

對b微分並令其為0可得 $\rightarrow 2b\sum X^2 + 2a\sum X - 2\sum XY = 0$

$$\rightarrow b\sum X^2 + na\bar{X} - \sum XY = 0$$

聯立求解可得 $\begin{cases} a = \bar{Y} - b\bar{X} \\ b = \dfrac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} \end{cases} \rightarrow \begin{cases} a = \bar{Y} - b\bar{X} \\ b = \dfrac{\sum (X-\bar{X})(Y-\bar{Y})}{\sum (X-\bar{X})^2} \end{cases} \rightarrow \begin{cases} a = \bar{Y} - b\bar{X} \\ b = r\dfrac{s_y}{s_x} \end{cases}$

線性迴歸的計算公式

- 線性迴歸方程式

$$\hat{Y} = a + bX$$

- 迴歸線的斜率

$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = r\frac{s_y}{s_x}$$

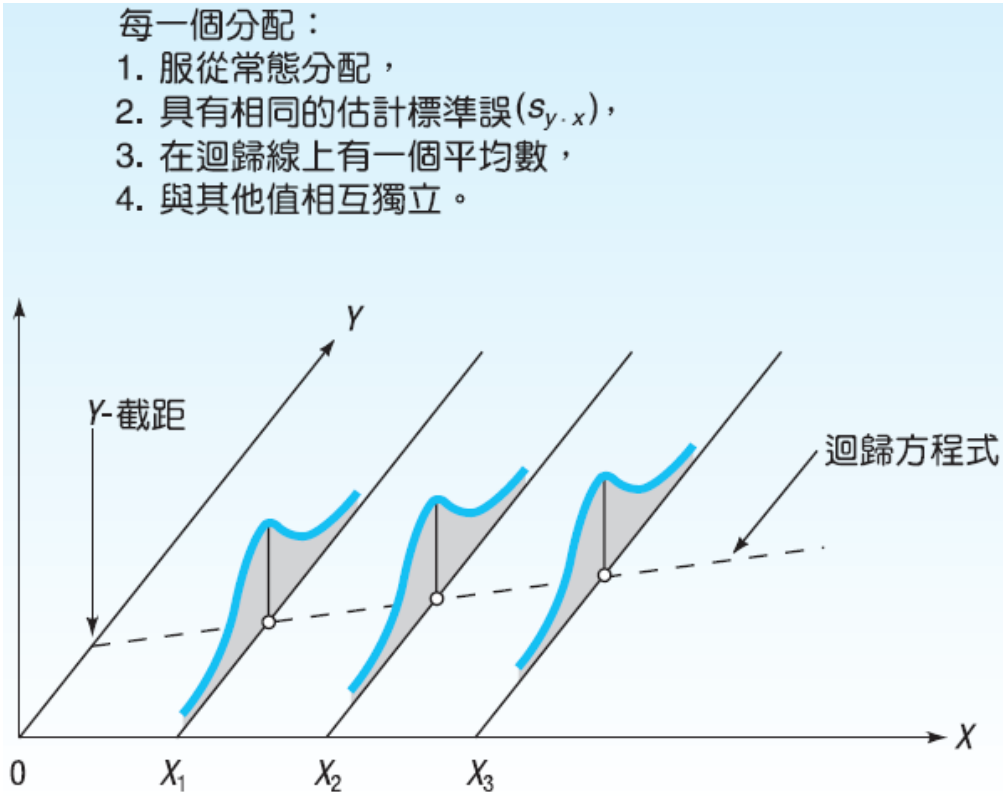
- Y-截距

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

r是相關係數
s_x是應變數X的標準差
s_y是應變數Y的標準差
$\bar{X}$為自變數的平均數
$\bar{Y}$為應變數的平均數

線性迴歸的假設

- 應用線性迴歸時，必須先滿足下列的假設條件：
  - 每個*X*值都會有其對應的一群*Y*值，且這些*Y*值服從常態分配
  - 這些常態分配的平均數皆落在迴歸線上
  - 這些常態分配的標準差都相同
  - *Y*值間相互獨立，亦即選取一特定*X*值作為樣本時，與其他X的值並無相關



迴歸分析(Ex.)

- 某公司為增加銷售業績，因此投入了廣告經費行銷產品，下表為最近四個月廣告支出與銷售收異的資料(以百萬美元為單位)。a.計算迴歸方程式 b.請估計3百萬美元廣告支出下的銷售收益

月份	廣告支出	銷售收益
7	2	7
8	1	3
9	3	8
10	4	10

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
2	7	2.5	7	-0.5	0	0.25	0	0
1	3	2.5	7	-1.5	-4	2.25	16	6
3	8	2.5	7	0.5	1	0.25	1	0.5
4	10	2.5	7	1.5	3	2.25	9	4.5
10	28	10	28	0	0	5	26	11

a.

$$b=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}=\frac{11}{5}\approx 2.2$$

$$a=\bar{Y}-b\bar{X}=7-2.2\times 2.5=1.5$$

$$\hat{Y}=a+bX=1.5+2.2X$$

b.

$$X=3\rightarrow\hat{Y}(3)=1.5+2.2\times 3=8.1$$

$\therefore$ 3百萬美元廣告支出下，銷售收益為8.1百萬美元

迴歸分析(Ex.)

- 隨機選取下列樣本資料。a.計算迴歸方程式 b.請計算 X=7時，$\hat{Y}$ 的值

X:	4	5	3	6	10
Y:	4	6	5	7	7

x	y	$\bar{x}$	$\bar{y}$	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
4	4	5.6	5.8	-1.6	-1.8	2.56	3.24	2.88
5	6	5.6	5.8	-0.6	0.2	0.36	0.04	-0.12
3	5	5.6	5.8	-2.6	-0.8	6.76	0.64	2.08
6	7	5.6	5.8	0.4	1.2	0.16	1.44	0.48
10	7	5.6	5.8	4.4	1.2	19.36	1.44	5.28
28	29	28	29	0	0	29.2	6.8	10.6

a.

$$b=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}=\frac{10.6}{29.2}\approx 0.363$$

$$a=\bar{Y}-b\bar{X}=5.8-0.363\times 5.6=3.7672$$

$$\hat{Y}=a+bX=3.7672+0.363X$$

b.

$$X=7\rightarrow \hat{Y}(7)=3.7672+0.363\times 7=6.3082$$

$$\therefore \hat{Y}=6.3082$$

TAINAN UNIVERSITY OF TECHNOLOGY

The end of this chapter.

Thank You !