統計學

第十一章 相關與迴歸分析

書號：512282

編著 江建良

普林斯頓國際有限公司

11-1 相關分析

係指一個統計量數來表示變數間的相關情形，並進一步來進行推論的統計方法。

其內容包括統計量數(即相關係數)的計算與統計推論兩部分。

相關的意義與種類：

- **相關的意義：**係指兩個變數或兩個以上變數間之關係。
- **相關的種類：**

(1) 依變數的多寡分：① 簡相關 ② 複相關

(2) 依相關的形態分：① 直線相關 ② 曲線相關

(3) 依相關的程度分：① 完全相關 ② 不完全相關

(4) 依相關的方向分：① 正相關 ② 負相關

11-1 相關分析

簡單直線相關：

散佈圖：

了解兩變數之間是否有關係，最簡單的方法是繪製散佈圖。

相關係數：

係指用以測量兩變數間直線相關的程度與方向的統計量數，一般以 r 表示之。

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2}\sqrt{\Sigma(Y_i - \overline{Y})^2}}$$

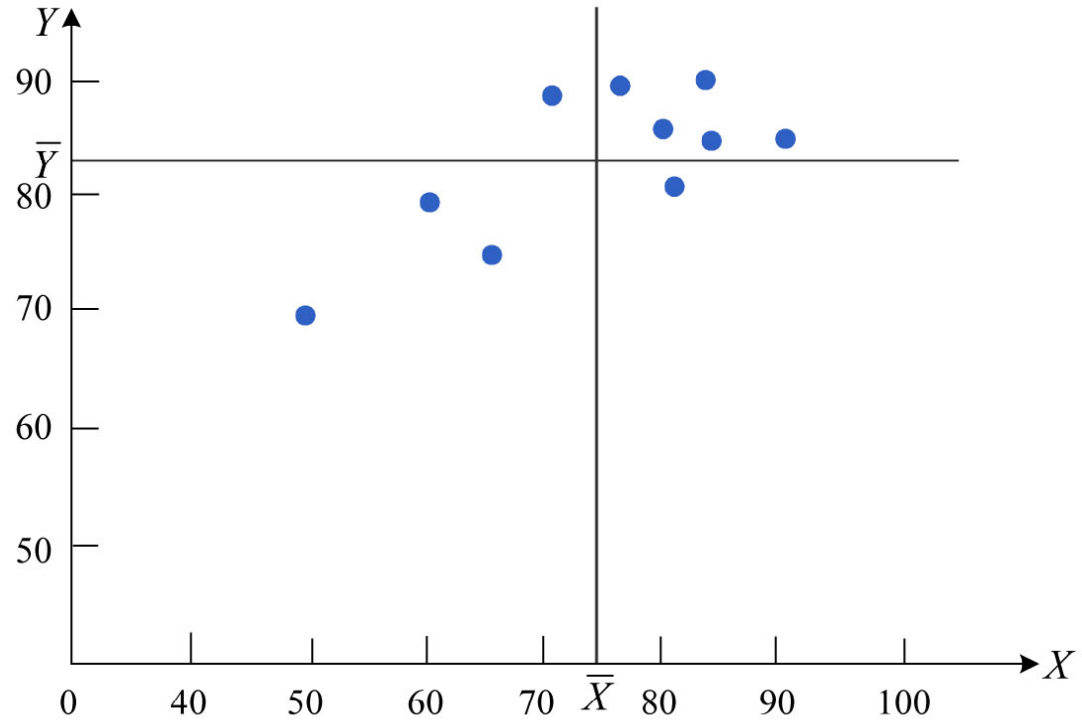
$$= \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{\sqrt{n\Sigma X_i^2 - (\Sigma X_i)^2}\sqrt{n\Sigma Y_i^2 - (\Sigma Y_i)^2}}$$

例題 11-1

下表是十名學生期末考數學與統計學的成績資料：

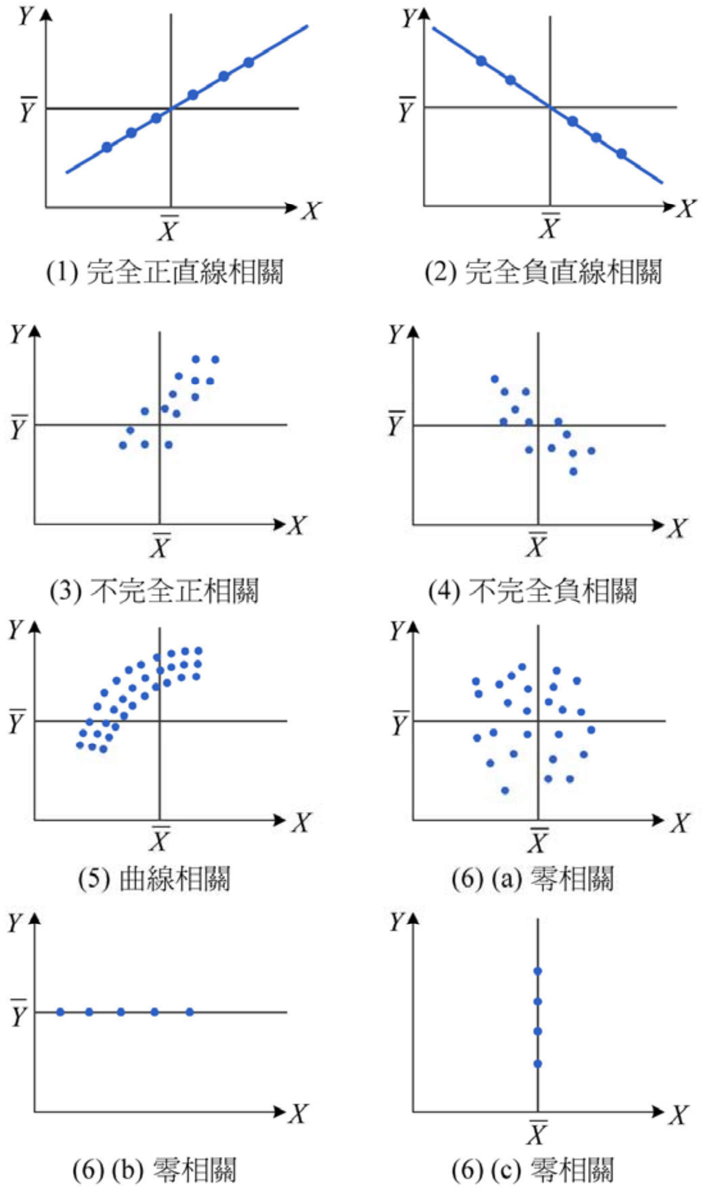
學生號碼	1	2	3	4	5	6	7	8	9	10
數　學 (X)	50	90	75	60	85	80	85	65	80	70
統計學 (Y)	70	85	90	80	85	80	90	75	87	88

解：



（ 註：為便於觀察分析，通常會將兩變數的平均值，加劃兩條平行於縱橫軸之直線，而此兩直線之相交點乃為散佈圖之中心點，見相關係數說明。）

我們可依相關之種類 (11-1.1)，繪製常見的散佈圖型態，如下所示：


(1) 完全正直線相關
(2) 完全負直線相關
(3) 不完全正相關
(4) 不完全負相關
(5) 曲線相關
(6) (a) 零相關
(6) (b) 零相關
(6) (c) 零相關

11-1 相關分析

- **簡單直線相關：**
  - **相關係數：**

    ① 相關係數 r 之變化範圍為 $-1 \leq r \leq 1$。

    ② r 為正數時表示正相關，r 為負數時表示負相關。

    $r = 0$ 表示零相關。

    $r = +1$ 表示完全正相關。

    $r = -1$ 表示完全負相關。

  - **母體相關係數：**

$$\rho = \frac{1}{N}\sum\left(\frac{X_i - \mu_x}{\sigma_x}\right)\left(\frac{Y_i - \mu_y}{\sigma_y}\right)$$

例題 11-2

試以例題 11-1 之例，求十名學生期末考數學與統計學成績之相關係數 r，並說明其性質。

解：(a) $\sum XY = (50)(70) + (90)(85) + \cdots\cdots + (70)(88) = 61{,}970$

(b) $\sum X = 740$, $\sum Y = 830$, $\sum X \sum Y = 614{,}200$

(c) $\sum X^2 = 56{,}200$, $\sum Y^2 = 69{,}288$

(d) $$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{(10)(61{,}970) - 614{,}200}{\sqrt{(10)(56{,}200) - (740)^2}\sqrt{(10)(69{,}288) - (830)^2}}$$

$$= \frac{5{,}500}{\sqrt{14{,}400}\sqrt{3{,}980}} = 0.72$$

$\because$ $0.72 > 0.7$ 又為正數，$\therefore$ 屬高度正相關。

例題 11-3

隨機抽樣共 9 組樣本，其資料如下，試求相關係數 r，並說明兩變數的關係。

X	0	5	10	0	5	10	0	5	10
Y	0	0	0	5	5	5	10	10	10

解：(a) $\sum XY = 225, \sum X = 45, \sum Y = 45$

(b) $\sum X^2 = 375, \sum Y^2 = 375,$

$n\sum XY - \sum X \sum Y = 9(225) - (45)(45) = 0$

(c) $$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{0}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} = 0$$

(d) $\because r = 0$ ，$\therefore$ 兩變數爲零相關。

11-1 相關分析

- **簡單直線相關：**

  - **母體相關係數ρ 的檢定：**

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

① 若 t_0 落在放棄域，則差異顯著，放棄 H_0，接受 H_1。

② 若 t_0 不落在放棄域，則差異不顯著，接受 H_0。

例題 11-4

試利用例題 11-1 之資料，檢定母體相關係數 ρ 是否為零？設 $\alpha = 0.05$。

解： (a) 假設：$H_0 : \rho = 0 ; H_1 : \rho \neq 0$（雙尾檢定）

(b) 顯著水準：$\alpha = 0.05$

(c) 放棄域：$t < -t_{(0.975,8)} = -2.306$ 及 $t > t_{(0.975,8)} = 2.306$

(d) 計算：本題樣本相關係數 $r = 0.72 , n = 10$，代入

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.72\sqrt{10-2}}{\sqrt{1-(0.72)^2}} = \frac{2.03}{0.7} = 2.90$$

(e) 判斷：$\because 2.90 > 2.306$，落在放棄域，即差異顯著，放棄 H_0，表示 X 與 Y 具有顯著相關（正相關）。

例題 11-5

X	4	5	9	14	18	22	24
Y	16	22	11	16	7	3	17

(a) 試就上列資料，求算相關係數。

(b) 設 $\alpha = 0.01$，檢定母體相關係數是否為零？

解：(a) 計算相關係數 r：

① $\sum XY = 1{,}097$, $\sum X = 96$, $\sum Y = 92$, $\sum X^2 = 1{,}702$, $\sum Y^2 = 1{,}464$

$$r = \frac{7(1{,}097) - (96)(92)}{\sqrt{7(1{,}702) - (96)^2}\sqrt{7(1{,}464) - (92)^2}}$$

$$= \frac{-1{,}153}{2{,}193.94} = -0.52 \text{（負相關）}$$

(b) ① 假設： $H_0:\rho=0\,;H_1:\rho\neq 0$

② 顯著水準： $\alpha=0.01$

③ 放棄域： $t<-t_{(0.995\,,5)}=-4.032$ 及 $t>t_{(0.995\,,5)}=4.032$

④ 計算： $n=7\,,r=-0.52$ ，代入

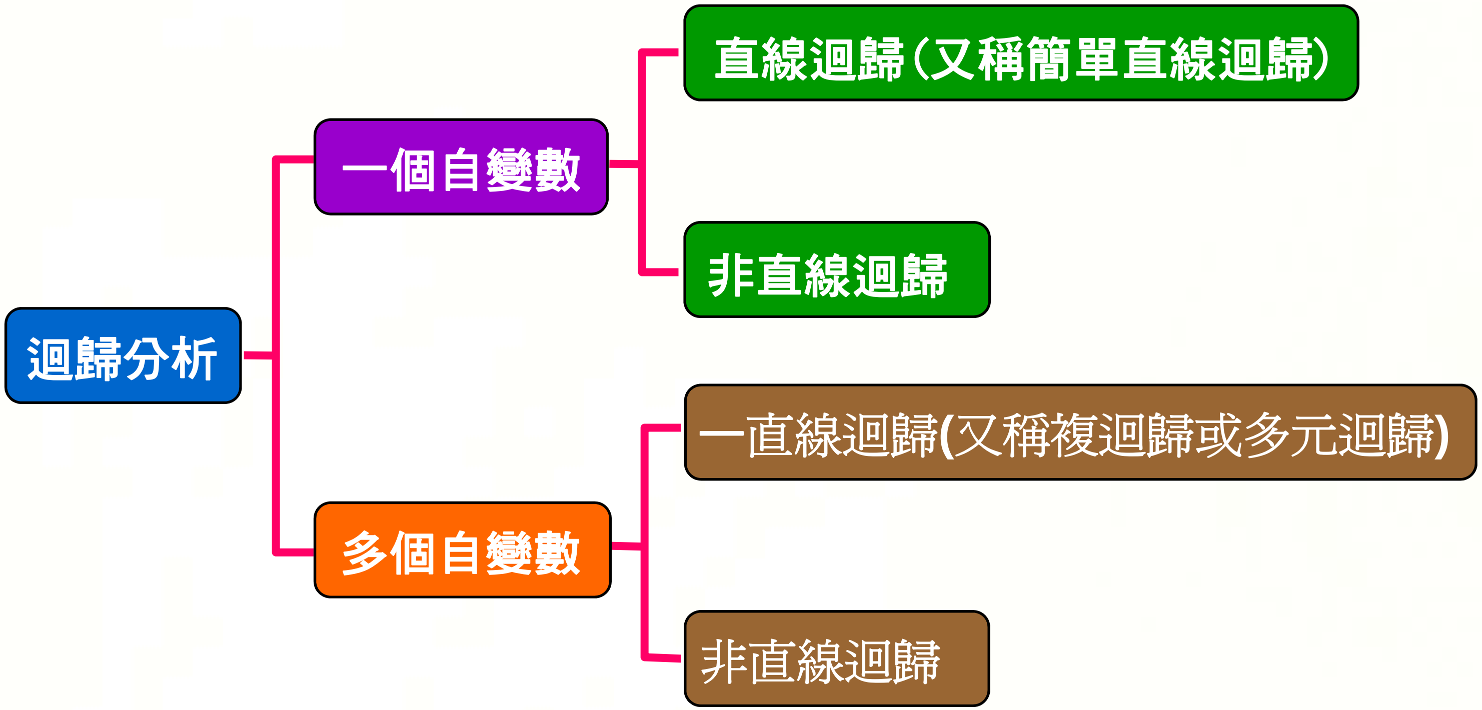
$$t_0=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}=\frac{(-0.52)(\sqrt{7-2})}{\sqrt{1-(-0.52)^2}}=-1.36$$

⑤ 判斷：$\because -1.36>-4.032$ ，落在接受域，差異不顯著，即接受 H_0 ，

表示 X 與 Y 可能無關。

11-2 迴歸分析

迴歸分析，係指利用一個或多個自變數 (independent variable，或稱預測變數) 來預測或推估依變數 (dependent variable，或稱準則變數) 分析方法。



11-3 簡單直線迴歸

- **迴歸的原始意義：**

"迴歸" 一詞，最早出現於英國科學 Galton 爵士 (1822～1911)
之遺傳學研究上，Galton 發現兒子的身高通常趨近於父母身高的平均數，而非趨近於某一極端值。

- **樣本迴歸直線的求算：**

令樣本迴歸線為 $\hat{Y} = a + bX$ (a 與 b 為迴歸係數)：

$$\begin{cases} b = \dfrac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \dfrac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} \\ a = \overline{Y} - b\overline{X} \end{cases}$$

例題 11-6

今隨機抽取 12 位學生，調查期末考數學與統計學之成績分數，如下表所示：

學　生	1	2	3	4	5	6	7	8	9	10	11	12
數　學 (X)	55	70	70	70	65	50	50	55	65	55	55	65
統計學 (Y)	74	91	98	87	90	74	76	81	94	85	76	85

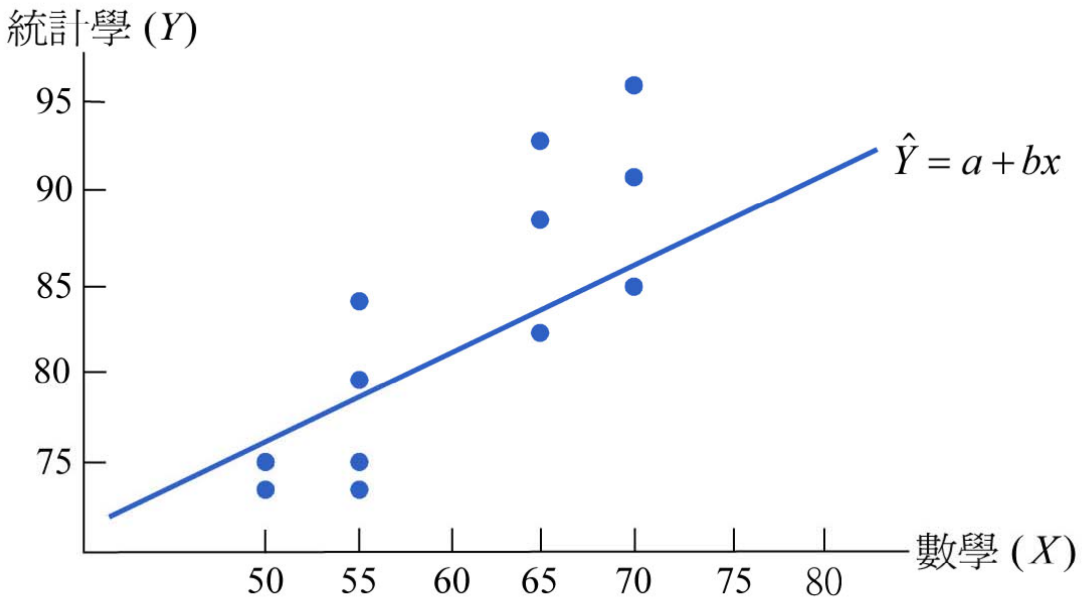
試求：

(a) 繪製散佈圖觀察 X 與 Y 兩變數的關係。

(b) 建立 Y 對 X 的迴歸直線（ 迴歸方程 ）。

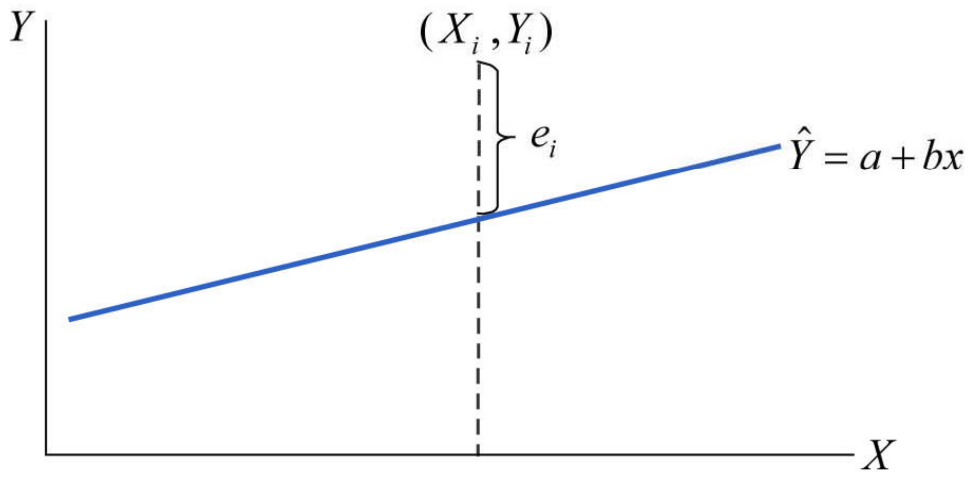
(c) 若隨機抽取一位學生其數學成績為 60 分，試預測其統計學分數為何？

解：(a) 首先繪製散佈圖，來觀察數學 (X) 與統計學 (Y) 兩者的關係，如下圖所示：



由散佈圖可知，$X = 50$ 對應 Y 之數為 74 與 76 二數；$X = 55$ 對應 Y 之數為 74, 81, 85, 76 四數；$X = 65$ 對應 Y 之數為 90, 94, 85 三數；$X = 70$ 對應 Y 之數為 91, 98, 87 三數。X 與 Y 的關係並非函數關係，而是統計關係。因此必須利用迴歸的方式來建立迴歸直線。

(b) 由散佈圖觀察得知，兩變數間有類似直線的關係。茲先導出直線的一般模式（即直線迴歸線）。設從母體 (X, Y) 資料中抽樣本大小爲 n 的隨機樣本 $(X_i, Y_i), i=1,2,\cdots,n$。令樣本迴歸線爲 $\hat{Y}=a+bX$（a 與 b 爲迴歸係數），我們利用**最小平方法 (least square method)** 的觀念（即各樣本觀察值到此直線的垂直離差平方和爲最小），導出最佳迴歸線，如下圖所示：



茲假設上圖 e_i 表示第 i 個觀察值到迴歸線的垂直離差，而最小平方法能使特定的 a 與 b 值，滿足樣本觀察值的離差平方和為最小。離差平方和一般稱為迴歸線的**誤差平方和 (sum of squares of the errors)**，以 SSE 表示，即求出 a、b 值，使下式為最小值

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \Sigma(Y-\hat{Y})^2 = \Sigma(Y-a-bX)^2$$

上式我們可藉由微分的方法，求出 a 與 b 值，而使 SSE 為最小，即得到最佳的迴歸線。經計算得下列兩個**常態方程式 (normal equation)**。

$$\begin{cases} \sum Y = na + b\sum X \\ \sum XY = a\sum X + b\sum X^2 \end{cases}$$

解聯立方程式，得

$$\begin{cases} b = \dfrac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \dfrac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(X-\overline{X})^2} \\ a = \overline{Y} - b\overline{X} \end{cases}$$

現在我們可利用 a 與 b 值的公式，求出迴歸線

$$\because\ \sum X = 725,\ \sum Y = 1{,}011,\ \sum XY = 61{,}685,\ \sum X^2 = 44{,}475$$

$$\overline{X} = 60.42,\ \overline{Y} = 84.25$$

$$\begin{aligned} b &= \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \\ &= \frac{(12)(61{,}685) - (725)(1{,}011)}{(12)(44{,}475) - (725)^2} \\ &= 0.897 \end{aligned}$$

$$a = \overline{Y} - b\overline{X} = 84.25 - (0.897)(60.42) = 30.06$$

$\therefore$ 迴歸線為 $\hat{Y} = a + bX = 30.06 + 0.897X$

式中 $a(=30.06)$ 代表截距，$b(=0.897)$ 代表斜率，表示 X 增加一單位，Y 平均增加 0.897 單位。

(c) $\because\ \hat{Y} = 30.06 + 0.897X$，將 $X = 60$ 代入此式，得

$$\begin{aligned} \hat{Y} &= 30.06 + (0.897)(60) \\ &= 83.88\,(\text{分}) \fallingdotseq 84\,(\text{分}) \end{aligned}$$

$\therefore$ 我們可預測某生數學 60 分，則統計學可能 84 分。

例題 11-7

茲就下列資料，試求：

(a) Y 對 X 迴歸直線的方程式 $\hat{Y}=a+bX$。

(b) 繪製散佈圖。

(c) $X=7$ 時，Y 為多少？

(d) a、b 的意義。

(e) 證明 $\sum e=\sum(Y-\hat{Y})=0$。

X	1	2	3	4	5	6
Y	6	4	3	5	4	2

解：(a) $\because \sum X = 21$, $\sum Y = 24$, $\sum XY = 75$, $\sum X^2 = 91$, $\sum Y^2 = 106$,

$\overline{X} = 3.5$, $\overline{Y} = 4$

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

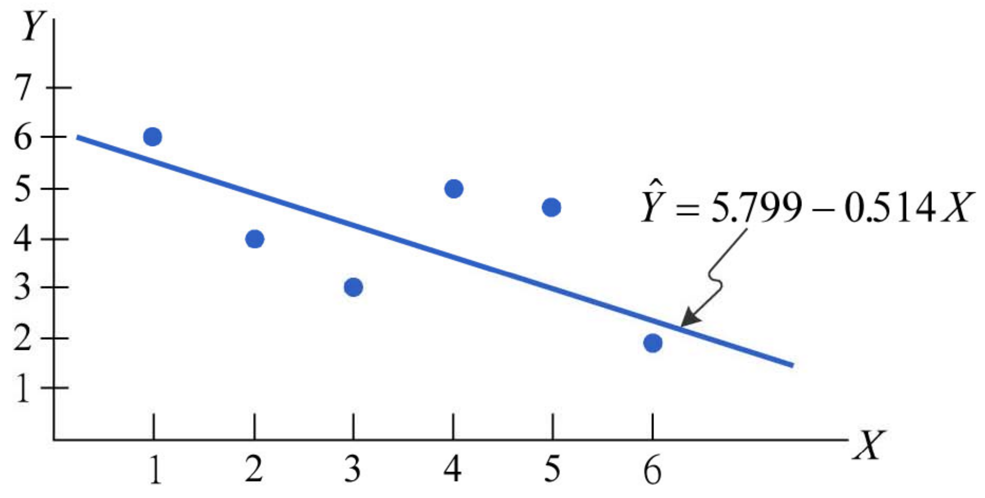
$$= \frac{(6)(75) - (21)(24)}{(6)(91) - (21)^2}$$

$$= -0.514$$

$$a = \overline{Y} - b\overline{X} = 4 - (-0.514)(3.5) = 5.799$$

$\therefore$ 迴歸線爲 $\hat{Y} = 5.799 - 0.514X$

(b)



(c) $X = 7 \Rightarrow \hat{Y} = 5.799 - 0.514(7) = 2.20$

(d) 截距 $a = 5.799$，表示當 $X = 0$ 時，Y 為 5.799；斜率 $b = -0.514$，表示 X 每增加一單位，Y 平均減少 0.514 單位。

(e)

X	Y	$\hat{Y} = 5.799 - 0.514X$	$e = Y - \hat{Y}$
1	6	5.285	0.715
2	4	4.771	−0.771
3	3	4.257	−1.257
4	5	3.743	1.257
5	4	3.229	0.771
6	2	2.715	−0.715
			0

$$\therefore \quad \sum e = \sum (Y - \hat{Y}) = 0$$

例題 11-8

愛之味公司為決定每週廣告支出與銷售額之關係，乃蒐集近年來廣告支出與銷售額之資料如下：(單位：萬元)

廣告支出 (X)	40	50	25	50	50	20	50	40	20	20	25	30
銷 售 額 (Y)	525	420	480	510	560	490	440	385	400	365	395	475

試求：(a) 按廣告支出預測銷售額的迴歸方程式。

(b) 若廣告支出為 60 萬元，則每週銷售額預測為多少？

解：(a) $\because \quad \sum X = 420, \sum Y = 5{,}445, \sum XY = 194{,}125, \sum X^2 = 16{,}550,$

$\sum Y^2 = 2{,}512{,}925, \bar{X} = 35, \bar{Y} = 453.75$

$$b = \frac{(12)(194{,}125) - (420)(5{,}445)}{(12)(16{,}550) - (420)^2} = 1.92$$

$$a = 453.75 - (1.92)(35) = 386.55$$

$$\therefore \quad \hat{Y} = 386.55 + 1.92X$$

(b) $X = 60, \hat{Y} = 386.55 + (1.92)(60) = 501.75$ (萬元)

例題 11-9

試就下列資料求算：

(a) Y 對 X 的迴歸直線。

(b) X 對 Y 的迴歸直線。

(c) 相關係數 r。

(d) 迴歸係數與相關係數的關係。

X	1	2	3	4	5	6
Y	6	4	3	5	4	2

解： $\sum X=21, \sum Y=24, \sum XY=75, \sum X^2=91, \sum Y^2=106, \overline{X}=3.5, \overline{Y}=4$

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

$$b=\frac{(6)(75)-(21)(24)}{(6)(91)-(21)^2}=-0.514$$

$$a=\overline{Y}-b\overline{X}=4-(-0.514)(3.5)=5.799$$

$$\hat{Y}=5.799-0.514X$$

(b) $\hat{X}=a'+b'Y$

$$b'=\frac{(6)(75)-(21)(24)}{(6)(106)-(24)^2}=-0.9$$

$$a'=35-(-0.9)(4)=7.1$$

$$\hat{X}=7.1-0.9Y$$

(c) $r=\dfrac{(6)(75)-(21)(24)}{\sqrt{(6)(91)-(21)^2}\sqrt{6(106)-(24)^2}}=-0.68$

(d) $\because b$ 與 b' 同爲負號，$\therefore r$ 爲負號

$$r=-\sqrt{b\times b'}=-\sqrt{(-0.514)(-0.9)}=-0.68$$

11-3 簡單迴歸分析

估計標準誤：

係指估計值 $\hat{Y}$ 與樣本觀察值 Y 之間的誤差平方和 (SSE) 除以自由度 n-2 的平方根，以 $\sqrt{\text{MSE}}$ 表示。

$$\hat{\sigma} = \hat{S}_{YX} = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

$$= \sqrt{\frac{\sum (Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

11-4 母體迴歸模式

- **母體與樣本迴歸直線的差異：**

	母體迴歸直線	樣本迴歸直線
模 式	$E(Y\|X_i) = \alpha + \beta X$ (α與β為未知迴歸母數)	$\hat{Y} = a + bX$ (α與β為迴歸係數)
線 型	迴歸直線	迴歸直線(迴歸方程)
型 態	機率分配 (Y_1 為隨機變數)	與機率分配無關 (Y_1 為觀察值)
誤差項	$\varepsilon_i = Y_i - E(Y\|X_i)$ (隨機誤差)	$e_i = Y_i - \hat{Y}_i$ (一般誤差)

11-5 迴歸母數的推論

- **母體迴歸直線常數項α的推論：**
  - **α的區間估計：**

信賴區間： $\alpha \sim a \mp t_{(1-\frac{\alpha}{2})} \times \hat{\sigma} \sqrt{\dfrac{\sum X^2}{n\sum(X-\overline{X})^2}}$

  - **α的檢定：**

$$t_0 = \frac{a-\alpha_0}{\hat{\sigma}\sqrt{\dfrac{\sum X^2}{n\sum(X-\overline{X})^2}}}$$

① 若 t_0 落在放棄域，則差異顯著，放棄 H_0，接受 H_1。

② 若 t_0 不落在放棄域，則差異不顯著，接受 H_0。

例題 11-10

已知下列資料適合迴歸分析，試作：

(a) Y 對 X 的迴歸方程式 $\hat{Y}=a+bX$。

(b) 對母體迴歸線 $E(Y|X)=\alpha+\beta X$ 之 α 作區間估計，設 $1-\alpha=0.95$，並檢定 α 是否為 0？

X	6	4	10	2	8
Y	7	1	9	3	5

解：(a) $\sum X=30,\sum Y=25,\sum X^2=220,\sum Y^2=165,\sum XY=182,n=5,\overline{X}=6,\overline{Y}=5$

$$b=\frac{(5)(182)-(30)(25)}{5(220)-(30)^2}=0.8$$

$$a=5-(0.8)(6)=0.2$$

$$\therefore\quad \hat{Y}=0.2+0.8X$$

(b) $\hat{\sigma} = \sqrt{\dfrac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$

$= \sqrt{\dfrac{165-(0.2)(25)-(0.8)(182)}{5-2}} = \sqrt{4.8} = 2.19$

$t_{\left(1-\frac{\alpha}{2},\, n-2\right)} = t_{(0.975,\,3)} = 3.182$

$\sum(X-\overline{X})^2 = \sum X^2 - n\overline{X}^2 = 220-(5)(6)^2 = 40$

$\alpha \sim a \mp t_{\left(1-\frac{\alpha}{2},\, n-2\right)} \cdot \hat{\sigma}\sqrt{\dfrac{\sum X^2}{n\sum(X-\overline{X})^2}}$

$= 0.2 \mp (3.182)(2.19)\left(\sqrt{\dfrac{220}{(5)(40)}}\right)$

$= 0.2 \mp 7.31$

$\therefore \quad -7.11 \le \alpha \le 7.51$

(c) ① 假設：$H_0 : \alpha = 0$；$H_1 : \alpha \neq 0$

② 顯著水準：$\alpha = 0.05$

③ 放棄域：$t < -t_{(0.975,3)} = -3.182$ 及 $t > t_{(0.975,3)} = 3.182$

④ 計算：$$t_0 = \frac{a - \alpha_0}{\hat{\sigma}\sqrt{\dfrac{\sum X^2}{n\sum(X - \overline{X})^2}}} = \frac{0.2 - 0}{2.19 \cdot \sqrt{\dfrac{220}{5(40)}}} = 0.087$$

⑤ 判斷：$\because 0.087 > -3.182$，落在接受域，差異不顯著，$\therefore$ 接受 H_0，即母體迴歸直線之截距可能為 0，即沒有截距。由 (b) 亦可看出，因為估計區間包括 0，差異不顯著。

11-5 迴歸母數的推論

母體迴歸直線常數項β的推論：

β的區間估計：

信賴區間： $\beta \sim a \mp t_{(1-\frac{\alpha}{2})} \times \hat{\sigma} \sqrt{\frac{1}{\Sigma(X-\overline{X})^2}}$

β的檢定：

$$t_0 = \frac{b-\beta_0}{\hat{\sigma}\sqrt{\frac{1}{\Sigma(X-\overline{X})^2}}}$$

① 若 t_0 落在放棄域，則差異顯著，放棄 H_0，接受 H_1。

② 若 t_0 不落在放棄域，則差異不顯著，接受 H_0，即β 可能為 0，自變數無法預測依變數。

例題 11-11

試就上例，作

(a) $E(Y|X)=\alpha+\beta X$ 之 β 的區間估計，設 $1-\alpha=0.95$。

(b) 檢定 β 是否為 0？其涵義為何？

解：(a) $\beta \sim b \mp t_{\left(1-\frac{\alpha}{2}\right)} \cdot \hat{\sigma}\sqrt{\frac{1}{\Sigma(X-\overline{X})^2}}$

$$= 0.8 \mp (3.182)(2.19)\sqrt{\frac{1}{40}}$$

$$= 0.8 \mp 1.10$$

$$\therefore \quad -0.30 \le \beta \le 1.90$$

(b) ① 假設：$H_0 : \beta = 0 ; H_1 : \beta \neq 0$

② 顯著水準：$\alpha = 0.05$

③ 放棄域：$t < -t_{(0.975,3)} = -3.182$ 及 $t > t_{(0.975,3)} = 3.182$

④ 計算：$$t_0 = \frac{b - \beta_0}{\hat{\sigma}\sqrt{\dfrac{1}{\Sigma(X - \overline{X})^2}}} = \frac{0.8 - 0}{2.19\sqrt{\dfrac{1}{40}}} = 2.309$$

⑤ 判斷：$\because 2.309 < 3.182$，落在接受域，差異不顯著，$\therefore$ 接受 H_0，表示 β 可能為 0，即母體迴歸直線可能與橫軸平行，換言之，自變數無法預測依變數，X 與 Y 可能無關。

11-6 迴歸預測

(1) 在定 X_0 值下，對依變數Y的平均值 $E(Y|X_0)$ 之預測。

(2) 在定 X_0 值下，對依變數 Y 的單一值 Y_0 之預測。

- **特定 X_0 值之下 $E(Y|X_0)$ 之預測：**

區間估計：$E(Y|X_0) \sim \hat{Y}_0 \mp t_{(1-\frac{\alpha}{2})} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\Sigma(X - \overline{X})^2}}$

- **特定 X_0 值之下，對依變數 Y 的單一值 Y_0 之預測：**

區間估計：$Y_0 \sim \hat{Y}_0 \mp t_{(1-\frac{\alpha}{2})} \times \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\Sigma(X - \overline{X})^2}}$

例題 11-12

承 11-5 之 [例題 11-10]，試求：

(a) 若 $X=5$ 時，其所對應之依變數平均值 $E(Y|5)$ 之 95% 的信賴區間。

(b) 若 $X=5$ 時，其所對應之依變數 Y 之 95% 的信賴區間。

解：(a) $X_0 = 5,\ \hat{Y}_0 = 0.2 + 0.8(5) = 4.2$

$$E(Y|X=5) \sim \hat{Y}_0 \mp t_{\left(1-\frac{\alpha}{2}\right)}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\Sigma(X-\overline{X})^2}}$$

$$= 4.2 \mp (3.182)(2.19)\sqrt{\frac{1}{5}+\frac{(5-6)^2}{40}} = 4.2 \mp 3.31$$

$$\therefore\quad 0.89 \le E(Y|X=5) \le 7.51$$

(b) $$Y_0 \sim \hat{Y}_0 \mp t_{\left(1-\frac{\alpha}{2}\right)}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\Sigma(X-\overline{X})^2}}$$

$$= 4.2 \mp (3.182)(2.19)\sqrt{1+\frac{1}{5}+\frac{(5-6)^2}{40}} = 4.2 \mp 7.71$$

$$\therefore\quad -3.51 \le Y_0 \le 11.91$$

例題 11-13

設甲班 9 個學生之統計學期中考 (X)，期末考 (Y) 成績如下：

期中考 (X)	77	70	81	96	67	95	94	72	50
期末考 (Y)	80	78	47	96	68	99	82	45	66

試求：

(a) Y 對 X 迴歸直線 $\hat{Y}=a+bX$。

(b) 期中考為 75 分而期末考成績平均數 $E(Y|75)$ 之 95% 的信賴區間。

(c) 若某生期中考為 75 分，但期末考缺考，試估計該生期末考統計學成績為多少？期末考成績之 95% 的信賴區間。

解：(a) $\sum X = 702$, $\sum Y = 661$, $\sum XY = 52{,}852$, $\sum X^2 = 56{,}640$, $\sum Y^2 = 51{,}439$,

$n = 9$, $\overline{X} = 78$, $\overline{Y} = 73.44$

$$b = \frac{(9)(52{,}852) - (702)(661)}{(9)(56{,}640) - (702)^2} = 0.69$$

$$a = \overline{Y} - b\overline{X} = 73.44 - (0.69)(78) = 19.62$$

$$\therefore \quad \hat{Y} = 19.62 + 0.69X$$

(b) $\hat{Y}_0 = 19.62 + (0.69)(75) = 71.37$

$t_{\left(1-\frac{\alpha}{2}, n-2\right)} = t_{(0.975, 7)} = 2.365$

$$\hat{\sigma} = \sqrt{\frac{51{,}439 - (19.62)(661) - (0.69)(52{,}852)}{9-2}} = 16.91$$

$$\sum (X - \overline{X})^2 = \sum X^2 - n\overline{X}^2 = 56{,}640 - (9)(78)^2 = 1{,}884$$

$$E(Y|X = 75) \sim \hat{Y}_0 \mp t_{\left(1-\frac{\alpha}{2}, n-2\right)} \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X - \overline{X})^2}}$$

$$= 71.37 \mp (2.365)(16.91)\left(\sqrt{\frac{1}{9} + \frac{(75-78)^2}{1{,}884}}\right)$$

$$= 71.37 \mp 13.62$$

$$\therefore \quad 57.75 \le E(Y|75) \le 84.99$$

(c) ① $\hat{Y}_0 = 19.62 + (0.69)(75) = 71.37$

② $Y_0 \sim \hat{Y}_0 \mp t_{\left(1-\frac{\alpha}{2},\, n-2\right)} \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X - \overline{X})^2}}$

$= 71.37 \mp (2.365)(16.91)\left(\sqrt{1 + \frac{1}{9} + \frac{(75-78)^2}{1{,}884}}\right)$

$= 71.37 \mp 40.11$

$\therefore \quad 31.26 \le Y_0 \le 111.48$

11-7 迴歸的變異數分析與判定係數

迴歸的變異數分析：

迴歸的變異數分析列表：

變異來源	SS	f	MS	F
迴 歸	SSR	1	MSR	$F=\dfrac{\text{MSR}}{\text{MSE}}$
誤 差	SSE	$n-2$	MSE	
總 和	SST	$n-1$		

$$\begin{aligned}\Sigma(Y-\overline{Y})^2 &= \Sigma(\hat{Y}-\overline{Y})^2+\Sigma(Y-\hat{Y})^2\\ &= b^2\Sigma(X-\overline{X})^2+\Sigma(Y-\hat{Y})^2\end{aligned}$$

例題 11-14

X	6	4	10	2	8
Y	7	1	9	3	5

試以 $\alpha=0.05$，檢定其母體迴歸直線 $E(Y|X)=\alpha+\beta X$ 之 β 是否為 0。

(a) 以迴歸之變異數分析表（F 檢定）進行，並說明其涵義。

(b) 以 t 檢定進行。

解：(a) ① 假設：$H_0 : \beta = 0$; $H_1 : \beta \neq 0$

② 顯著水準：$\alpha = 0.05$

③ 放棄域：$F > F_{(0.95,1,3)} = 10.128$

④ 計算：

$$\sum X = 30 , \sum Y = 25 , \sum X^2 = 220 , \sum Y^2 = 165 ,$$

$$\sum XY = 182 , n = 5 , \overline{X} = 6 , \overline{Y} = 5$$

$$\text{SST} = \sum (Y - \overline{Y})^2 = \sum Y^2 - n\overline{Y}^2 = 165 - (5)(5)^2 = 40$$

$\text{MSE} = \hat{\sigma}^2 = 4.8$ (詳見 11-5 之例題 11-10 的計算)

$$\therefore \quad \text{SSE} = (5-2)(4.8) = 14.4$$

$$\text{SSR} = \text{SST} - \text{SSE} = 40 - 14.4 = 25.6$$

或 $\text{SSR} = b^2 \sum (X - \overline{X})^2$

$$= b^2 [\sum X^2 - n\overline{X}^2]$$

$$= (0.8)^2 (40)$$

$$= 25.6$$

⑤ 迴歸分析之變異數分析表：

變異來源	SS	f	MS	F
迴　　歸	25.6	1	25.6	$F=\dfrac{25.6}{4.8}=5.33$
誤　　差	14.4	3	4.8	
總　　和	40	4		

⑥ 判斷：$\because 5.33<10.128$，落在接受域，差異不顯著，接受 H_0，即 β 可能爲 0，表示母體迴歸直線可能與橫軸平行，自變數 X 對依變數 Y 解釋能力低。

(2) t 檢定請參考 11-5 之例題 11-11 說明。其中 t 值 $=2.309$, $(t=2.309)^2 \ =5.33=F$ 值，故兩者方法所得結果相同。

11-7 迴歸的變異數分析與判定係數

判定係數：

係指依變數 Y 的總變異中，可由自變數 X 的變異(迴歸變異)所解釋的變異。

公式：

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}\ ,\ 0 \le r^2 \le 1$$

例題 11-15

試利用上例求：

(a) 判定係數並解釋其涵義。

(b) 非判定係數並解釋其涵義。

(c) 相關係數。

解：(a) $r^2=\frac{SSR}{SST}=\frac{25.6}{40}=0.64=64\%$，即依變數 Y 的變異可由自變數 X 的變異解釋 64%。

(b) $1-r^2=1-64\%=36\%$，即依變數 Y 的變異有 36% 無法由自變數 X 所解釋。

(c) $r=\pm\sqrt{r^2}=\pm\sqrt{0.64}=0.8$ $(\because b=0.8>0, \therefore r$ 取爲正數$)$。

另外我們亦可直接由相關係數 r 的公式計算

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{(5)(182) - (30)(25)}{\sqrt{5(220) - (30)^2}\sqrt{5(165) - (25)^2}}$$

$$= 0.8$$

11-8 複迴歸（選讀）

例題 11-16

味全公司想了解售價與廣告支出對銷售量的影響，乃蒐集十天的相關資料，如下所示：

銷售量 (Y)	60	75	80	92	96	100	110	105	120	125
售價 (X_1)	9	8	8	7	6	6	5	5	4	4
廣告支出 (X_2)	2	3	4	5	5	5	6	6	7	8

試求：(設 $\alpha=0.05$，母體迴歸模式為 $E(Y|X_1, X_2)=\alpha+\beta_1X_1+\beta_2X_2$)

(a) 複迴歸方程 $\hat{Y}=a+b_1X_1+b_2X_2$。

(b) 偏迴歸係數的估計標準誤。

(c) 迴歸的變異數分析表。

(d) 檢定 $\beta_1=\beta_2=0$ 是否成立？

(e) 檢定 $\beta_1=0$ 是否成立？

(f) 檢定 $\beta_2=0$ 是否成立？

(g) 複判定係數 R^2。

(h) 修正複判定係數 $\bar{R}^2$。

(i) 複相關係數。

(j) 偏判定係數。

(k) 偏相關係數。

解： 因爲複迴歸手算複雜，本題以統計軟體解題，請讀者參考。

(a) 複迴歸方程：$\hat{Y} = 99.353 - 5.471X_1 + 6.052X_2$

(b) 偏迴歸係數的估計標準誤，係指個別自變數的迴歸係數的估計標準誤：

$$S(b_1) = 1.526\,,\ S(b_2) = 1.491$$

(c) 迴歸的變異數分析表：其一般格式爲：

變異來源	SS	f	MS	F
迴　　歸	SSR	k	MSR	$F = \dfrac{\text{MSR}}{\text{MSE}}$
誤　　差	SSE	$n-k-1$	MSE	
總　　和	SST	$n-1$		

表中 k 爲自變數的個數。

本題迴歸的變異數分析為：

變異來源	SS	f	MS	F
迴　　歸	3,685.519	2	1,842.759	395.951
誤　　差	32.581	7	4.654	
總　　和	3,718.1	9	413.12	

(d) 檢定 $\beta_1=\beta_2=0$ 是否成立？

$$\begin{cases} H_0: \beta_1=\beta_2=0 \\ H_1: \text{不全為 } 0 \end{cases}$$

$$F=\frac{\dfrac{\text{SSR}}{k}}{\dfrac{\text{SSE}}{n-k-1}}=\frac{\dfrac{3{,}685.519}{2}}{\dfrac{32.581}{10-2-1}}=\frac{1{,}842.759}{4.654}=395.951$$

$\because\ 395.95>F_{(0.95,2.7)}=4.7374$

$\therefore$ 差異顯著，放棄 H_0，即 β_1,β_2 有可能不全為 0。

(e) 檢定 $\beta_1 = 0$ 是否成立？

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

$$t = \frac{b_1 - \beta_1}{S(b_1)} \sim t_{(1-\frac{\alpha}{2},\, n-3)}$$

$$t = \frac{-5.471 - 0}{1.526} = -3.585$$

$$\because\ -3.585 < -t_{(1-\frac{\alpha}{2},\, n-3)} = -t_{(0.975,\, 7)} = -2.365$$

∴ 差異顯著，放棄 H_0，即 β_1 有可能不為 0。

(f) 檢定 $\beta_2 = 0$ 是否成立？

$$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$$

$$t = \frac{b_2 - \beta_2}{S(b_2)} \sim t_{(1-\frac{\alpha}{2},\, n-3)}$$

$$t = \frac{6.052 - 0}{1.491} = 4.059$$

$$\because\ 4.059 > t_{(0.975,\, 7)} = 2.365$$

∴ 差異顯著，放棄 H_0，即 β_1 有可能不為 0。

(g) 複判定係數 R^2 ：複判定係數 R^2 是用來測定依變數 Y 的總變異可由所有自變數 X_i 組合所能解釋的能力，其公式為：

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{3{,}685.519}{3{,}718.10} = 0.9912$$

即銷售量的不同可由售價及廣告支出所影響，高達 99.12%。

(h) 修正複判定係數 $\overline{R}^2$ ：一般而言，當自變數增加時，R^2 會增大，使我們容易以為多增加一些自變數來解釋依變數 Y 的變異，是有用的。但可惜的是，增加一些不相干的自變數，R^2 仍然是增大，反而無法明確分析。因此 R^2 必須適當的調整，以修正複判斷係數 $\overline{R}^2$ 表示：

$$\overline{R}^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{4.654}{413.12} = 0.9887$$

(i) 複相關係數：複相關係數是測定依變數 Y 與自變數 X_i 之相關程度，以 R 表示，恆為正數 $(0 \le R \le 1)$

$$R = \sqrt{R^2} = \sqrt{0.9912} = 0.995 \text{ (設以未修正之 } R^2 \text{ 為例)}$$

(j) 偏判定係數：所謂**偏判定係數 (coefficient of partial determination)**，係指在已有的自變數之迴歸方程中，再加入一個自變數，可增加對依變數 Y 解釋變異的能力。(註：複判定係數 R^2 是表示全部自變數 X_i 對依變數 Y 變異的解釋能力)，其公式如下：

$r^2_{Y2\cdot1}$：表示迴歸直線已包括 X_1 的情況下，再加入 X_2 而能解釋依變數 Y 變異的比例：

$$r^2_{Y2\cdot1}=\frac{\mathrm{SSE}(X_1)-\mathrm{SSE}(X_1,X_2)}{\mathrm{SSE}(X_1)}=1-\frac{\mathrm{SSE}(X_1,X_2)}{\mathrm{SSE}(X_1)}$$

茲先求出 Y 對 X_1 的迴歸方程爲 $Y=167.196-11.435X_1$，其迴歸之變異數分析表如下：

變異來源	SS	f	MS	F
迴　　歸	3,608.817	1	3,608.817	264.189
誤　　差	109.283	8	13.660	
總　　和	3,718.1	9	413.122	

$$r^2_{Y2\cdot1}=\frac{\mathrm{SSE}(X_1)-\mathrm{SSE}(X_1\,,X_2)}{\mathrm{SSE}(X_1)}$$

$$=\frac{109.283-32.581}{109.283}$$

$$=0.702$$

當我們只考慮一個 X_1 變數時，其 SSE 爲 109.283，但當加入 X_2 後，SSE 降爲 32.581，即增加 70.2% 對銷售量的解釋變異。(比較 (c) 與 (j))

(k) 偏相關係數：將偏判定係數開方根，即得偏相關係數

$$r_{Y2\cdot1}=\pm\sqrt{r^2_{Y2\cdot1}}=\pm\sqrt{0.702}=0.838$$

(其正負號與其所對應之偏迴歸係數同，本題因 X_2 之偏迴歸係數爲正數，∴ 偏相關係數取正號)