第 3 章

診斷與矯正之測量

Diagnostics and Remedial Measures

本章內容

- 當簡單線性迴歸模型（2.1）被引用時，通常無法事前即正確判斷該模型是否合適，如線性或誤差項之常態分配性等
- 因此，對模型進行推論之前，應該先檢查一些有關於資料是否合適於模型之問題。
- 對於一些模型中常見的適當性問題，透過圖形與正式的檢定方法來進行討論，同時也將介紹一些有關於資料不合適於模型下的矯正方法。
- 本章基本上雖然是以簡單線性迴歸模型（2.1）作為對象，但是其基本原則仍然可以適用到整本書中所討論的全部模型 。

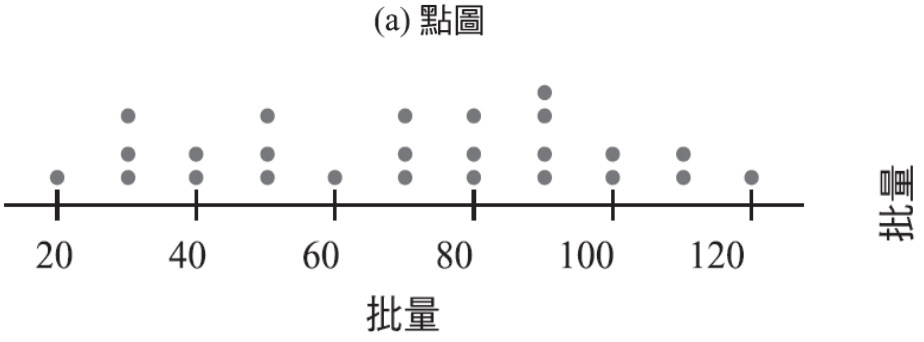
3.1 預測變數的診斷

Diagnostics for Predictor Variable

- 透過預測變數之圖形診斷，該診斷訊息可以提供有關於離群的X值對所配適之迴歸直線之影響，同時研究該圖形X水準之全距與集中程度，將有助於迴歸分析的適用範圍確認。

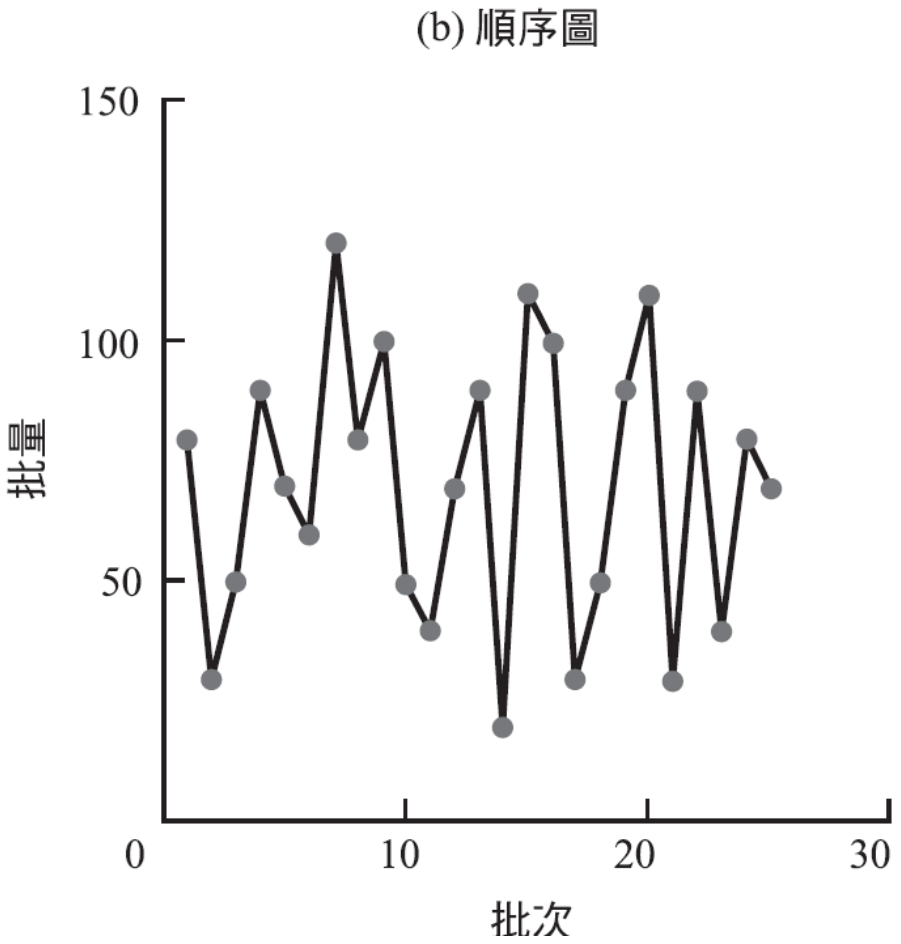
- 圖3.1a 是圖1.10 Toluca公司案例中有關批量的**點圖（dot plot）**，對於觀測值不多的情形下相當有用，圖3.1a 的點圖顯示批量X的最小值與最大值分別為20與120，亦即批量之大小均在此一範圍內，同時並無離群值的情形發生

圖 3.1
MINITAB 與 SYGRAPH 對於預測變數之診斷|



＊可利用EXCEL 繪出直方圖

- 第二種有關預測變數之圖形診斷工具為**順序圖（sequence plot）**，圖3.1b 是Toluca公司案例中有關批量的時間順序圖。
- 將圖中各個觀測點連接起來可以更清楚地顯示時間順序。如果有順序性的產生可以利用順序圖以提供診斷訊息。



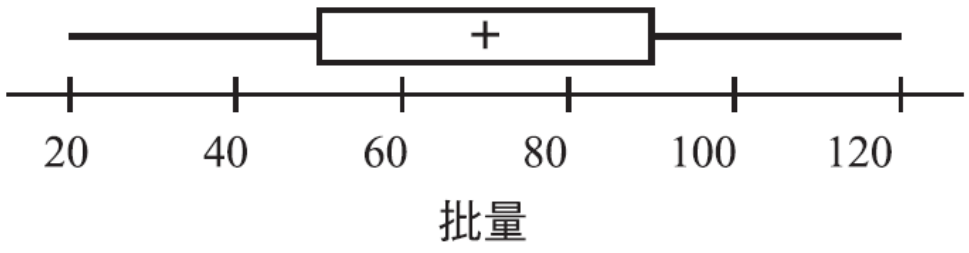
＊可利用EXCEL 繪出順序圖

- 圖3.1c 為**莖葉圖（stem-and-leaf plot）**，提供了與次數直方圖相同之訊息，圖形特別標示出中位數（M）以及第一四分位數與第三四分位數（H）的位置。
- 圖3.1d 的**盒型圖（box plot）**顯示最大批量、最小批量、批量的第一四分位數、批量的第三四分位數與批量的中位數，圖中可以看出一半的批量資料介於50至90之間，而相當對稱於中位數，盒型圖對於大量的觀測值時特別有用。

(c) 莖葉圖

2	0
3	000
4	00
5H	000
6	0
7M	000
8	000
9H	0000
10	00
11	00
12	0

(d) 盒型圖



＊可利用EXCEL 繪出直方圖

3.2 殘差 Residuals

- 直接對於反應變數Y畫出診斷圖，其實對於迴歸分析之診斷幫助不大，因為反應變數之觀測值將受到預測變數水準高低之影響，因此對於反應變數Y之診斷，通常是間接經由殘差分析達成。
- 在定義(1.16)中，殘差 e_i 是觀測值 Y_i 與配適值之間的差：

$$e_i = Y_i - \hat{Y}_i \tag{3.1}$$

殘差可以被視為觀測誤差，而迴歸模型中未知的真實誤差則是：

$$\varepsilon_i = Y_i - E\{Y_i\} \tag{3.2}$$

- 在迴歸模型(2.1)中誤差項假定誤差項 ε_i 假定為獨立之常態隨機變數平均數為 0，變異數為常數 ，如果資料滿足模型之假設，則殘差 e_i 應該會反映出誤差項 ε_i 的性質，這就是**殘差分析**(residual analysis)的基本想法，也是一種用來檢驗統計模型之適當性的有效方法。

殘差之性質

- **平均數**

在簡單線性迴歸模型(2.1)中，n個殘差e_i的**平均數**可以透過(1.17)得到：

$$\bar{e} = \frac{\sum e_i}{n} = 0 \tag{3.3}$$

- **變異數**

迴歸模型(2.1)中n個殘差e_i的變異數定義如下：

$$s^2 = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE \tag{3.4}$$

如果模型適合，則 *MSE* 為誤差項變異數的不偏估計量。

- **不獨立性**

因為殘差 e_i 牽涉到來自相同迴歸直線的配適值 $\hat{Y}_h$ 之影響，所以它們並非相互獨立之隨機變數(亦即 e_i 的總合必然為 0)。如果樣本數夠大，則殘差 e_i 間的不獨立性將相對地不重要而可以被忽略。

半學生化殘差

- 有時殘差分析會將殘差經過 標準化 而產生較佳之效果，由於誤差項 ε_i 的標準差 σ 一般習慣用 $\sqrt{MSE}$ 估計，可以考慮如下之標準化：

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}} \qquad (3.5)$$

- 如果 $\sqrt{MSE}$ 是殘差 e_i 之標準差的估計值，則 e_i^* 稱為**半學生化之殘差**，不過實際上 e_i 的標準差複雜且各 e_i 的標準均不相同，$\sqrt{MSE}$ 只是殘差 e_i 之標準差的近似值，因此 (3.5) 的 e_i^* 稱為**半學生化之殘差** (semistudentizd residual)
- 不過無論是學生化之殘差（第五章）或是半學生化之殘差在辨識離群資料時均很有用。

透過殘差研究發現模型偏離

六種利用殘差來檢驗**簡單線性迴歸模型 (2.1)**

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

中有關常態誤差假設之情形：

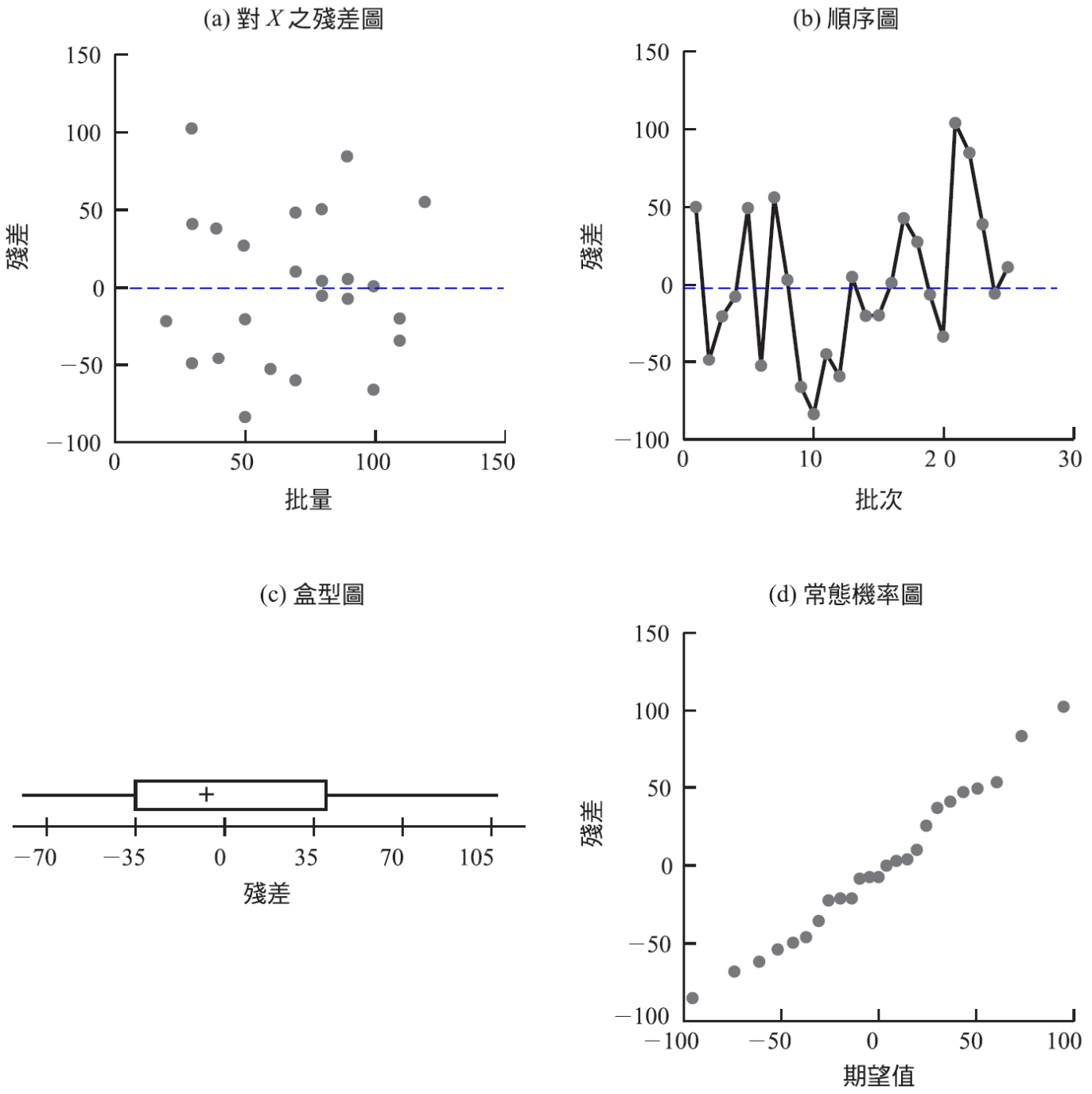
1. 迴歸函數不是直線形式
2. 誤差項不滿足常數變異數
3. 誤差項不獨立。
4. 除了少數離群值外，模型配適適當。
5. 誤差項不滿足常態
6. 在模型中沒有考慮到一個或多各重要的預測變數。

3.3 殘差診斷 Diagnostics Residuals

- 進行一些非正式的殘差診斷圖之討論，以提供上述六種偏離簡單線性迴歸模型(2.1)是否存在的情形一些訊息，
- 包括以下幾種殘差或半學生化之殘差圖：

1. 對預測變數之殘差圖。
2. 對預測變數之殘差絕對值或殘差平方值圖。
3. 對配適值之殘差圖
4. 殘差對時間或其他順序圖。
5. 對於被忽略之預測變數的殘差圖。
6. 殘差盒型圖。
7. 殘差的常態機率圖(normal probability plot)。

圖 3.2
MINITAB 與 SYGRAPH 之殘差診斷圖－Toluca 公司案例。



- 在Toluca公司案例中，圖3.2畫出了對預測變數之殘差圖、對時間順序之殘差圖、殘差盒圖與殘差的常態機率圖，這些圖形都能夠顯示使用迴歸模型(2.1)適當性

非線性迴歸函數

可以透過對預測變數之殘差圖來判斷線性迴歸函數是否可以使用於所要分析之資料，或者探討對配適值之殘差圖，雖然迴歸函數的非線性問題之研究也可以利用散佈圖來看出，但是並不如殘差圖有用，圖3.3a 的散佈圖與所配適之迴歸直線

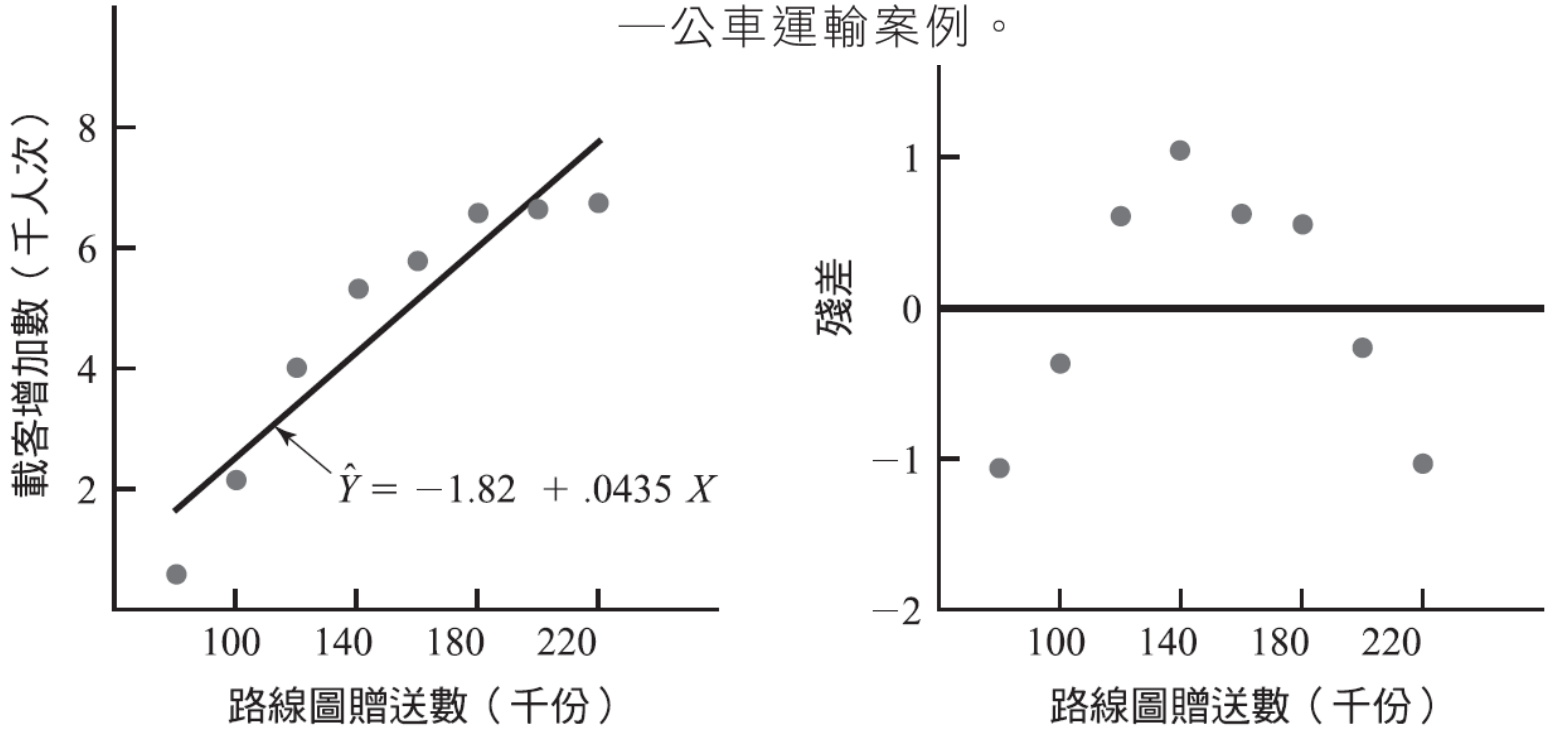


圖 3.3
散佈圖與殘差圖說明非線性迴歸函數—公車運輸案例。

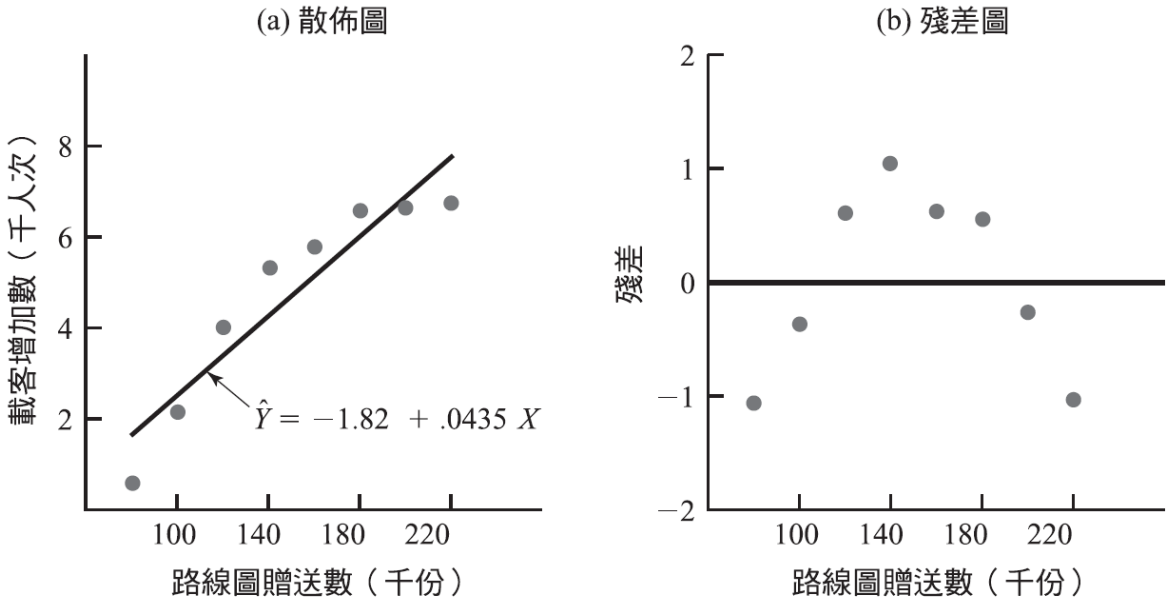
- 表 3.1資料來自對於八個城市所進行的贈送公車路線圖與公車載客數關係之研究，X 表示免費送出之公車路線圖份數，Y 表示每日尖峰時刻之平均載客增加數，原始資料與所配適之結果顯示於表 3.1 的第一、二、三欄位，同時**圖 3.3b** 也顯示了直線迴歸函數的不適合性（如下頁）。

表 3.1
路線圖贈送數與載客增加數—公車運輸案例。

(1)	(2)	(3)	(4)
城市	載客增加數（千人次）	路線圖贈送數（千份）	配適值	殘差
i	Y_i	X_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i = e_i$
1	.60	80	1.66	−1.06
2	6.70	220	7.75	−1.05
3	5.30	140	4.27	1.03
4	4.00	120	3.40	.60
5	6.55	180	6.01	.54
6	2.15	100	2.53	−.38
7	6.60	200	6.88	−.28
8	5.75	160	5.14	.61

$$\hat{Y} = -1.82 + .0435\,X$$

- 圖3.3b 是表3.1第四欄位的殘差對預測變數之關係圖形，該圖比圖3.3a 的散佈圖更能顯示出直線迴歸函數的不適合性，
- 此處需特別注意的是殘差以對稱的方式偏離 0 的位置
- 當批量較大或是較小時，殘差為負值，而當批量大小適中時，殘差為正值。



城市 i	(1) 載客增加數（千人次） Y_i	(2) 路線圖贈送數（千份） X_i	(3) 配適值 $\hat{Y}_i$	(4) 殘差 $Y_i - \hat{Y}_i = e_i$
1	.60	80	1.66	−1.06
2	6.70	220	7.75	−1.05
3	5.30	140	4.27	1.03
4	4.00	120	3.40	.60
5	6.55	180	6.01	.54
6	2.15	100	2.53	−.38
7	6.60	200	6.88	−.28
8	5.75	160	5.14	.61

$\hat{Y} = -1.82 + .0435\ X$

- 從上面兩個圖形中可以發現殘差圖比散佈圖更具有某些優點
  - 首先殘差圖可以更方便容易地使用來檢驗模型的適切性
  - 其次有時散佈圖會因座標尺度的大小而不易判斷模型的適切性，特別是當值線斜率很大時，但是殘差圖在此種情形下並不會因此受到任何系統性之影響。

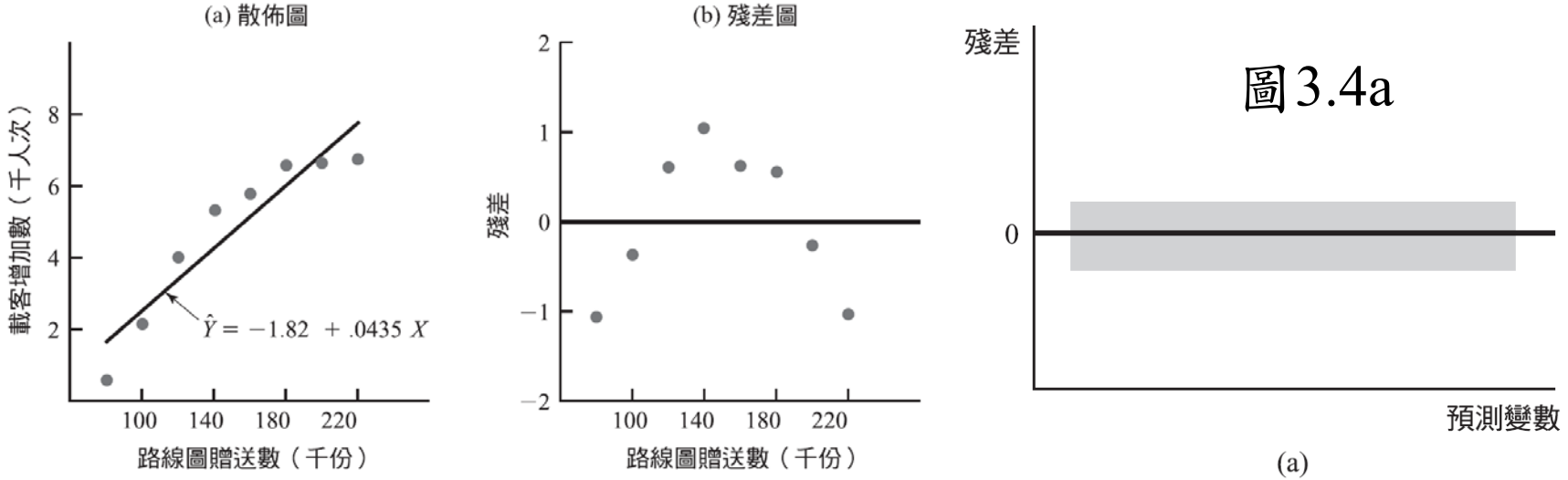
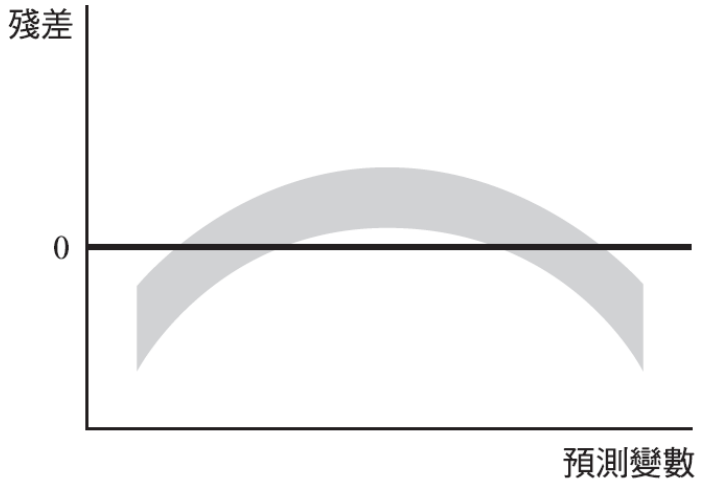


圖3.4a

- 當線性迴歸模型適當時，殘差對預測變數之離型將如圖3.4a所示，此時殘差將落在以 0 為中心線的區塊內並且不會有任何系統性的型態顯示。

- 圖3.4b 為模型偏離了線性迴歸的情形，此時可能必須考慮採用曲線迴歸函數，在圖中出現殘差的正負具有系統性地類型，是偏離線性。
- 在簡單直線迴歸模型中對於配適值 $\hat{Y}_h$ 之殘差圖所能提供之訊息，等同於預測變數X之殘差圖，因此不需要另外畫出對預測變數X之殘差圖，因為配適值 $\hat{Y}_h$ 是預測變數值 X_i 的直線函數，不過，當模型為曲線迴歸或複迴歸之情形下，同時畫出對於配適值 $\hat{Y}_h$ 之殘差圖以及對預測變數X之殘差圖，將有助診斷模型



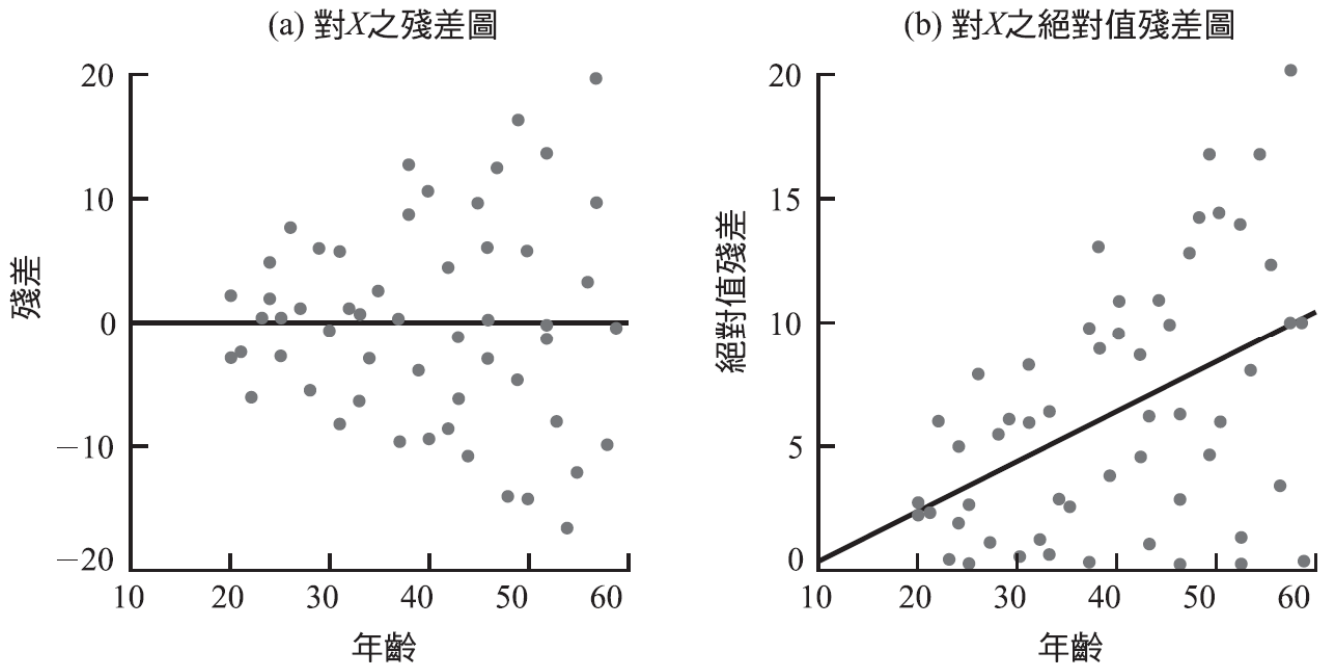
(b)

圖 3.4
殘差圖雛型

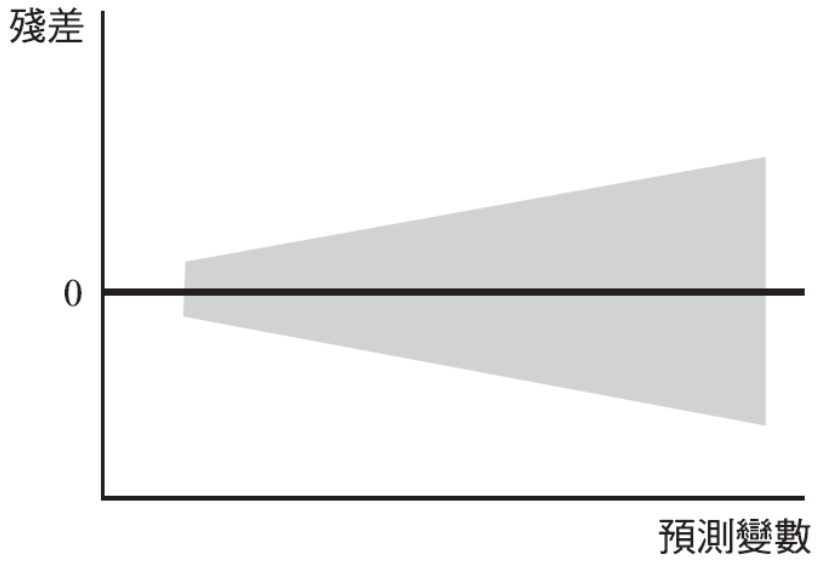
非常數誤差變異數

- 對於配適值或預測變數分別畫出殘差圖，不僅可以用來檢驗直線迴歸函數是否適當，也可以用來檢查誤差項之變異數是否常數，殘差圖3.5a 來自一個關係研究血壓(Y)與女性年齡(X)關係的研究案例，圖中可看出女性年齡越大，殘差散佈範圍越廣，由於血壓與女性年齡之關係為正向比例，所以年紀較大之女性其誤差項之變異數也較年輕女性大。

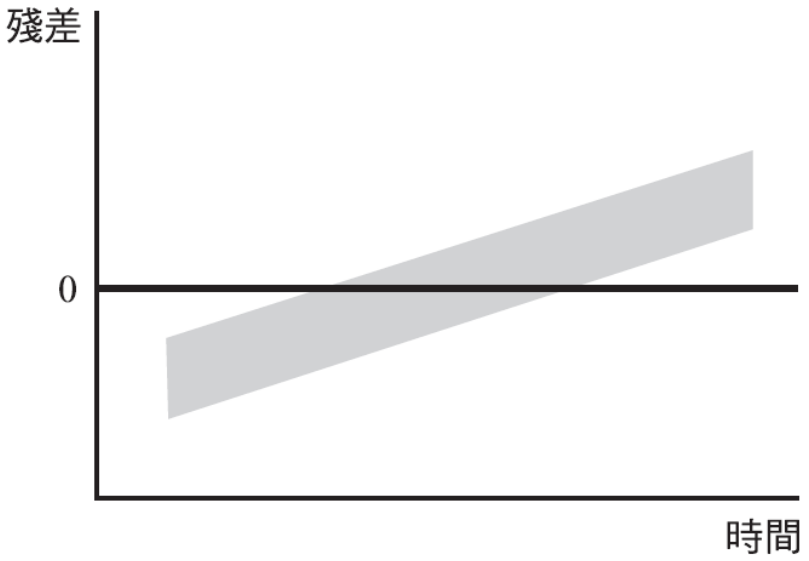
圖 3.5
殘差圖解釋非常數誤差變異數。



- 再回到**圖3.4a** 的情形來看，該圖中所顯示的是當誤差項變異數為常數之雛形，而圖3.2a Toluca公司正符合此一情形，所以在該案例中之誤差項變異數為常數。
- **圖3.4c**的情形則是當誤差項變異數預測變數X遞增，即誤差項變異數偏離常數而越來越大，呈現如一個「傳聲筒」之形狀，例如上圖3.5a 血壓之案例，當然有些時候會出現誤差項變異數偏離常態而越來越小，或出現其他複雜形狀之變化。

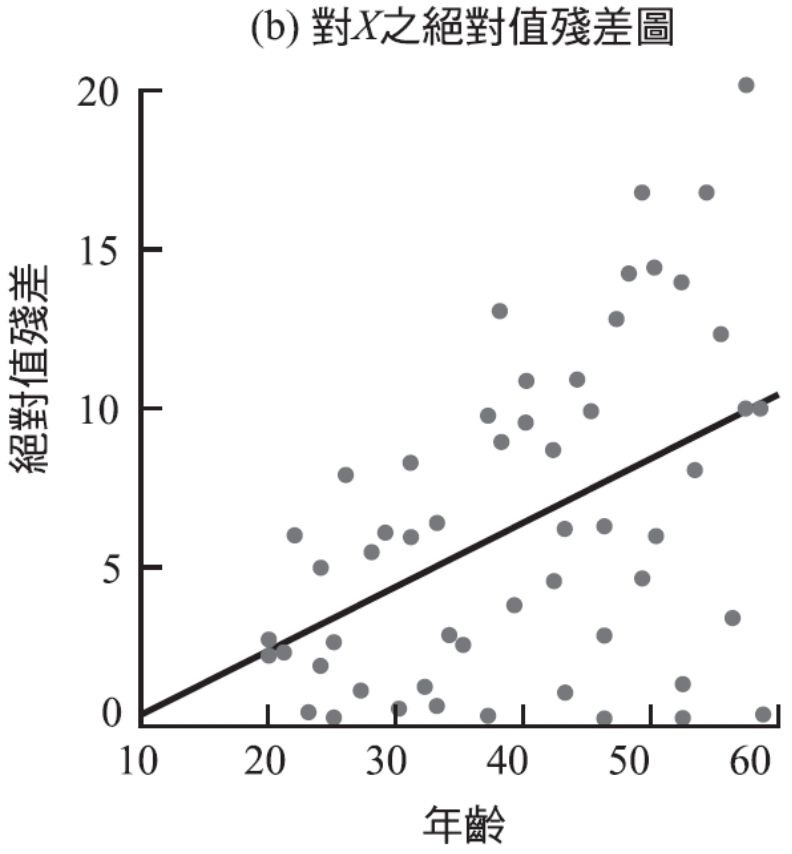


(c)



(d)

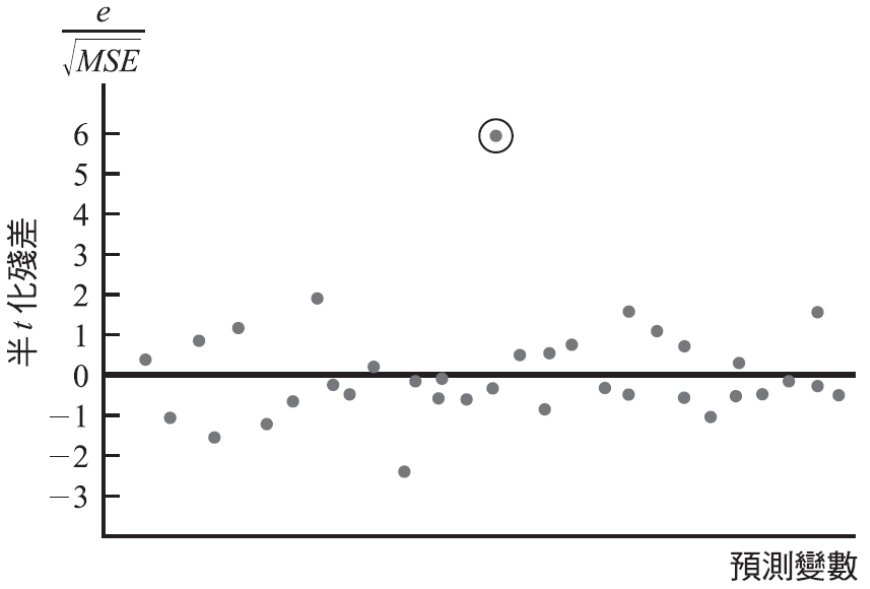
- 透過殘差取絕對值或平方殘差後，對配適值$\hat{Y}_h$或預測變數X分別畫出殘差圖，也有助於診斷誤差項變異數是否為常數，因為當在檢驗誤差項變異數偏離常數時，殘差的正負符號是不具有任何意義的。
- 當觀測個數不多時特別有用，因為此時有關誤差量都會出現在水平的零以上，所以更容易看出殘差大小隨預測變數X或配適值$\hat{Y}_h$而改變之情形。
- 圖3.5b 為血壓與女性年齡關係的研究案例中，對於殘差取絕對值與年齡X的關係圖，透過該圖更清楚地顯示了高齡女性較有的絕對值殘差量。



離群值

- 所謂的離群值就是極端的觀測值，可以藉由對配適值 $\hat{Y}_h$或預測變數X分別畫出殘差圖來看出離群值其殘差之分布情形，或是藉由**盒形圖、莖葉圖與點圖**以提供診斷訊息，
- 以半學生化殘差圖（或學生化殘差圖）來辨識離群值的觀測值較為有用，透過該殘差圖可以很明確找出距離0的數個標準差之標準差之離群值，如果觀測的樣本夠多的話，一般是取半學生化殘差後之絕對值大於四作為離群值之定義。

- 圖3.6為一各半學生化殘差圖，圖中被圈出之點便是所謂的離群值，此一離群值距離配適值約有六個標準差之遠



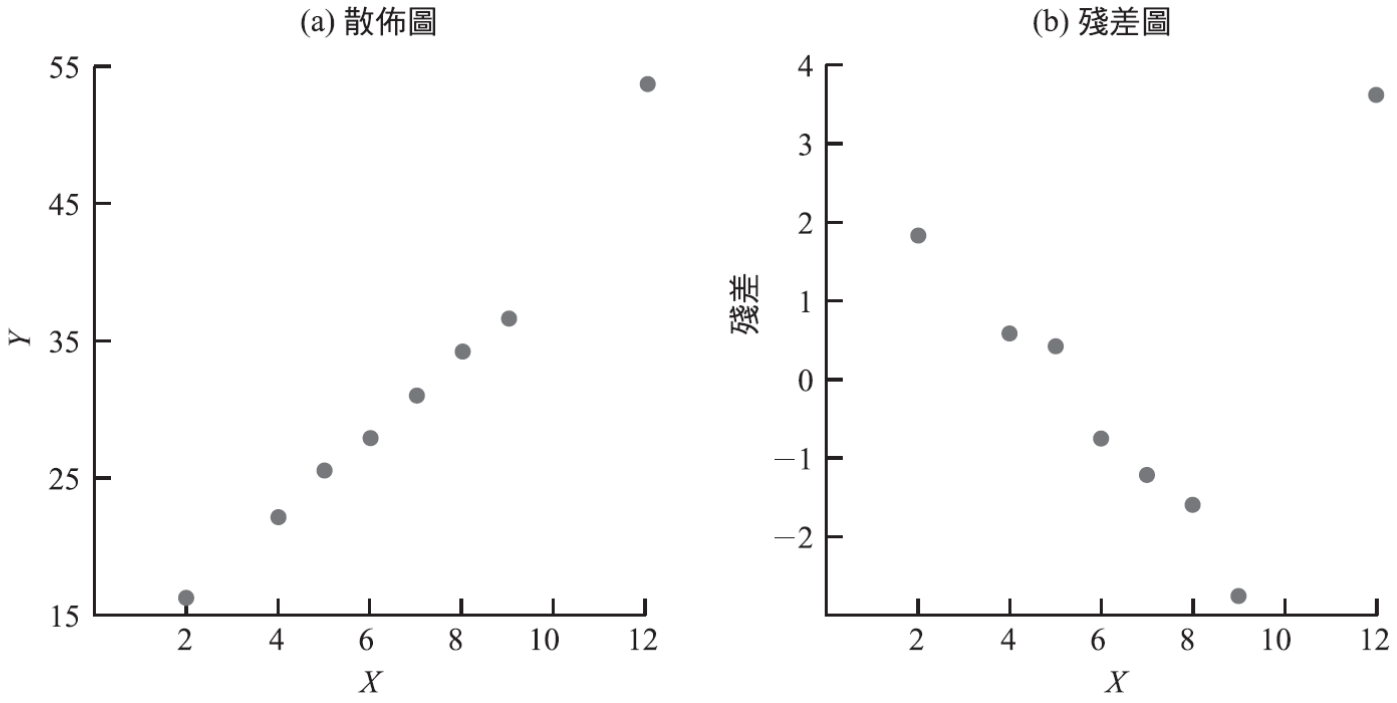
離群值

- 離群值的出現將造成許多困擾，首先必須懷疑此一離群值是否是因為某種錯誤或外在因素而引起，或許可以考慮放棄該離群值，以使得平方離差和最小，因為根據最小平方法的原理所配適出的迴歸直線是否會因此而偏向該離群值，
- 如果確實是因某種錯誤或外在因素引起，這樣配適出的結果將造成誤解。
- 但是從另一方面來看，離群值也可能隱含了一些重要的訊息，例如當離群值的產生是因為有一個預測變數沒有被放入迴歸模型中，
- 所以建議一個安全規則，就是只有當存在直接證據顯示該離群值是因為紀錄錯誤、計算錯誤、設備故障或其他類似情形下，才能將它放棄。

離群值

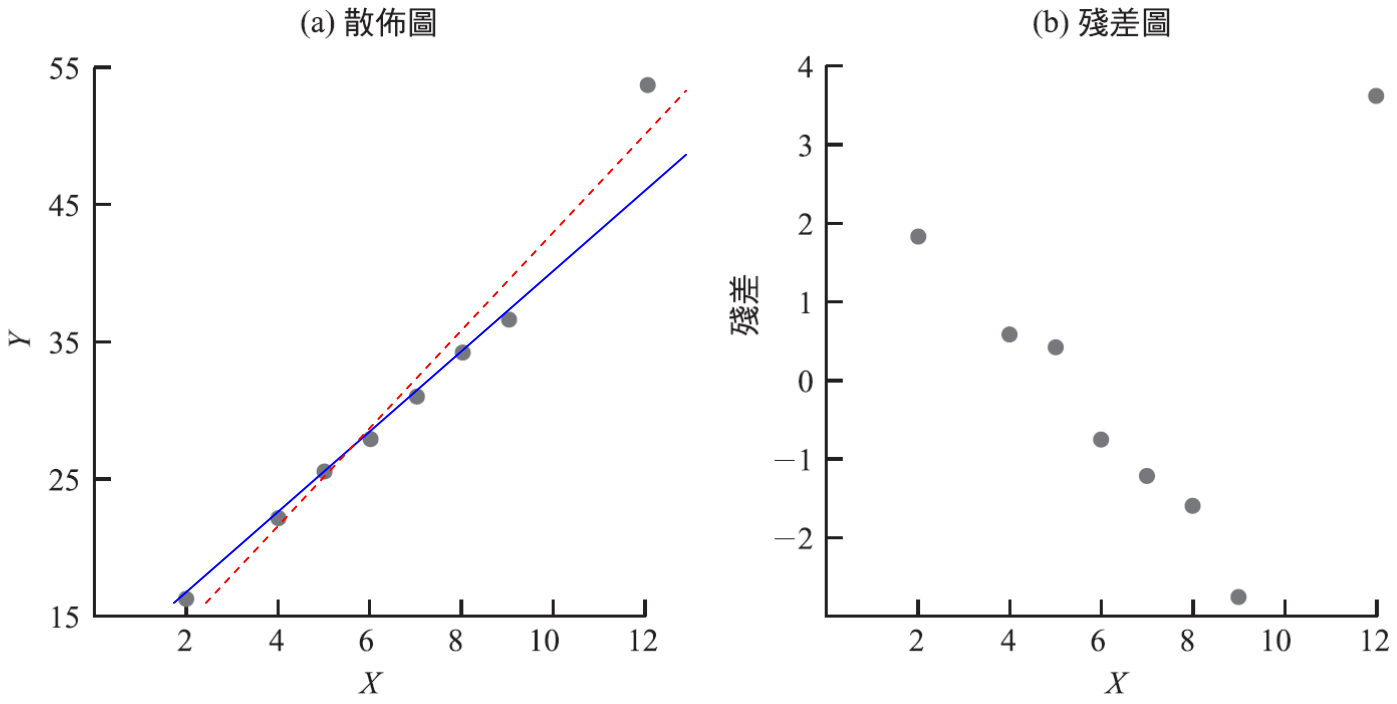
- 當觀測樣本數不大時所配適出之迴歸直線被發現有離群值時，所配適出之迴歸直線將被嚴重扭曲，
- 此時殘差圖除了標示出離群值外，也會告訴我們不適當的模型配適。
- 圖3.7 便說明了此種情形。

圖 3.7
迴歸直線配適因一個離群值而造成殘差圖異常的不良影響。



- 圖3.7a 的散佈圖顯示了多數的觀測值均落在統計關係的直線附近，但是有一個離群值遠離此一直線，如果將此一資料全部考慮進去配適模型，則此一離群值顯然會造成迴歸直線的偏移，以致影響了其它關測值樣本與迴歸直線間發生系統性之變化，所顯示出的卻是不適當的模型配適，如圖3.7b所示。

圖 3.7
迴歸直線配適因一個離群值而造成殘差圖異常的不良影響。

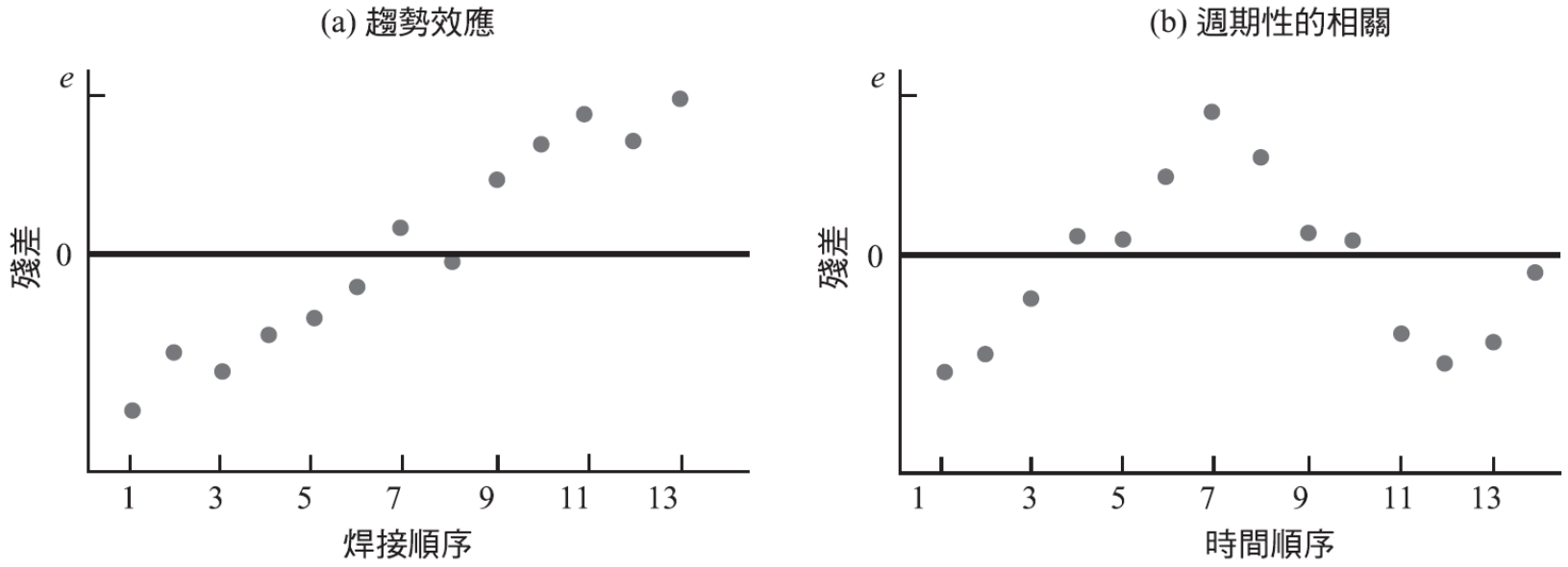


非獨立之誤差項

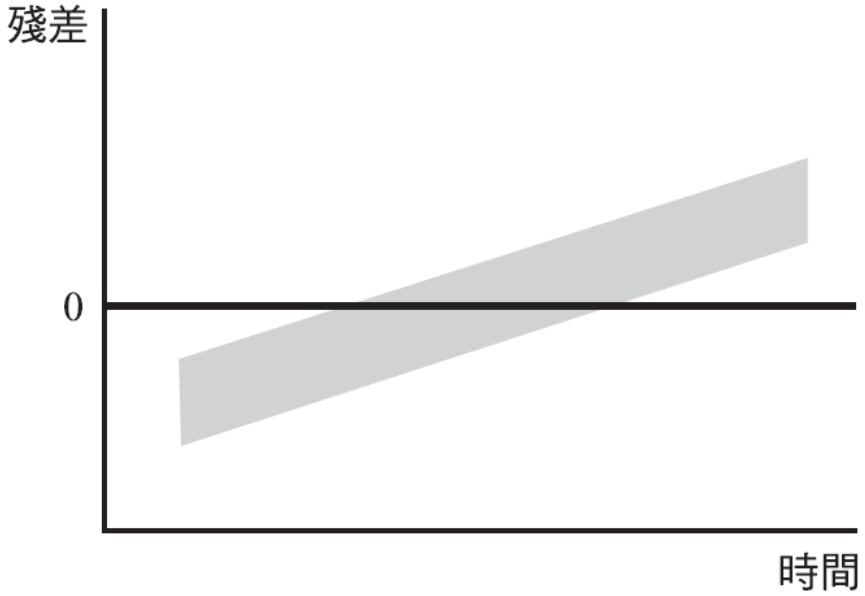
- 當資料的取得是根據時間或某一種形式的順序，如相鄰之地理位置，此時殘差順序圖將是一種好的診斷工具，其目的在觀察順序鄰近的誤差項間是否存在關係。

- 圖3.8a 為一個研究焊接面直徑(X)與焊接點剪刀強度(Y)關係之案例，圖中顯示殘差的時間順序資料彼此有顯著的相關性存在，實驗剛開始多為負值的殘差，但是後來便轉為正值的殘差
- 所以其中隱含了與時間順序的效果，可能焊接的學習效果或是設備的改變，而造成後來地焊接有較大的剪刀強度。

圖 3.8
殘差的時間順序圖說明了誤差項間的非獨立性。



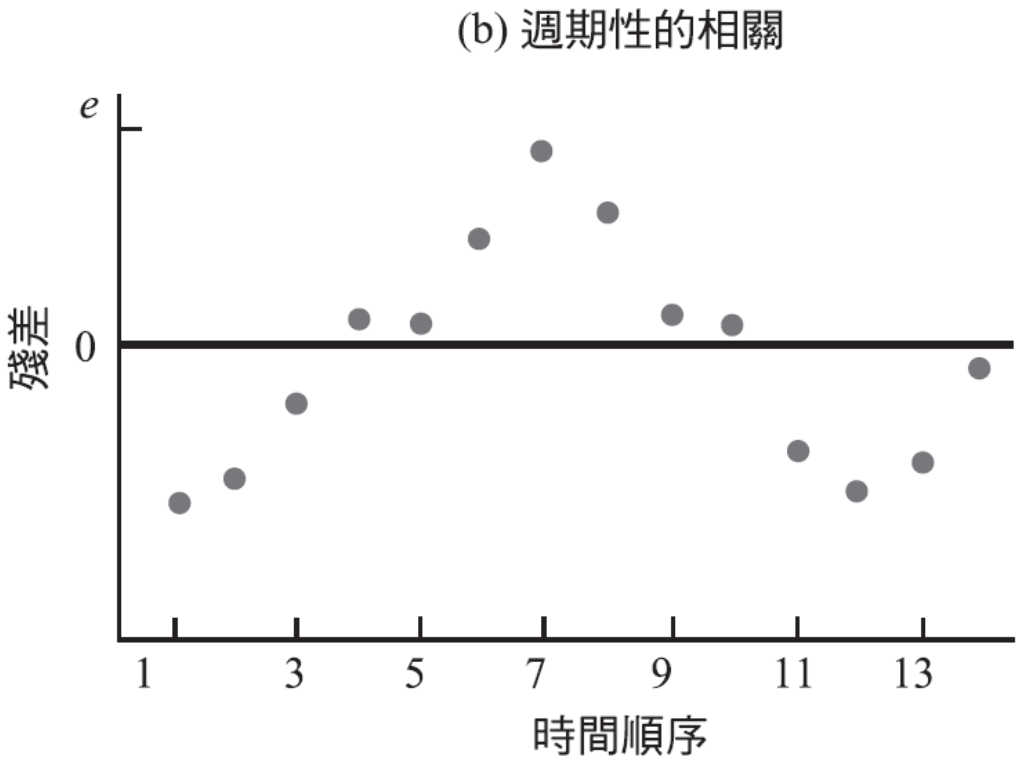
- 圖3.4d 有關時間相關趨勢效果的殘差圖雛形，有時候可以將誤差間的相關性看成是某一個重要的變數(例如:時間)被忽略了。



(d)

- 另一種有關誤差項間不獨立的類型如圖3.8b，誤差間的相關性存在著週期性而無趨勢效果；當誤差項獨立時，順序圖中的殘差可以預測期會在基準線 0 的上下間隨機地跳動。如圖 3.2b Toluca 公司的情形；

- 如果沒有出現基準線 0 第上下間隨機地跳動，可能的情形是在基準線 0 的上下間交錯次數過多或過少，實務上，交錯次數過多的情形較不常發生，常見的例子是如圖3.8a 焊接的案例中交錯次數過少。



•透過對X畫出之殘差圖如3.3b 的公車運輸案例，可能並沒有隨機性的散佈情形，然而問題並不在誤差項缺乏隨機性，而是如圖3.3a 所描述的迴歸函數配適不良。

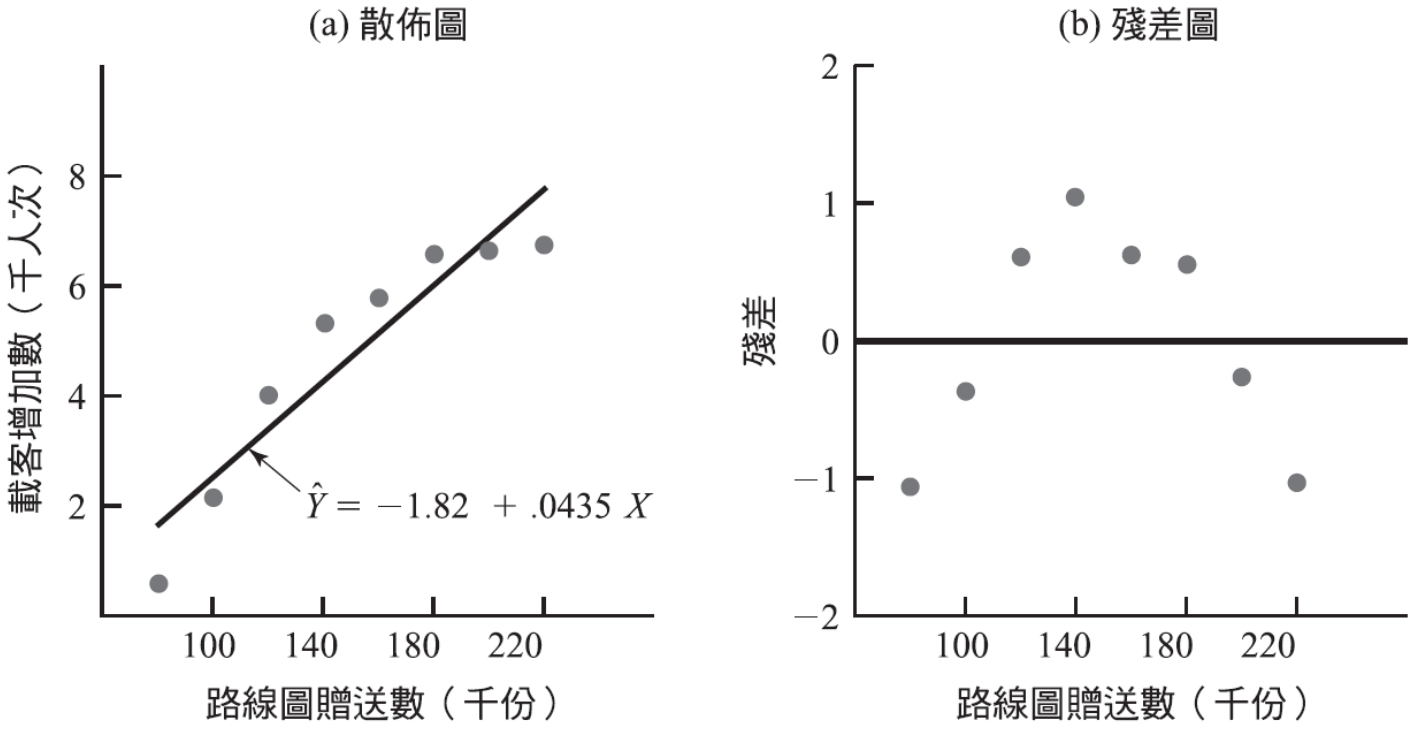


圖 3.3
散佈圖與殘差圖說明非線性迴歸函數—公車運輸案例。

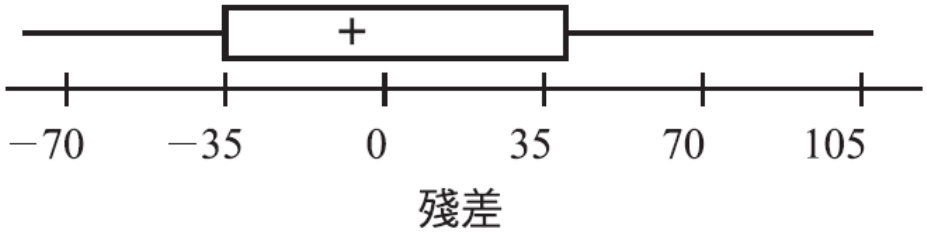
非常態之誤差項

- 有關誤差項的常態性可以由某些非正式的殘差圖形來檢驗。

1. 分配圖

- 透過殘差的盒形圖可以幫助確認殘差的對稱性與發現可能存在的離群值，以圖3.2c Toluca公司的情形為例，圖中並未有偏離殘差對稱性的情形發生，

(c) 盒型圖



- 利用殘差的直方圖、點圖或是莖葉圖可以幫助我們用來偵測常態假設的偏離程度，不過如果想透過這些圖形來表現誤差項地分配情形，則用於迴歸研究之樣本數必須夠大才可以。

2.比較次數

- 在樣本數夠大(Z分佈）的情形下考慮另一種方式來檢驗常態性假設，理論上約有 68% 的殘差落在1個 $\pm (MSE)^{1/2}$ 範圍內，或有 90% 的殘差會落在 $\pm 1.645\ (MSE)^{1/2}$ 範圍內。
- 當樣本數適中時，可以利用所對應的 t 值進行比較，如Toluca公司案例表3.2的第一欄位是殘差在圖 2.2，$\sqrt{MSE} = 48.82$ 根據 t 分配可得知約有 90% 的殘差會落在 t（0.95;23）= 1.714 *（48.82）範圍內。

亦即 -83.68與 83.68 之間，在實際的資料中有22個殘差值（約佔88%）落在該範圍內，

表 3.2
常態假設下殘差與對應之期望值—Toluca 公司案例。

批次 i	(1) 殘差 e_i	(2) 等級 k	(3) 常態假設下所對應之期望值
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
...	...	...	...
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

- 類似的情形，約有 60% 的殘差會落在 -41.89 與 41.89 之間，而實際的資料中有 52% 在該範圍內，所以本案例算是相當符合常態假設。

3.常態機率圖

- 還有一種方法可以用來檢驗常態性假設，就是畫出殘差的常態機率圖(normal probability plot)，圖中的每一個殘差都對應到它在常態分配下之期望值，當圖形越接近直線表示常態性假設越符合，反之，如果明顯偏離直線則表示誤差的分配悖離常態。

- 一樣是Toluca公司的案例，表3.2的第一個欄位是殘差值，利用迴歸模型(2.1)中，誤差項期望值為 0，標準差之估計值為 $\sqrt{MSE}$ ，統計理論上可以得證，n 個樣本來自一個期望值為 0，標準差之估計值為 $\sqrt{MSE}$ 之常態分配，則樣本中第 k 小之觀測值的期望值近似如下：

$$\sqrt{MSE}\left[z\left(\frac{k-.375}{n+.25}\right)\right] \qquad (3.6)$$

  z(A) 之意義表示標準常態分配的第(A)100個百分位數。
  例如：見表 B-1 (PB-3)，Z分佈，Z(0.9222)=1.4+0.2=1.42

- 利用(3.6)近似公式可計算殘差在常態假設下的期望值，參考P 3-12 例題說明殘差的「等級」(rank)，及在常態分配的假設下此殘差之「期望值」。

表 3.2
常態假設下殘差與對應之期望值—Toluca 公司案例。

批次 i	(1) 殘差 e_i	(2) 等級 k	(3) 常態假設下所對應之期望值
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
...	...	...	...
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

- 表 3.2 的第二個欄位是殘差的等級(rank)，而最小的殘差等級定為 1；

- 第一批生產殘差值為e_1 =51.02，它的等級排在22，表示在25各殘差中，它是第22小而k=22，表2.1有MSE=2.384，所以

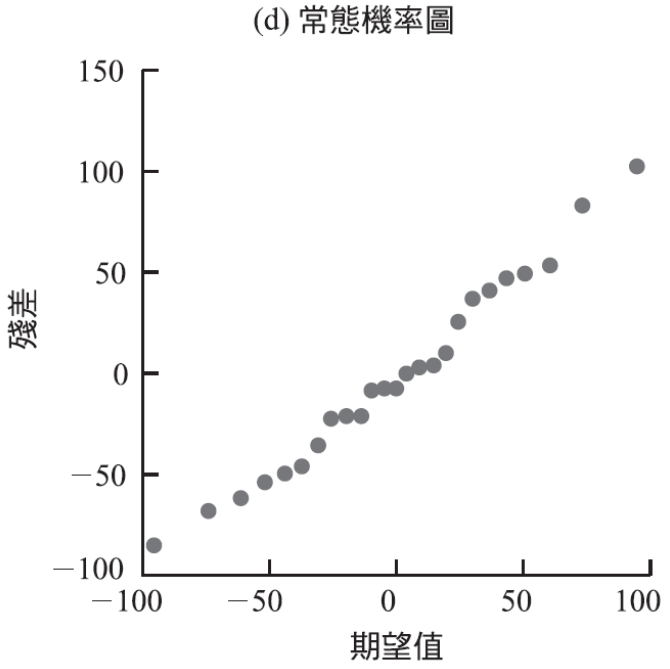
$$\frac{k-0.375}{n+0.25}=\frac{22-0.375}{25+0.25}=\frac{21.625}{25.25}=0.8564$$

- 此殘差之期望值在常態分配的假設下為:

$$\sqrt{2384}[z(0.8564)]=\sqrt{2384}(1.064)=51.95$$

	(1)	(2)	(3)
批次 i	殘差 e_i	等級 k	常態假設下所對應之期望值
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
...	...	...	...
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

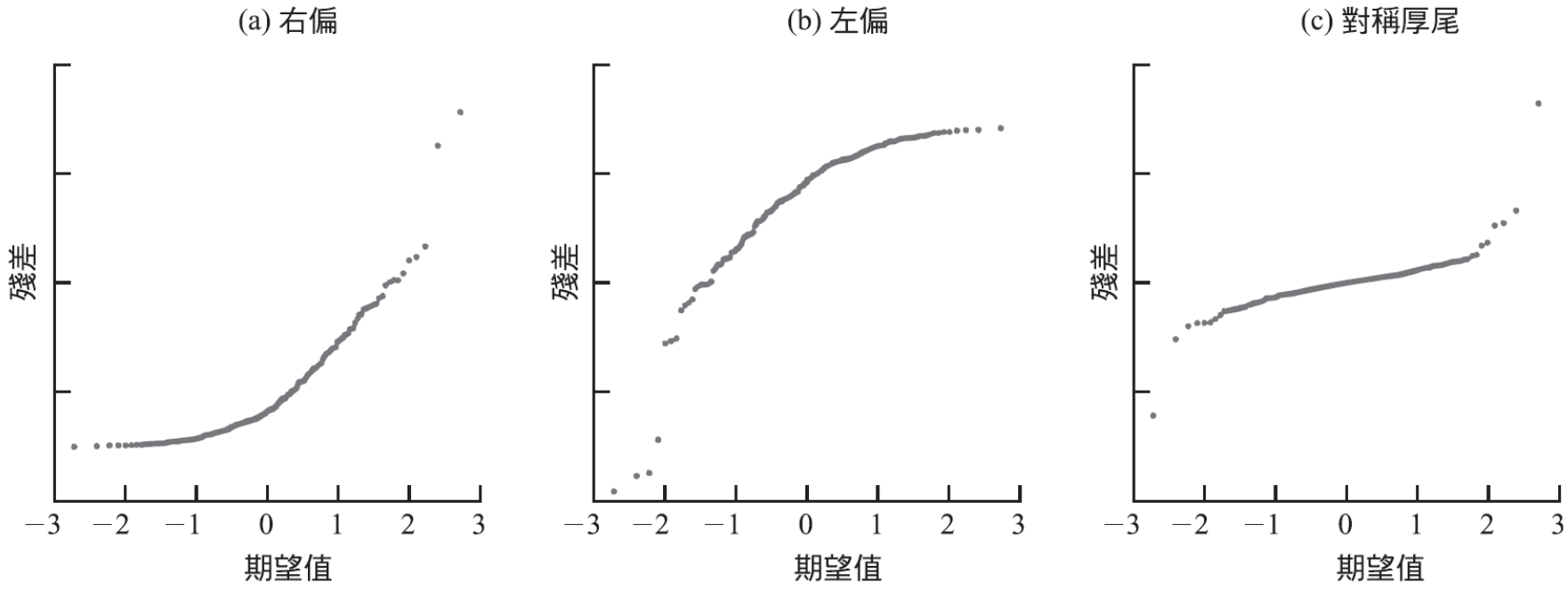
- 表3.2 的第三個欄位是 25個殘差在常態分配下之期望值
- 圖3.2d 則是殘差與所對應期望值之關係圖，圖中之各點連線將近似於一條直線，表示誤差地分配符合常態性假設。



- 圖3.9 表示當誤差項之分配明顯悖離常態分配的假設時，三種可能的常態機率圖。圖3.9a 誤差項分配高度右偏，圖形之凹口向上，圖3.9b 是誤差項分配高度左偏時，圖形之凹口向下；圖3.9c 雖然是對稱圖形，不過尾巴過厚，表示此分配在左右兩尾的部分機率大於常態分配。

圖 3.9

當誤差項悖離常態分配時的常態機率圖。

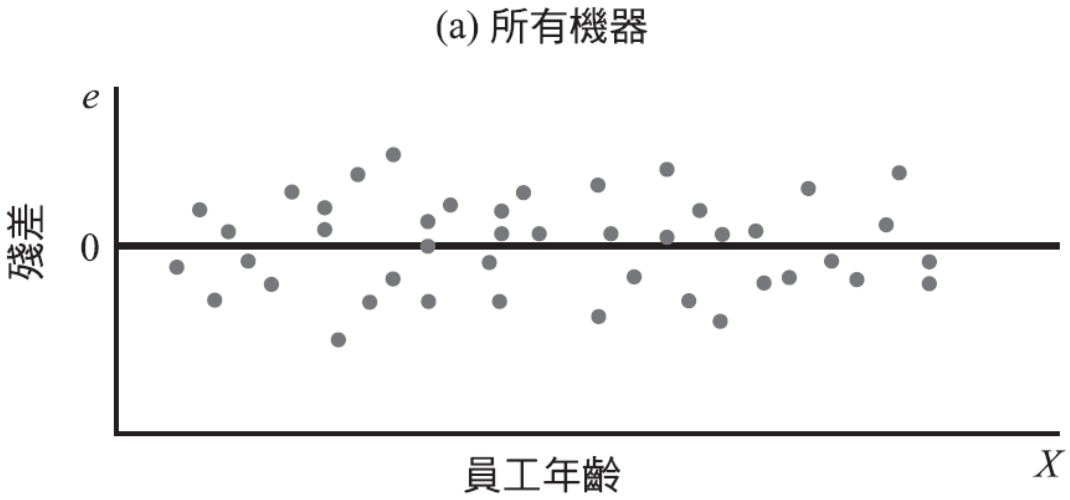


評估常態性的難處

- 對於模型是否偏離了常態性假設的分析，在許多方面確實是比其他型態的偏離要困難，
- 首先，除非樣本數夠大，否則隨機變異將造成研究機率分配相當程度的困擾；
- 其次，更嚴重的是因為其他方面偏離而影響了殘差的分配，例如有時會因為使用了不適當的迴歸函數，或是誤差項變異數不是常數，而使得誤差項的分配看起來不像是常態，
- **因此在關心誤差項是否為常態分配之前，最好先檢查其他型態的偏離。**

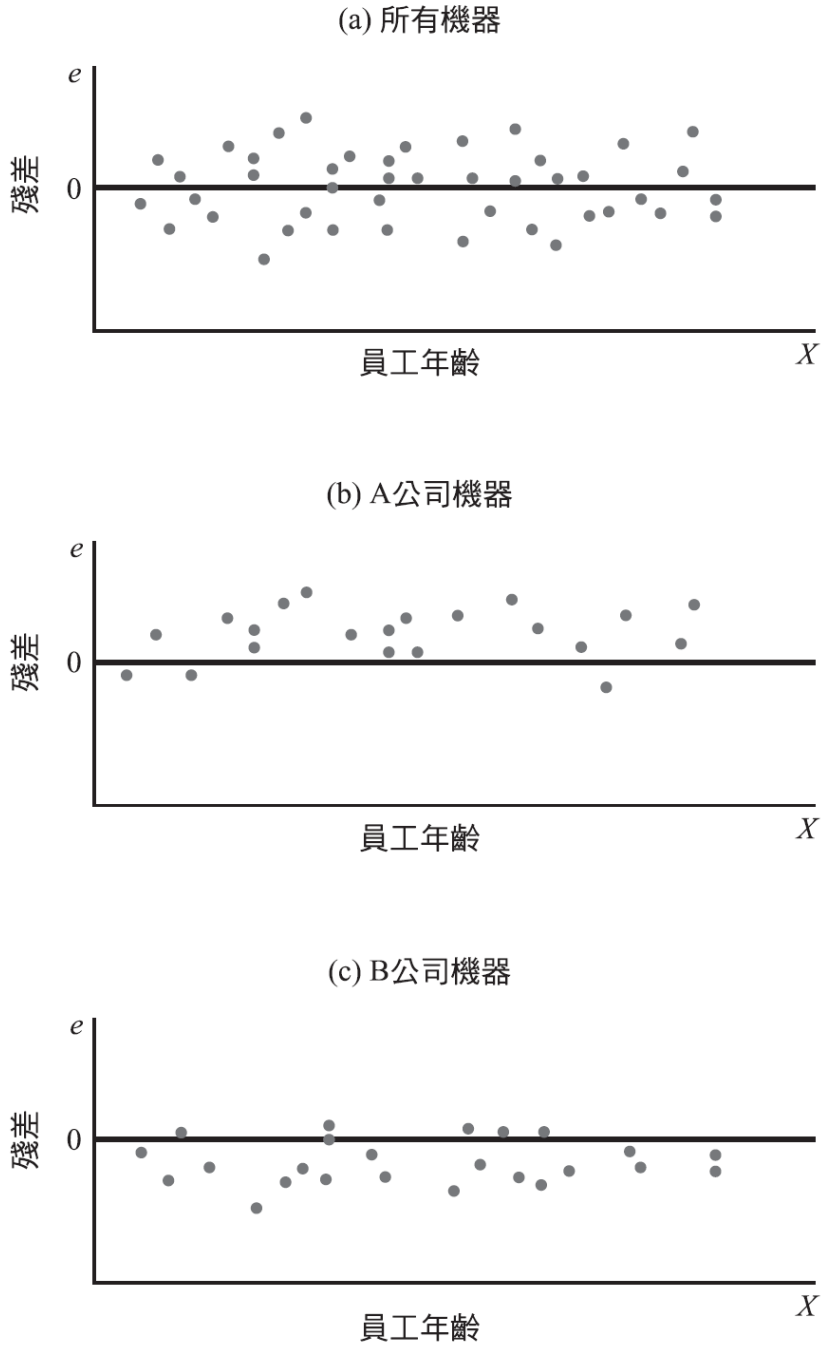
當重要地預測變數被忽略掉時

- 某些不在模型中的變數可能也會對於反應變數造成影響，在前面焊接的案例中，當懷疑時間變數也會是一個影響反應變數的因子時，可以透過對時間變數畫出殘差圖來分析。
- 這種附加的分析工作，其目的在於是否還有其他關鍵性變數，可以透過它們提供給模型額外的描述或預測能力。
- 一個有關於按件計酬的作業員其產量預測之研究，抽取一組員工研究其產量(Y)與年齡(X)兩者之間之關係，



- 圖3.10a 為X之殘差圖，圖中隱含了直線迴歸函數的適當性與常數誤差變異數的合理性

圖 3.10
對於可能忽略重要預測變數的殘差圖－生產力案例。



- 由於裝配之機器來自A與B兩家公司，所以不同公司的裝配機器可能有不同的產量，圖3.10b 與3.10c 分別針對A、B兩家公司的裝配機器劃出X之殘差圖。
- 此時圖3.10b 顯示了A公司的裝配機器其殘差多數為正值，而B公司的裝配機器其殘差多數為負值，因此似乎不同的公司所產生的機器對生產力有一些影響，如果模型中加入此一變數將可使產量的預測更為精確。

- 剛才所引的例子其預測變數是屬質變數(兩種機器)，但是所附加上屬量變數的殘差分析是相類似的，透過殘差圖形可以檢視此一預測變數是否會對殘差造成系統性之影響。

說明

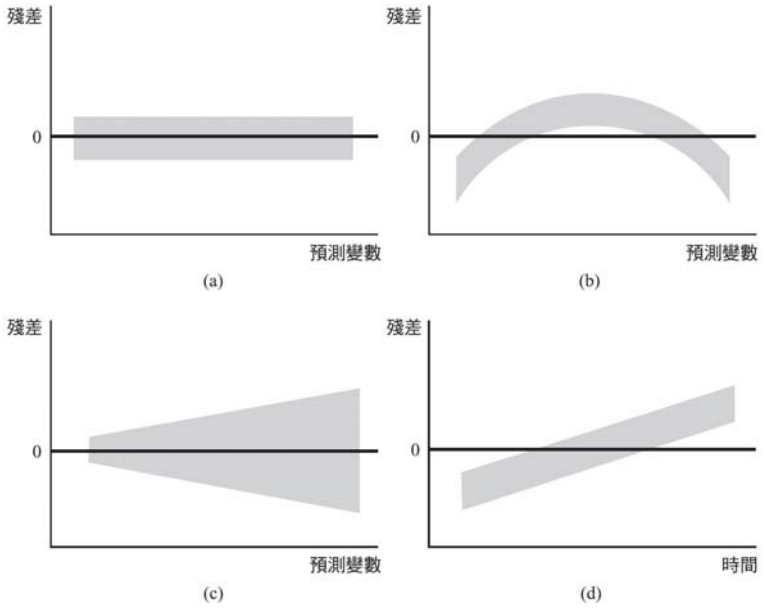
- 即使增加一個或數個預測變數可以有實質改善，也不能代表原來的模型就是錯的，
- 實際上任一反應變數的許多因子中，只有少數預測變數能被明顯地納入迴歸模型中，因此對於確認其他重要預測變數的殘差分析，其主要目的在於**檢驗模型之適切性**以及考慮**是否需要多增加預測變數**來改善結果。

最後幾點說明

1.關於模型的偏離，可能同時會有下列數種情形，逐項分析。例如直線迴歸可能配適不良，誤差項之變異可能也不是常數，在此情形下，圖3.4 的雛形經過複合仍然適用。

2.雖然殘差圖形的分析並非正式方法，卻足以檢驗多數情形下模型之適切性。

3.基本的殘差分析不僅適用於簡單線性迴歸模型，也適用於更複雜的迴歸或其他統計模型。



4.藉由殘差分析工具的診斷可以確認多數簡單線性模型的偏離，關於模型的非線性或重要預測變數未被納入，而導致參數估計或變異估計偏差的問題，在3.9節繼續討論

3.4 殘差檢定概述 Overview of Tests Involving Residuals

透過主觀的殘差圖形方法分析，似乎較為容易進行模型適切性的判斷。對多數的統計檢定而言，觀測值的獨立性常是必須的條件，有時因樣本數夠大時，殘差間微弱的相依性是可以被忽略的。

- 隨機性檢定

當將殘差依照時間順序排列後，可以透過**連續檢定**(runs test)來檢定殘差間是否不具有隨機性

- 常數變異數檢定

✓ 當殘差圖形中依照時間顯示變異數可能隨X或E{Y}出現系統性之變化，進行殘差絕對值與預測變數間的**等級相關檢定** (rank correlation test)

✓ 另外特別針對誤差項是否為常數之兩種檢定方法分別為Brown-Forsythe檢定與Breusch-Pagan檢定，將於3.6節中再討論之。

- 離群值檢定

當欲檢定某一樣本觀測值是否是離群值時，有一簡單的檢定方法，那就是先將 (*n*-1) 個觀測值配適一條迴歸直線，將所懷疑地離群值視爲新的樣本觀測值，然後計算如果有 *n* 個觀測值時，會出現一個剛才那種離群情形下的觀測樣本之機率有多大，如果機率小到一個程度，則可以說在此顯著水準下，除非此一觀測值已經有足夠的證據被視為離群值，否則，應該繼續保留此觀測值。

- 常態性檢定

✓配適度的檢定方法可以幫助檢定誤差項的常態性，例如**卡方檢定**(chi-square test)、Kolmogorov-Smirnov檢定或其他修訂版本Lilliefors檢定，均可於殘差分析中檢定誤差項之常態性

✓另外一個以常態機率圖為基礎的簡單檢定方法，將於下一節中進行說明。

3.5 常態性之相關檢定 Correlation Test for Normality

- 常態性之相關檢定

(a) 除了透過評估常態機率圖上之點是否近似直線外

(b) 也可以透過計算殘差 e_i 與其所對應常態期望值兩者間之相關係數，來進行較為正式的檢定，而相關係數值越高表示越符合常態分配。

1. 表 B6 (附錄B-16) 分別計算了在不同的樣本大小下，當誤差值確實是服從常態分配時，經過排序後之殘差與對應之常態期望值兩者間之相關係數的臨界值(百分位數)

2. 如果給訂一個顯著水準α，當觀測結果所計算出之相關係數至少達到了表中的對應值大小，則我們可以做出誤差項近似常態分配之結論。

表3.2的Toluca公司案例中，經過排序後之殘差與對應之常態期望值兩者間之相關係數為 .991 ，給定所控制之風險$\alpha = .05$ ，在表B.6中$n = 25$下之臨界值為 .959 ，顯然所觀測之相關係數大於此一水準，所以結論與前面相同，都是誤差項並未嚴重偏離常態分配。

3.6 常數誤差變異數之檢定
Tests for Constancy of Error Variance

Brown-Forsythe（BF）檢定(1)

- 由於檢定是利用殘差的變異性為基礎，所以當誤差之變異性越大，則殘差之變異應該也會越大，Brown-Forsythe檢定是利用 X 水準之大小將資料分成兩群，使得某一群的 X 水準較低，而另一群的 X 水準較高，如果誤差項變異數隨 X 而遞增或遞減，則此兩個群體之變異數將會明顯不同
- 為使檢定之穩健性更佳，檢定透過殘差對中位數的絕對離差為基礎來進行，依據 (A.67) –附錄A15的檢定統計量做出兩組樣本之 t 檢定，以資判斷是否兩群體間的平均絕對離差已經達到顯著差異，雖然殘差之絕對離差並非服從常態分配，但是在誤差項之變異相等條件下，當兩群體之樣本數夠大時，已經被證明出了t*近似 t 分配。

- *Brown-Forsythe*（BF）檢定 (2)

  下面以符號 e_{i1} 表示第一個群體的第 i 個殘差，符號 e_{i2} 表示第二個群體的第 i 個殘差，兩群體之樣本大小分別為 n_1 與 n_2，其中

$$n = n_1 + n_2 \tag{3.7}$$

- 同時利用 $\tilde{e}_1$ 與 $\tilde{e}_2$ 分別表示兩群體殘差之中位數，則兩群體殘差對於中位數之絕對離差 d_{i1} 與 d_{i2} 分別為：

$$d_{i1} = \left| e_{i1} - \tilde{e}_1 \right| \qquad d_{i2} = \left| e_{i2} - \tilde{e}_2 \right| \tag{3.8}$$

(A.67)的 t 檢定統計量為：

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.9}$$

其中$\bar{d}_1$與$\bar{d}_2$分別是 d_{i1} 與 d_{i2}之樣本平均數，而 (A.63) 的混合變異數 s^2 為：

混合變異數 s^2 為：

$$s^2 = \frac{\sum\left(d_{i1} - \bar{d}_1\right)^2 + \sum\left(d_{i2} - \bar{d}_2\right)^2}{n-2} \qquad (3.9a)$$

使用符號 t^*_{BF} 來表示 *Brown-Forsythe* 的檢定統計量。

- 如果誤差項滿足常數變異數之假設條件，且樣本數 n_1 與 n_2 並不會太小，則 t^*_{BF} 將會近似自由度 $(n\text{-}2)$ 之 t 分配，所以當 t^*_{BF} 之絕對值很大時，表示誤差項變異數不會是常數。

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

例題

利用 Brown-Forsythe 檢定來看看 Toluca 公司的案例中，誤差項變異數是否會隨X之水準不同而改變，由於案例中X之水準分散情形相當均勻（見圖3.1a），我們可以將25個樣本資料分割成兩群，使得這兩群樣本之全距接近，於是第一群將有 13 個批次，批量範圍從 20 到 70 ，第二群有 12 個批次，批量範圍從 80 到 120 ，表 3.3 列出各群之部分資料，在第一個欄位與第二個欄位重複了表1.2的批量與殘差。從表3.3中可以看出第一群殘差之中位數為$\tilde{e}_1 = -19.88$，第二群殘差之中位數為$\tilde{e}_2 = -2.68$，而第三個欄位是殘差對於該群之中位數的絕對離差，例如：

表 3.3
誤差項為常數變異數之 *Brown-Forsythe* 檢定計算－Toluca 公司案例。

第一群					
		(1)	(2)	(3)	(4)
			殘差		
i	批次	批量	e_{i1}	d_{i1}	$(d_{i1} - \bar{d}_1)^2$
1	14	20	−20.77	.89	1,929.41
2	2	30	−48.47	28.59	263.25
...	...	...	...	...	...
12	12	70	−60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
	總合			582.60	12,566.6

$\tilde{e}_1 = -19.88 \qquad \bar{d}_1 = 44.815$

表 3.3
誤差項為常數變異數之 *Brown-Forsythe* 檢定計算－Toluca 公司案例。

第二群					
		(1)	(2)	(3)	(4)
			殘差		
i	批次	批量	e_{i2}	d_{i2}	$(d_{i2} - \bar{d}_2)^2$
1	1	80	51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
...	...	...	...	...	...
11	20	110	−34.09	31.41	8.76
12	7	120	55.21	57.89	866.71
	總合			341.40	9,610.2
			$\tilde{e}_2 = -2.68$	$\bar{d}_2 = 28.450$	

$$d_{11} = |e_{11} - \tilde{e}_1| = |-20.77 - (-19.88)| = .89$$

$$d_{12} = |e_{12} - \tilde{e}_2| = |51.02 - (-2.68)| = 53.70$$

絕對離差之平均數為：

$$\bar{d}_1 = \frac{582.60}{13} = 44.815 \qquad \bar{d}_2 = \frac{341.40}{12} = 28.450$$

最後，第四個欄位是d_{i1}與d_{i2}對於該群之平均數取離差後之平方，例如：

$$\left(d_{11} - \bar{d}_1\right)^2 = (.89 - 44.815)^2 = 1{,}929.41$$

$$\left(d_{12} - \bar{d}_2\right)^2 = (53.70 - 28.450)^2 = 637.56$$

於是我們可以算出(3.9)之統計量：

表 3.3
誤差項為常數變異數之 *Brown-Forsythe* 檢定計算－Toluca 公司案例。

第一群

		(1)	(2)	(3)	(4)
			殘差		
i	批次	批量	e_{i1}	d_{i1}	$(d_{i1} - \bar{d}_1)^2$
1	14	20	−20.77	.89	1,929.41
2	2	30	−48.47	28.59	263.25
...	...	...	...	...	...
12	12	70	−60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
	總合			582.60	12,566.6

$\tilde{e}_1 = -19.88 \qquad \bar{d}_1 = 44.815$

第二群

		(1)	(2)	(3)	(4)
			殘差		
i	批次	批量	e_{i2}	d_{i2}	$(d_{i2} - \bar{d}_2)^2$
1	1	80	51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
...	...	...	...	...	...
11	20	110	−34.09	31.41	8.76
12	7	120	55.21	57.89	866.71
	總合			341.40	9,610.2

$\tilde{e}_2 = -2.68 \qquad \bar{d}_2 = 28.450$

$$s^2 = \frac{12,566.6 + 9,610.2}{25 - 2} = 964.21$$

$$s = 31.05$$

$$t^*_{BF} = \frac{44.815 - 28.450}{31.05\sqrt{\frac{1}{13} + \frac{1}{12}}} = 1.32$$

當風險控制在$\alpha = .05$時，$t(.975;23) = 2.069$，此時決策依據為：

若$\left|t^*_{BF}\right| \le 2.069$, 則結論為誤差項變異數是常數

若$\left|t^*_{BF}\right| > 2.069$, 則結論為誤差項變異數不是常數

因$\left|t^*_{BF}\right| = 1.32 \le 2.069$，於是我們認為誤差項變異數並沒有嚴重偏離常數之假設，亦即沒有明顯地隨X之水準不同而改變，此一檢定之雙尾機率值為 .20。

- ***Breusch-Pagan（BP）檢定***
- 此一檢定適用於大樣本之情形
- 假設條件為誤差項 ε_i 之間彼此獨立且服從變異數 σ_i^2 之常態分配，而 σ_i^2 與 X_i 之關係如下：

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i \qquad (3.10)$$

- 式子(3.10)隱含了 σ_i^2 會隨著 X_i 之水準遞增或遞減是取決於 γ_1 之正負符號而定，固定的誤差變異數時是對應到 $\gamma_1 = 0$ 。
- $H_0 : \gamma_1 = 0$ 對 $H_a : \gamma_1 \neq 0$ 之檢定過程是透過將殘差平方後得到的 σ_i^2 對 X_i 配適一般之迴歸模型，以 *SSR* 表示所配適之迴歸平方和，而檢定之統計量 X_{BP}^2 為：
- $$X_{BP}^2 = \frac{SSR^*}{2} \div \left(\frac{SSE}{n} \right)^2 \qquad (3.11)$$

上式中 SSR* 為配適之迴歸平方和，但是 SSE 是針對 Y 對 X 所配適之迴歸模型中的誤差平方和

- 如果 H_o：$\gamma_1 = 0$ 成立，且 n 也夠大，則統計量 X^2_{BP} 將近似於自由度 1 之卡方分配，而大的 X^2_{BP} 值將傾向於接受 H_α：$\gamma_1 \neq 0$，亦即誤差項不是常數變異數。

利用 Breusch-Pagan 檢定來看看 Toluca 公司的案例，我們將表 1.2 中的第五個欄位關於殘差平方的部分，對 X_i 配適一般之迴歸模型，可以得到 $SSR^* = 7{,}896{,}128$，而在圖 2.2 中可以得到 $SSE = 54{,}825$，因此可以計算出(3.11)式中的檢定統計量 X^2_{BP} 為：

表 B.3 (附錄B-6)

$$X^2_{BP} = \frac{7{,}896{,}128}{2} \div \left(\frac{54{,}825}{25}\right)^2 = .821$$

當風險控制在 $\alpha = .05$ 時，$\chi^2(.95;1) = 3.84$，因 $X^2_{BP} = .821 \leq 3.84$，所以我們的結論無法拒絕 H_0 之虛無假設，可以將誤差項視為常數變異數，此一檢定之機率值為 .64，相當符合常數變異數之條件要求。

3.7 配適不佳之 *F* 檢定

F Test for Lack of Fit

- 討論有關於特定型態的迴歸函數是否可以適合配適資料之正式檢定，在此以直線迴歸函數為例探討說明。
- 假設
- 配適不良之檢定之基本假設有三：(1) 獨立，(2) 服從常態分配，(3) Y 之分配變異數相同。
- 配適不良之檢定需要在一個或多個 X 水準下**重複多個觀測值**
- 對於非實驗性之資料，只能是偶然發生的情形，而對於實驗性之資料，研究者可以透過設計實驗來收集重覆之觀測值，
- 對於相同水準下之預測變數進行重複之實驗，稱之為**重複實驗**(replication)，而所得到之觀測值為**重複觀測值**(replicate)。

例題

研究對象為某家商銀十二個分散在各地之分行，研究調查分行的支票存款戶，存戶在開立新帳戶時可以獲贈禮物，但有最低的初次存款金額限制，同時規定的最低初次存款金額與禮品之價值成正比例，在研究之實驗中採用六種不同最低初次存款金額之規定，以及成正向比例價值之禮品，以便研究與新開戶數之間所存在之關係，並且對於每一種規定的存款金額均進行兩家分行之實驗，不過在實驗進行的過程中，由於其中一家分行不幸失火，而必須被剔除在觀測的樣本之外，表 3.4a 的資料是實驗進行後之結果，其中 X 為最低存款金額，Y 則是新開戶數。

表 3.4
資料與變異數分析表－銀行案例。

(a) 資料

分行	最低存款金額（元）	新開戶數	分行	最低存款金額（元）	新開戶數
i	X_i	Y_i	i	X_i	Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

透過一般的直線迴歸模型配適函數，得到如下之結果：

$$\hat{Y} = 50.72251 + .48670X$$

表 3.4b 為該案例之變異數分析表，圖 3.11 則顯示散佈圖以及所配適之迴歸直線，從該圖中可以發現直線迴歸之模型並不適合本案例中之資料，如果要進行更為正式的檢定，則可以應用本書 2.8 節之一般線性檢定法。

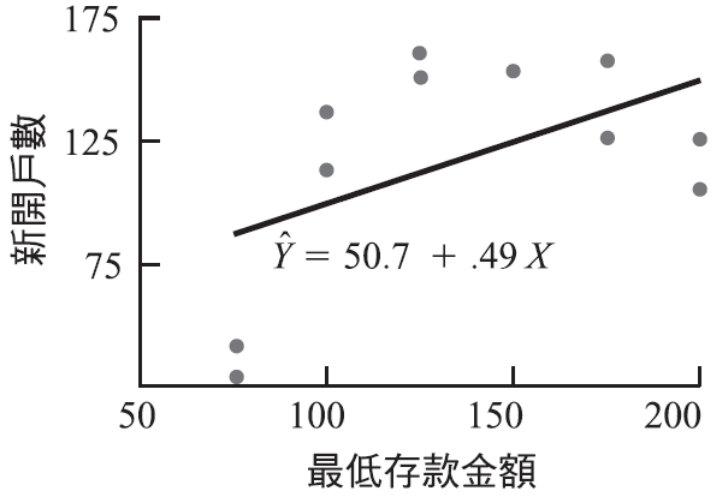


圖 3.11
散佈圖與所配適之迴歸直線－銀行案例。

(b) 變異數分析表

變異來源	*SS*	*df*	*MS*
迴歸	5,141.3	1	5,141.3
誤差	14,741.6	9	1,638.0
總合	19,882.9	10	

表 3.4
資料與變異數分析表－銀行案例。

符號

- 由於同一水準之 X 出現瞭重複實驗之觀測樣本資料，所以有必要另外定義符號以配合模型的表示。表3.5 的資料與表3.4a 的資料一樣，只是表3.5 的資料是依據重複實驗之順序與最低存款金額而進行的排列，不論實驗是否進行重複，對於不同的 X 水準用 $X_1 \ldots X_c$ 表示，

表 3.5
資料依重複實驗之順序與最低存款金額排列－銀行案例。

	最低存款金額（元）					
重複實驗之順序	$j = 1$ $X_1 = 75$	$j = 2$ $X_2 = 100$	$j = 3$ $X_3 = 125$	$j = 4$ $X_4 = 150$	$j = 5$ $X_5 = 175$	$j = 6$ $X_6 = 200$
$i = 1$	28	112	160	152	156	124
$i = 2$	42	136	150		124	104
平均 $\bar{Y}_j$	35	124	155	152	140	114

重複實驗之順序	最低存款金額（元）					
	$j = 1$ $X_1 = 75$	$j = 2$ $X_2 = 100$	$j = 3$ $X_3 = 125$	$j = 4$ $X_4 = 150$	$j = 5$ $X_5 = 175$	$j = 6$ $X_6 = 200$
$i = 1$	28	112	160	152	156	124
$i = 2$	42	136	150		124	104
平均 $\bar{Y}_j$	35	124	155	152	140	114

- 在本案例中因為共有六種不同最低初次存款金額之規定，所以c＝6，其中五個 X 水準分別有兩個觀測樣本，而 $X_4 = 150$ 時僅有一個觀測值樣本，令X的第 j 個水準重複觀測 n_j 個樣本，在本例中 $n_1 = n_2 = n_3 = n_5 = n_6 = 2$，而 $n_4 = 1$，所以總觀測樣本數為：

$$n = \sum_{j=1}^{c} n_{\cdot} \qquad (3.12)$$

- 反應變數 Y 在 X 的第 j 個水準下所進行的第 i 次實驗值用符號 $\mathbf{Y_{ij}}$ 表示，其中i＝1,..,n_j，j＝1,…,c，在本案例中(3.5)，Y_{11}=28，Y_{21}=42， Y_{12}=112。當X=X_j水準下觀測值之平均數以 $\bar{Y}_j$ 表示，例如 $\bar{Y}_1 = (28+42)/2 = 35$，而 $\bar{Y}_4 = 152/1 = 152$。

1. 完全模型 （2. 減縮模型， 3. 檢定統計量）

- 一般線性檢定法首先須指定**完全模型**如下，除了線性迴歸關係之假設外，其餘假設均相同於簡單線性迴歸模型

(2.1)： $Y_{ij} = \mu_j + \varepsilon_{ij}$ 完全模型 (3.13)

其中 μ_j為參數，$j = 1,\ldots,c$；

ε_{ij}為互相獨立並服從 $N(0,\sigma^2)$ 之隨機變數。

- 目前的檢定問題中，不再假設直線部分的線性關係。完全模型與與(2.1)簡單線性迴歸相同，由於誤差項 ε_{ij} 之期望值為零，所以： $E\{Y_{ij}\} = \mu_j$ (3.14)

而參數 μ_j 表示 $X=X_j$ 時觀測樣本的平均反應。

- 完全模型(3.13)說明了一個反應值Y都是由兩個部分所組成，亦即$X=X_j$時，觀測樣本的平均反應以及所伴隨的隨機誤差項

- 不同於簡單線性迴歸模型(2.1)的地方，是(3.13)的完全模型對於平均反應 μ_j 沒有做出任何假設限制，而簡單線性迴歸模型(2.1)的平均反應則與X有線性關係(亦即 $E\{Y\} = \beta_0 + \beta_1 X$)。
- 對資料配適的完全模型需要求出參數 μ_j 的最小平方估計量或是最大概似估計值，這些估計量就是樣本平均數 $\bar{Y}_j$：

$$\hat{\mu}_j = \bar{Y}_j \tag{3.15}$$

- 因此對觀測值 Y_{ij} 所估計值正是 $\bar{Y}_j$，而**完全模型的誤差平方和**為

$$SSE(F) = \sum_j \sum_i \left(Y_{ij} - \bar{Y}_j\right)^2 = SSPE \tag{3.16}$$

- 在此有關配適不良的檢定中，完全模型(3.13)的誤差平方和經常被稱為 **純誤差平方和** (pure error sum of squares)，用 ***SSPE*** 表示。

- *SSPE* 是根據各個 X 水準的離差平方和加總而成的，為：

$$SSE(F)=\sum_j\sum_i\left(Y_{ij}-\bar{Y}_j\right)^2=SSPE \qquad \sum_i\left(Y_{ij}-\bar{Y}_j\right)^2 \qquad (3.17)$$

- 然後將所有X水準的離差平方和加總，如上述銀行的案例：

$$SSPE=(28\text{-}35)^2+(42\text{-}35)^2+(112\text{-}124)^2+(136\text{-}124)^2+$$
$$(160\text{-}155)^2+(150\text{-}155)^2+(152\text{-}152)^2+(156\text{-}140)^2$$
$$+(120\text{-}140)^2+(124\text{-}114)^2+(140\text{-}114)^2=1{,}148$$

	最低存款金額（元）					
重複實驗之順序	$j=1$ $X_1=75$	$j=2$ $X_2=100$	$j=3$ $X_3=125$	$j=4$ $X_4=150$	$j=5$ $X_5=175$	$j=6$ $X_6=200$
$i=1$	28	112	160	152	156	124
$i=2$	42	136	150		124	104
平均 $\bar{Y}_j$	35	124	155	152	140	114

- 值得注意的是，沒有進行重複觀測值實驗的X水準對於 *SSPE* 是沒有貢獻的，因為 $\bar{Y}_j = Y_{ij}$ 例如上述銀行的答案案例中，當 j＝4時，$(152\text{-}152)^2=0$。

- 在(3.17)中每一個 X = X_j 下 n_j 個觀測值之離差平方和 $\sum_i (Y_{ij} - \bar{Y}_j)^2$ 其自由度 n_j-1，因為 *SSPE* 是對於所有X水準的離差平方和加總，所以其自由度為： $df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c$ (3.18)

  在銀行案例中，自由度 df_F = 11-6 = 5。對於沒有進行重複觀測實驗的X水準而言，其自由度為 n_j-1=1-1=0，所以對於全部的自由度而言是沒有貢獻的。

2. 精簡模型

E{Yij }= μj

- 其次考慮在H_o之下的精簡模型，對於檢定直線線性迴歸關係適當與否的問題，可以有如下之檢定：

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X$$

$$H_a : E\{Y\} \neq \beta_0 + \beta_1 X$$

E{Y_{ij} }= μ_j

- 亦即 H_o 主張完全模型 $Y_{ij} = \mu_j + \varepsilon_{ij}$ (3.13)下的 μ_j 與 X_j 有線性關係：

$$\mu_j = \beta_o + \beta_1 X$$

- 所以在 H_o 下的精簡模型為：

$$Y_{ij} = E\{Y\} = \beta_o + \beta_1 X_j + \varepsilon_{ij} \qquad \text{精簡模型} \qquad (3.20)$$

- 此處的精簡模型其實就是簡單線性迴歸模型(2.1)，由於在迴歸模型(2.1)中觀測值 Y_{ij} 之估計期望值為配適 $\hat{Y}_{ij}$：

$$\hat{Y}_{ij} = b_0 + b_1 X_j \qquad (3.21)$$

- 因此，精簡模型下的誤差平方和 SSE(R) 即是以前所提的誤差平方和 ***SSE***：

$$SSE(R) = \sum\sum \left[Y_{ij} - \left(b_0 + b_1 X_j \right) \right]^2 = \sum\sum \left(Y_{ij} - \hat{Y}_{ij} \right)^2 = SSE \qquad (3.22)$$

同時 *SSE(R)* 之自由度為： $df_R = n\text{-}2$

- 上述銀行案例中從表3.4b可以得到： $SSE(R) = SSE = 14.741.6$ ， $\boldsymbol{df_R} = 9$

3. 檢定統計量

- 在(2.70)中一般線性檢定之統計量為：

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

在此修改為：

$$F^* = \frac{SSE - SSPE}{(n-2)-(n-c)} \div \frac{SSPE}{n-c} \qquad (3.23)$$

- 而這兩個誤差平方和相減後所得之為配適不良平方和(***lack of fit sum of squares***)，以符號 *SSLF* 表示：

$$SSLF = SSE - SSPE \qquad (3.24)$$

所以檢定統計量可以表示成：

$$F^* = \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE} \qquad (3.25)$$

- 其中，*MSLF*表示配適不良均方(*lack of fit mean square*)，*MSPE*表示純誤差均方(*pure error mean square*)。

- 在一般線性檢定中，F^*值越大結論越傾向H_a，而決策規則 (2.71) 則成為

$$\text{若 } F^* \le F(1-\alpha;c-2,n-c), \text{則結論為 } H_0$$
$$\text{若 } F^* > F(1-\alpha;c-2,n-c), \text{則結論為 } H_a \quad (3.26)$$

- 上述銀行的案例中，檢定統計量經計算後可以得到：

  $SSPE$ = 1,148.0　　n-c=11-6=5

  SSE = 14,741.6

  $SSLF$ = 14,741-1,148.0=13,536　　c-2=6-2=4

  F^* = (13593.6 / 4) / (1148.0 / 5) = 33984 / 229.6 =14.80

- 當顯著水準 α =.01時，$F(0.99;4,5)$ = 11.4＜14.80 = F^*，因此結論為 H_a：迴歸函數不是線性，此一結果與圖3.11所給的看法相同，在本案例中之 P-值為 0.006。

$$H_0: E\{Y\} = \beta_0 + \beta_1 X$$
$$H_a: E\{Y\} \neq \beta_0 + \beta_1 X$$

變異數分析表

- 在(3.24)中配適不良平方和*SSLF*的定義可以理解為將誤差平方和拆解成兩部分：

$$SSE=SSPE+SSLF \tag{3.27}$$

此一分解原理源自於下列恆等式：

$$\underbrace{Y_{ij}-\hat{Y}_{ij}}_{\text{誤差離差}}=\underbrace{Y_{ij}-\overline{Y}_{j}}_{\text{純誤差離差}}+\underbrace{\overline{Y}_{j}-\hat{Y}_{ij}}_{\text{配適不良離差}} \tag{3.28}$$

- 上面的恆等式表示誤差平方和是由純誤差成分與配適不良之成分所構成，在圖3.12中表示了銀行的案例裡，觀測樣本$X_3=125$, $Y_{13}=160$的誤差量之分解情形。

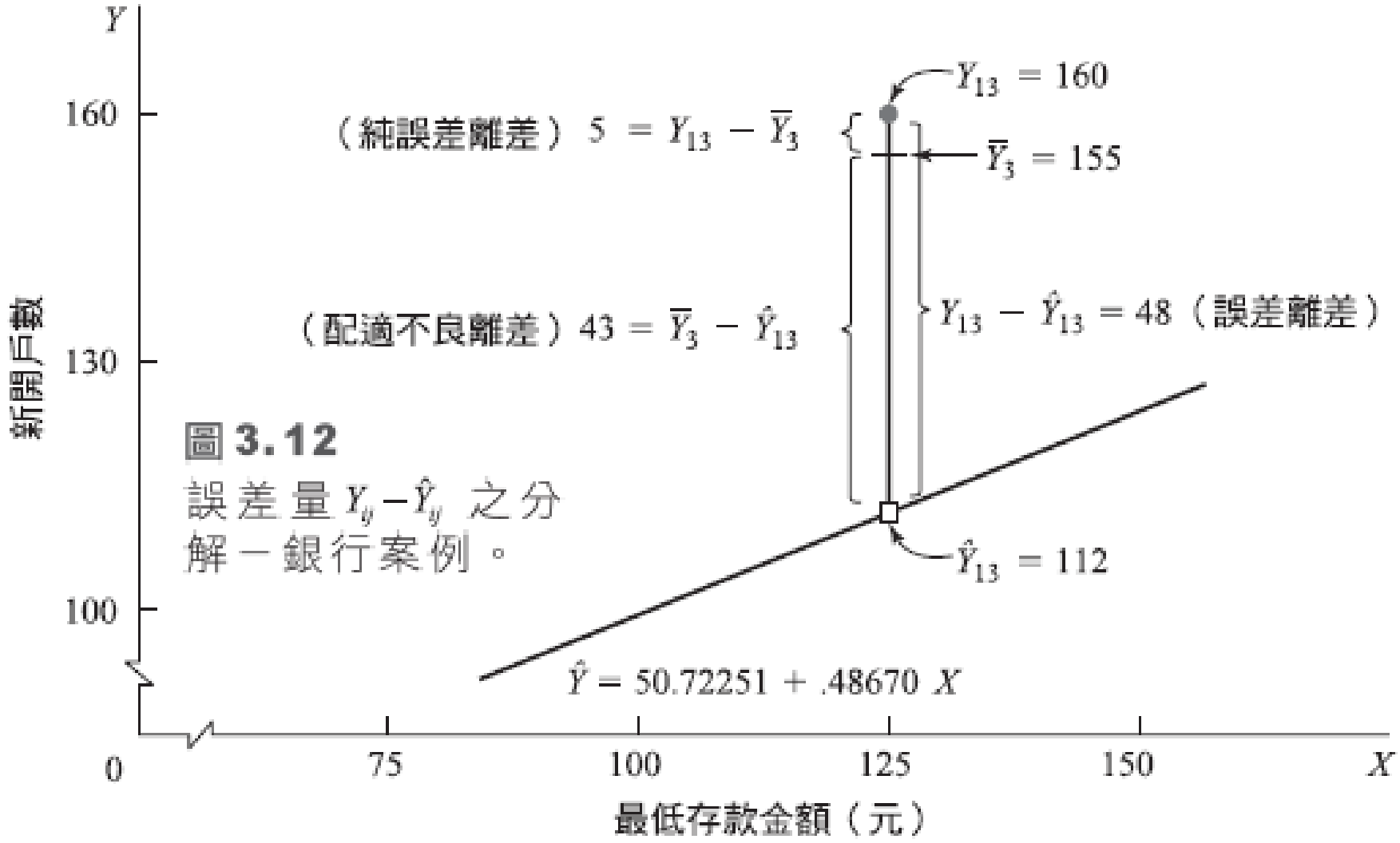


圖 3.12
誤差量 $Y_{ij} - \hat{Y}_{ij}$ 之分解－銀行案例。

圖3.12中表示觀測樣本X_3=125, Y_{13}=160的誤差量之分解情形

$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\text{誤差離差}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\text{純誤差離差}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\text{配適不良離差}}$$

- 將(3.28)取平方後對所有觀測之樣本加總，由於交叉乘積項之總合為零，所以(3.27)可以解釋成：

$$\sum\sum\left(Y_{ij}-\hat{Y}_{ij}\right)^2=\sum\sum\left(Y_{ij}-\bar{Y}_j\right)^2+\sum\sum\left(\bar{Y}_j-\hat{Y}_{ij}\right)^2 \qquad (3.29)$$

$$SSE \quad = \quad SSPE \quad + \quad SSLF$$

- 因此可以直接定義**配適不良之平方和**如下：

$$SSLF=\sum\sum\left(\bar{Y}_j-\hat{Y}_{ij}\right)^2 \qquad (3.30)$$

- 在水準 X_j 下之**所有**觀測值 Y_{ij} 都有相同的配適值，所以(3.30) 又可以表示成：

$$SSLF=\sum_j n_j\left(\bar{Y}_j-\hat{Y}_j\right)^2 \qquad (3.30a)$$

(3.30a)式清楚地指出了爲何 *SSLF* 可以代表配適不良之程度，如果線性之迴歸函數適當，則 $\bar{Y}_j$ 將會接近所估計的線性迴歸函數值 $\hat{Y}_j$ ，如此一來，*SSLF* 將會很小；反之如銀行的案例，*SSLF*將會偏高。

- (3.30a)式同時也告訴了我們為何 *SSLF* 具有 c-2 個自由度，由於在平方和中有c個平均數 $\overline{Y}_j$，而其中有兩各自由度因為估計了 β_0 與 β_1 而損失掉了。
- 對於 *SSE* 之分解也可以透過變異數分析(ANOVA)表來顯示，表3.6a 是一般常見的 ANOVA 表，包含上述 *SSE* 之分解，而表3.6b 則是銀行案例中的 ANOVA 表。

(a) 變異數分析表

變異來源	平方和	自由度	均方
迴歸	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
誤差	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$
配適不良	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c - 2$	$MSLF = \frac{SSLF}{c-2}$
純誤差	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n - c$	$MSPE = \frac{SSPE}{n-c}$
總合	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n - 1$	

(b) 銀行案例之變異數分析表

變異來源	平方和	自由度	均方
迴歸	5,141.3	1	5,141.3
誤差	14,741.6	9	1,638.0
配適不良	13,593.6	4	3,398.4
純誤差	1,148.0	5	229.6
總合	19,882.9	10	

表 3.6
檢定簡單線性迴歸函數是否配適不良之變異數分析表以及銀行案例之變異數分析表。

- 說明

1.在銀行的案例中，要進行配適不良之檢定，其實並不需要X的所有水準均具有重覆之觀測樣本，不過至少需要一個X的水準有重覆之觀測樣本。

2.對於檢定簡單線性迴歸函數是否為線性之問題中，可以證明出均方 *MSLF* 與 *MSPE* 之期望值：

$$E\{MSPE\} = \sigma^2 \tag{3.31}$$

$$E\{MSLF\} = \sigma^2 + \frac{\sum n_j \left[\mu_j - (\beta_0 + \beta_1 X_j)\right]^2}{c-2} \tag{3.32}$$

「純誤差」名詞之由來是因為不論真實的迴歸函數如何，*MSPE* 永遠都是誤差項變異數σ^2的不偏估計量，當迴歸函數真是直線時，因為$\mu_j = \beta_0 + \beta_i X_j$，所以(3.32)的第二項將為零，於是*MSLF*的期望值也會是σ^2，不過若$\mu_j \neq \beta_0 + \beta_i X_j$，則$E\{MSLF\}$將大於$\sigma^2$，因此$F^*$值越接近1表示模型越接近線性函數，而$F^*$值越大表示模型越不可能是線性函數。

4.假設在對模型的配適度進行分析之前，先配適了線性迴歸模型並進行β_1=0之檢定，在銀行的案例中，根據檢定統計量(2.60)：　$F^* = (MSR/MSE) = 5141.3 / 1638.0 = 3.13$

當顯著水準$\alpha = 0.10$時，$F(0.90; 1, 9) = 3.36$，因此結論並未強烈到足以拒絕H_o，亦即可以認定β_1=0的事實，或者說是最低存款金額與新開戶數，兩者之間不存在明顯的線性關係，不過，這並不代表兩者之間不存在關係，在本案例此二變數可能存在有關係，只不過迴歸函數之關係不是線性罷了。這種情形說明了在進行任何結論之前，先檢驗分析適當的模型之重要性。

5. 一般而言，*SSLF*具有$c - p$個自由度，其中p為迴歸函數中的參數各數，在簡單線性函數之檢定中，因為迴歸函數有兩個參數，分別為β_0與β_1，所以$p = 2$。

6. 在(3.19)中的H_α包含了直線之外的所有迴歸函數，例如二次函數或是對數函數，當做出H_o之結論後，可以透過對於殘差之研究來找出合適的迴歸函數。

8. 關於同一水準X下的重複觀測，是指誤差項須具備有獨立的實驗結果。例如，當我們想研究合金之硬度(Y)與碳含量(X)兩者間之關係，在模型中的誤差項必須包含有分析人員測量硬度時之誤差，以及隨著不同樣本，無法控制且會影響硬度的製程變異，假設分析人員針對同一樣本測量兩次硬度，因為製程變異是固定的，所以並不能算是重複觀測；如果要進行重複觀測，則分析人員應針對相同的碳含量(X)，但是不同的兩個樣本來測量硬度，這樣誤差項的作用才具備有隨機變化的意義。

3.8 矯正測量概述 Overview of Remedial Measures

- 當簡單線性迴歸模型(2.1)不適合所要分析的資料時，可能選擇有二：

1. 放棄線性迴歸模型(2.1)，並找出更為適合之模型。
2. 對資料進行轉換工作，希望轉換後之結果，迴歸模型(2.1)能適合所要分析的資料。

- 上述兩種選擇各有其優劣，第一種選擇或許可以找到對問題之描述更為清楚之模型，但是因模型的複雜度增加，可能對於參數之估計程序將更為困難，第二種選擇如果真能將資料轉換成適合迴歸模型(2.1)之結果，則估計方法將較為簡單，所牽涉之參數也較少，對於小樣本之研究是一大優點，不過有時變數間的轉換，可能當中便隱含了變數間的基本關係。

■ 非線性迴歸函數

- 若迴歸函數不是直線，直線的做法之一是改變迴歸模型(2.1)的函數形式，例如考慮二次迴歸函數：

$$E\{Y\} = \beta_o + \beta_1 X + \beta_2 X^2$$

- 或指數迴歸函數：

$$E\{Y\} = \beta_o \beta_1^X$$

- 變數變換法是將一個非線性函數透過某種轉換方式，使其線性化或接近線性函數，關於這點，我們將於3.9節中討論；如果函數的性質未知，可以應用探索性分析(exploratory analysis)，這種不需特別「指定函數型態」的分析方法-於 3.10 節討論

■ 非常數誤差變異數

- 如果具備有系統性地型態，可以直接修改函數模型以配合使用，並利用加權最小平方法(weighted least squares)得到參數的估計量。有時**變數轉換**對於穩定變異數也有幫助，我們將於 3.9 節中討論之。

■ 非獨立之誤差項

- 當誤差項之間彼此具有關聯性時，可以使用第12章第相關誤差項模型來修正此一問題，有一種簡單的做法是將原始資料進行一階差分之後，再進行分析工作，關於這點，一樣會在第12章中討論之。

■非常態之誤差項

- 誤差項不服從常態分配或誤差變異數不是常數，這兩種情形經常同時發生，不過幸運的是，透過**穩定**變異數之轉換，經常也可以使誤差項近似於常態分配，因此我們可以優先考慮對資料進行穩定誤差變異數的變數轉換，然後再探討誤差項是否仍然嚴重違背常態分配，關於這點，我們將於 3.9 節中討論。

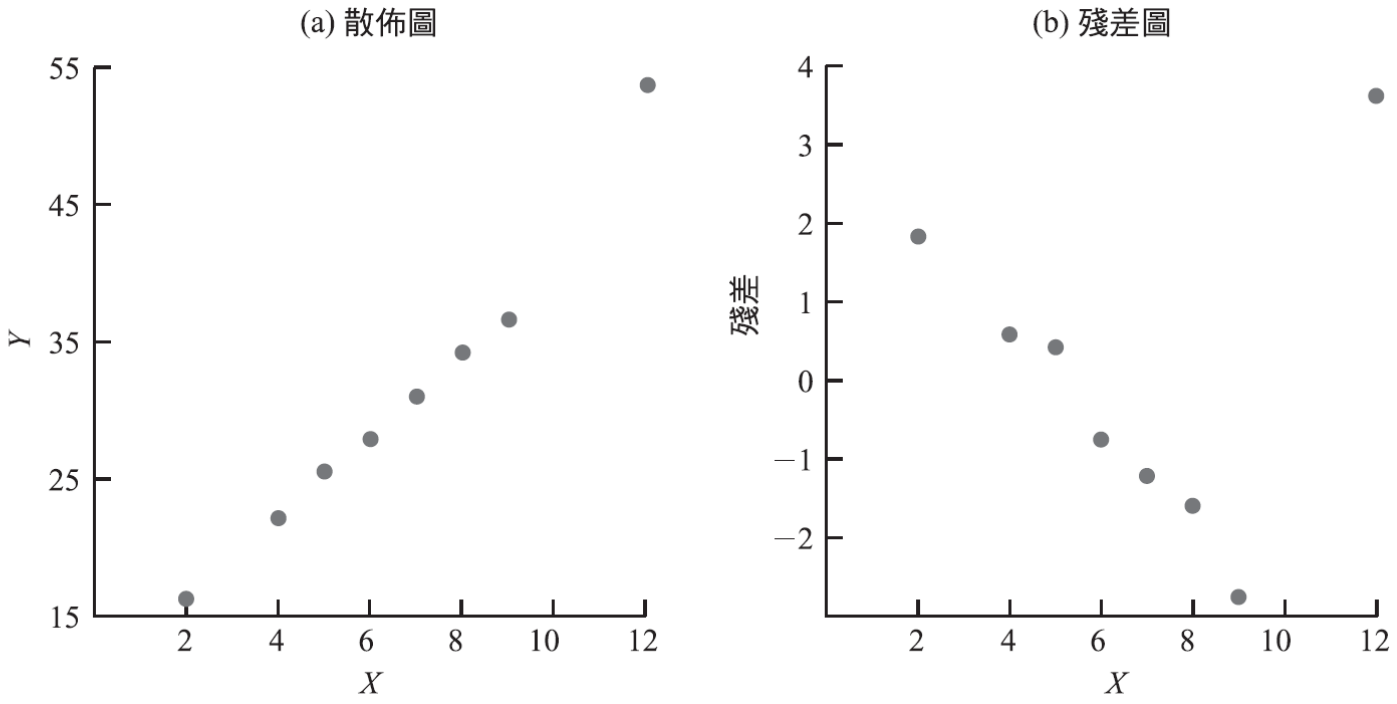
■當重要的預測變數被忽略掉時

- 如果殘差分析後顯示出有某些重要的預測變數未被納入模型中，此時可能需要修正模型以解決此一問題，在第 6 章以後，將會進行兩個以上預設變數的複迴歸分析(multiple regression analysis)。

■ 離群值

- 當出現如圖 3.7a 之離群值時，透過最小平方法或最大概似估計量 (1.10) 來估計迴歸模型 (2.1)，將可能嚴重扭曲迴歸函數，如果確定離群觀測值並非來自錯誤的記錄，則此時不該將其剔除在資料外，此時可以另外採用一些估計程序來針對離群觀測值進行分析，第10章將討論此一穩健的估計程序。

圖 3.7
迴歸直線配適因一個離群值而造成殘差圖異常的不良影響。



3.9 資料轉換 Transformations

透過簡單的轉換程序，常能使得簡單線性迴歸模型適合轉換後的資料。

僅存在非線性關係之轉換

- 首先考慮當誤差項接近常態分配，且其變異數近似常數，不過此時迴歸函數卻表現出非線性的情形，可以嘗試先對 X 進行轉換，因為如果一開始便對 Y 進行轉換，例如 $\acute{Y}=(Y)^{1/2}$ ，則結果將對於誤差項的分配有嚴重之影響，可能不再是常態分配，或變異數偏離常數

- 圖 3.13 列出了一些具備有常數變異數但卻爲非線性迴歸關係之雛型，並提出了在不同的情形時，在不影響 Y 之分配前提下，對 X 進行轉換後，將有助於迴歸關係線性化的一些轉換方式，有時候必須要嘗試多種不同的轉換，並分別畫出散佈圖及殘差圖來檢視分析，方能決定最有效之轉換方式。

圖 3.13
常數誤差變異數下的非線性迴歸型態之雛型以及對 X 進行的簡單轉換方式。

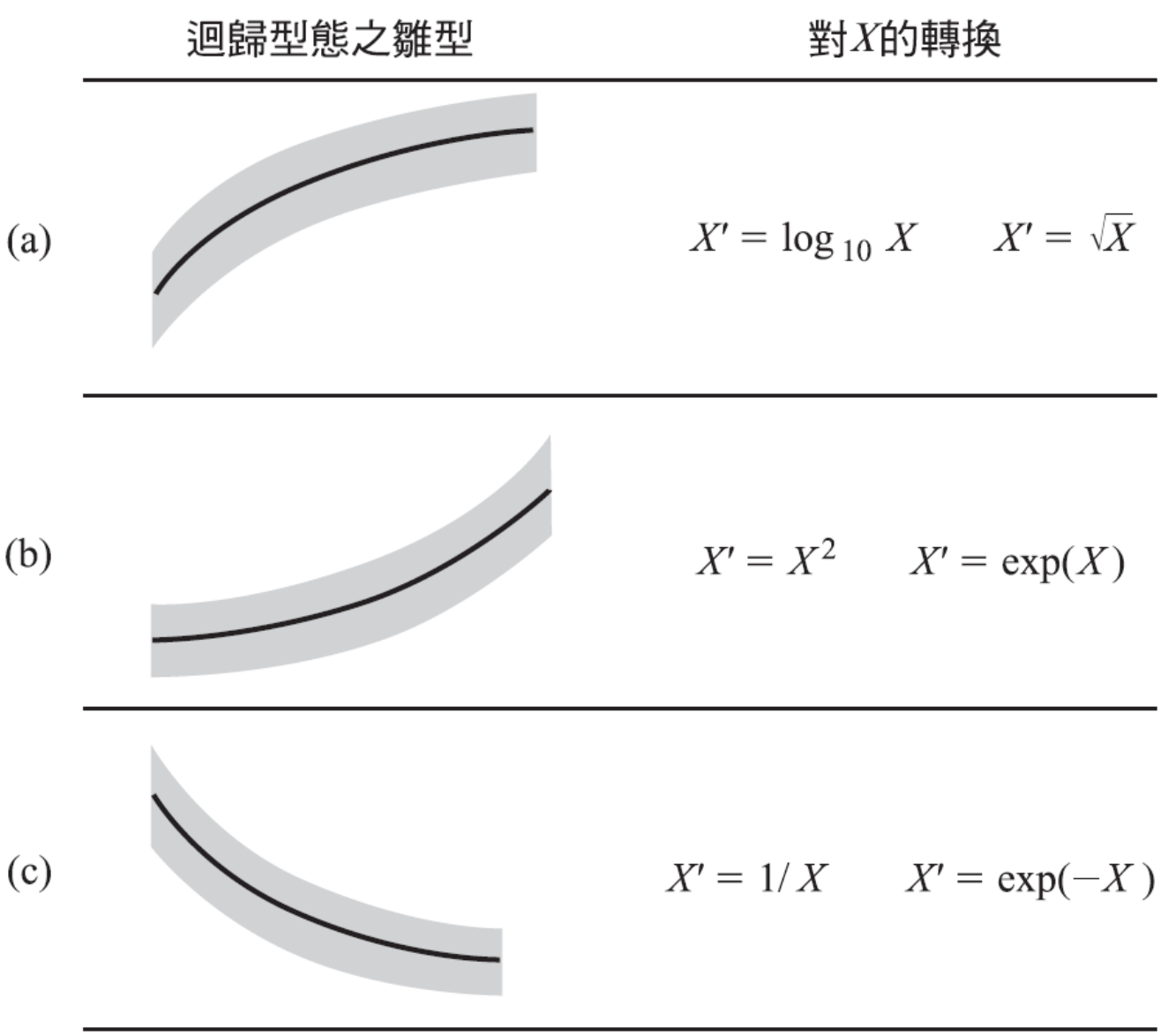


表3.7的第一個欄位與第二個欄位之資料，分別是接受訓練的天數(X)與銷售業績(Y)之表現結果，總共有十名參與測試的人員，圖3.14a為資料之散佈圖，我們可以看出曲線型態的迴歸關係，所以並不適用簡單線性迴歸模型(2.1)，由於在不同的X水準下，其變異性相當接近常數，所以我們可以考慮對X進行轉換，根據圖3.13a的雛型，首先考慮平方根轉換$X'=\sqrt{X}$，經轉換後的資料列於表 3.7 的第三個欄位。

對轉換後的資料$X'=\sqrt{X}$畫出之散佈圖，從圖3.14b中可以合理的假設其線性關係，由於在不同的X水準下，資料散佈之範圍並未改變，所以並不需要對 Y 進行轉換。

現在進一步檢驗簡單線性迴歸模型(2.1)是否適用，先對轉換後的X資料進行迴歸配適，所需計算與前面相同，只是預測變數改成X'，配適函數之結果如下：

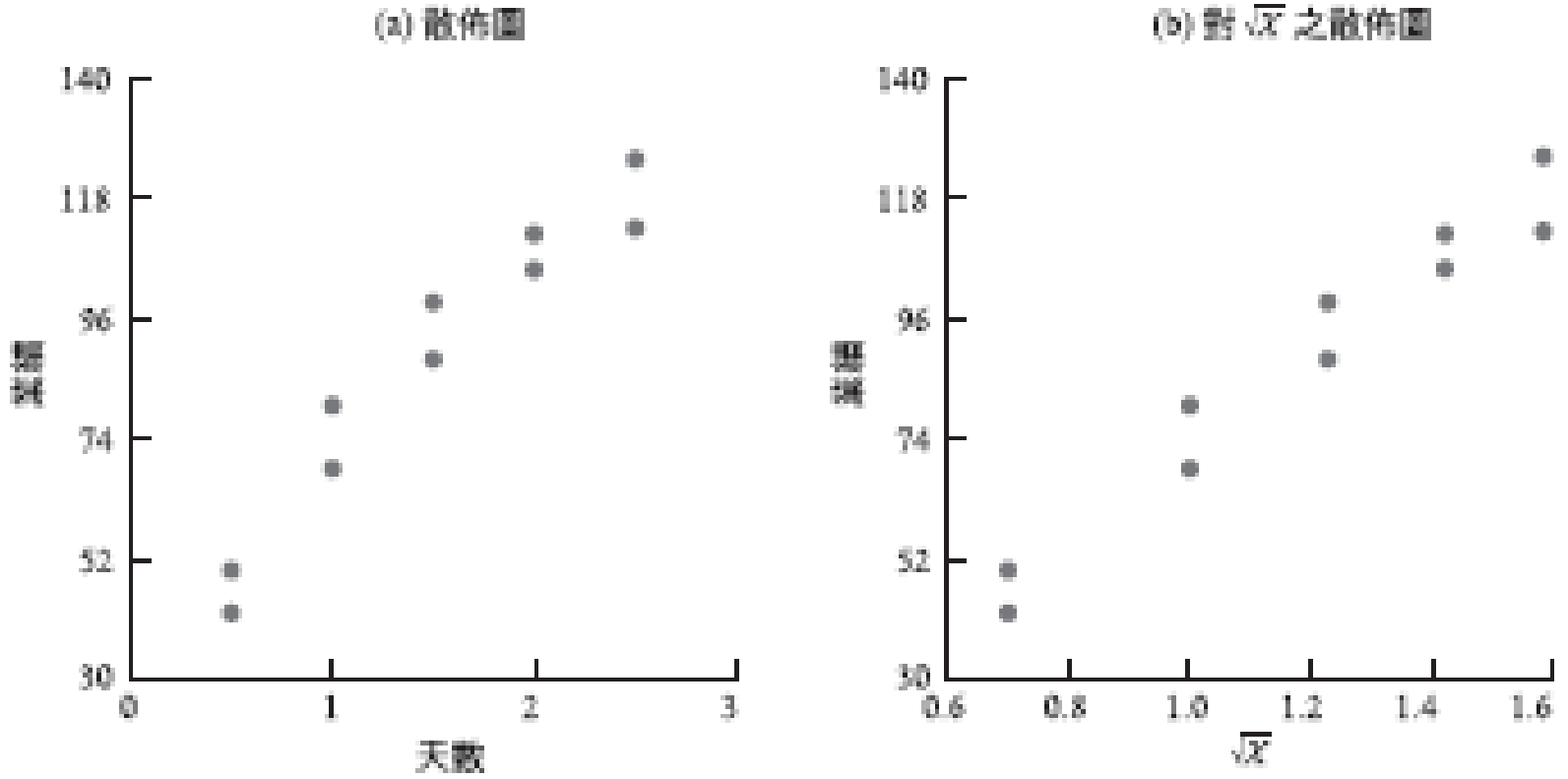
$$\hat{Y}=-10.33+83.45X'$$

表 3.7

對 X 進行平方根轉換後的線性迴歸關係－訓練天數與業績關係案例。

	(1)	(2)	(3)
測試人員	受訓天數	業績	
i	X_i	Y_i	$X'_i = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

圖3.14
散佈圖與殘差圖—訓練天數與業績關係案例。



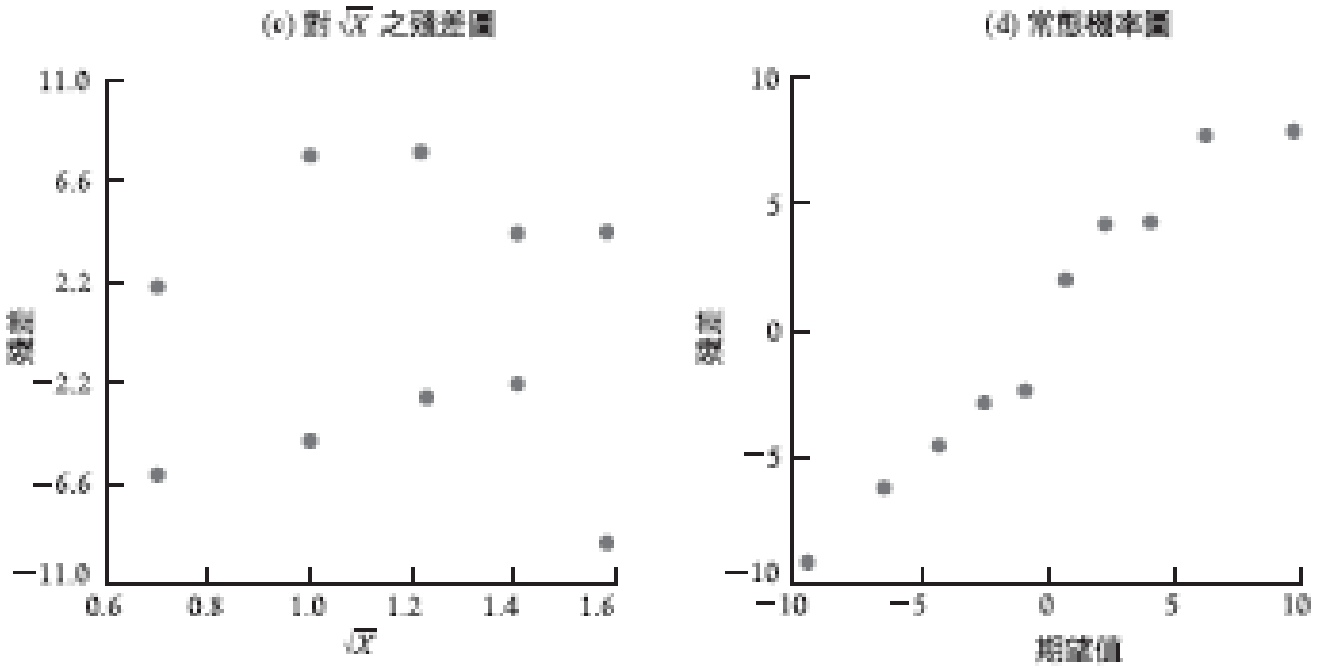


圖 3.14c 為對 X' 畫出之殘差圖，從圖形上來看並無明顯的證據指出不良配適，同時誤差變異數也在合理範圍內，圖 3.14d 是殘差的常態機率圖，同樣的並無偏離常態分配之明顯態勢，對殘差排序後的結果，與相對應常態分配之期望值，兩者間之相關係數高達 .979，在 $\alpha = .01$ 下，表 B.6顯示臨界值為 .879，於是我們可以更為支持常態分配之合理性，所以簡單線性迴歸模型(2.1)對此轉換後的資料應該是合適的，於是我們可以表示出對原始 X 所配適出的迴歸函數如下：

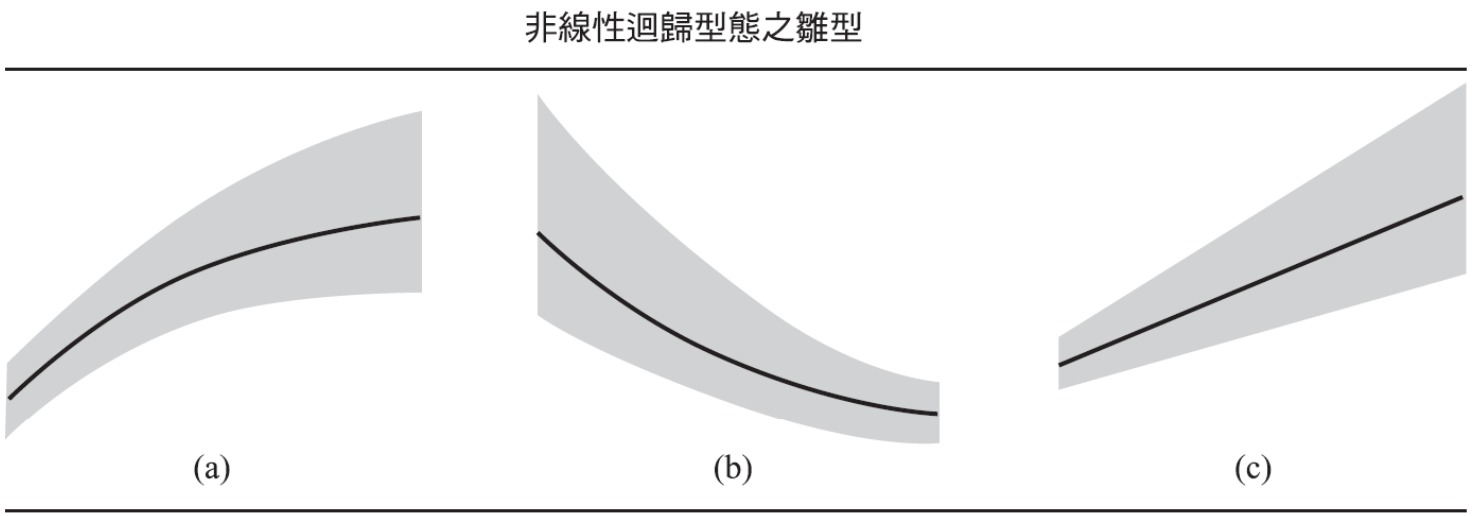
$$\hat{Y} = -10.33 + 83.45\sqrt{X}$$

有關非常態以及相異誤差變異數之轉換

- 誤差變異數不是常數或誤差項不服從常態分配，這兩種情形經常同時發生，爲矯正此種偏離簡單線性迴歸模型 (2.1) 的情形，我們此時可以針對 Y 進行轉換，同時也可能有助於曲線迴歸關係之線性化，另外有時可以同時針對X進行轉換，已獲取或保持原有的線性迴歸關係。
- 誤差項不服從常態分配或變異數不是常數的現象經常是遞增偏態(increasing skewness)，或誤差項之變異數平均反應 E｛Y｝變大而增加，例如，關於研究家庭渡假之全年支出(Y)，與家庭的收入所得之迴歸關係中，可以發現對於高收入所得的家庭會有較大的變化與較高的正偏態(positive skewness)，亦即全年的家庭渡假支出會高於更線性比例之預期，而低收入所得的家庭在渡假支出上總是顯得較低且變化不大。

- 圖3.15為幾種誤差變異數及偏態隨平均反應E{Y}變大而增加的迴歸型態之雛型，同時也同時也列出幾種有關Y的簡單轉換，有時嘗試多種不同的轉換後，並分別畫出散佈圖及殘差圖來檢視分析，方能決定最有效之轉換方式。

圖 3.15
非常數誤差變異數下的迴歸型態雛型以及對 Y 的簡單轉換。



對 Y 的轉換

$Y' = \sqrt{Y}$

$Y' = \log_{10} Y$

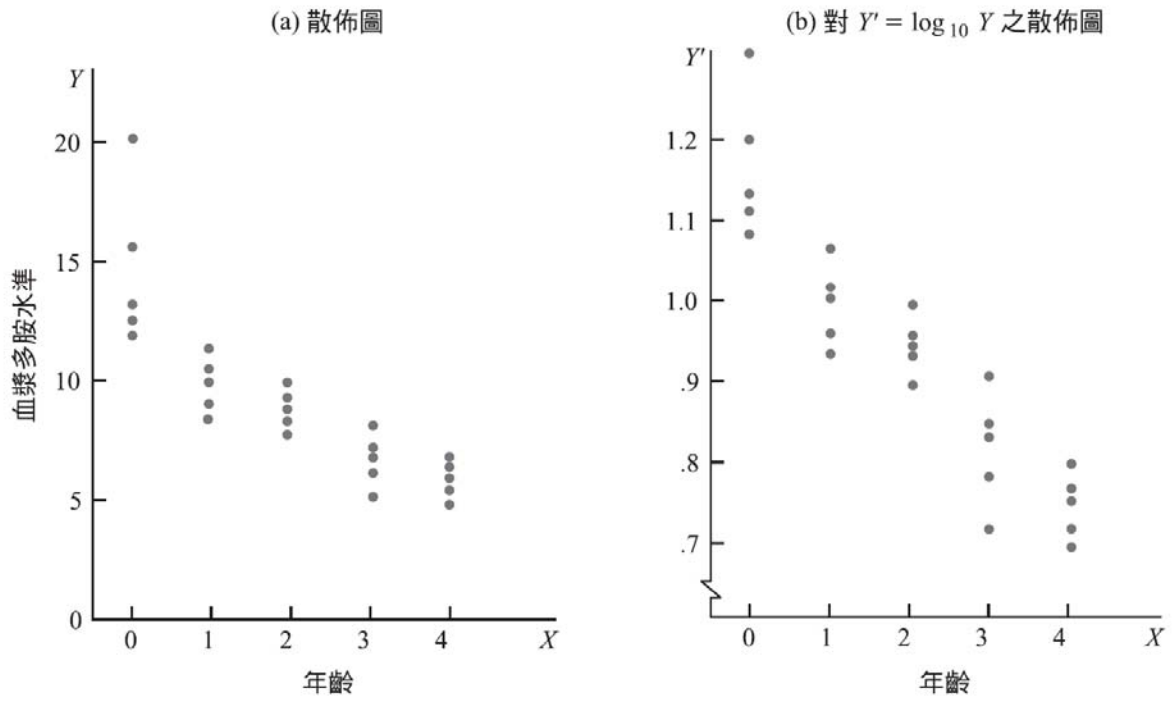
$Y' = 1/Y$

註：同時針對 X 進行轉換經常有用且必要。

一項針對25名健康幼兒所進行之研究，在表3.8的第一個欄位與第二個欄位之資料，分別是他們的年齡(X)與血漿多胺水準(Y)，圖3.16a為資料之散佈圖，圖形顯示出曲線型態的迴歸關係，而且年齡越小，資料的變異性越大。

根據圖3.15b的迴歸型態雛型，可先考慮利用對數轉換 $Y' = \log_{10} Y$，轉換後之結果在表3.8的第三個欄位，圖3.16b為其散佈圖，轉換後的資料其線性關係顯的較為合理，同時在不同的X水準下，其變異性也相當接近常數。

圖 3.16
散佈圖與殘差圖－血漿多胺案例。



幼兒 i	(1) 年齡 X_i	(2) 血漿多胺水準 Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 （新生兒）	13.44	1.1284
2	0 （新生兒）	12.84	1.1086
3	0 （新生兒）	11.91	1.0759
4	0 （新生兒）	20.09	1.3030
5	0 （新生兒）	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

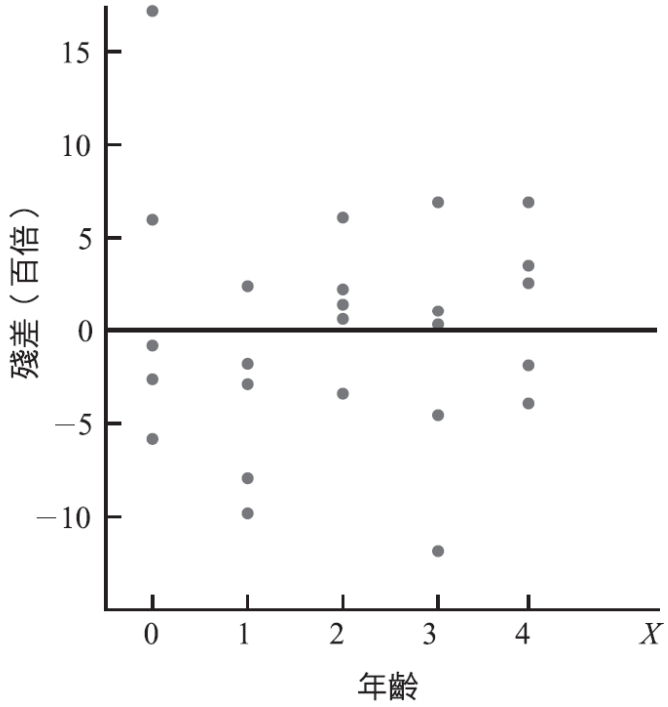
表 3.8
以 Y 的對數轉換線性化迴歸關係並穩定誤差變異數－血漿多胺案例。

為了更進一步檢驗 $Y' = \log_{10} Y$ 的轉換是否適當，可以對轉換後之 Y 資料配適簡單線性迴歸模型(2.1)，結果如下：

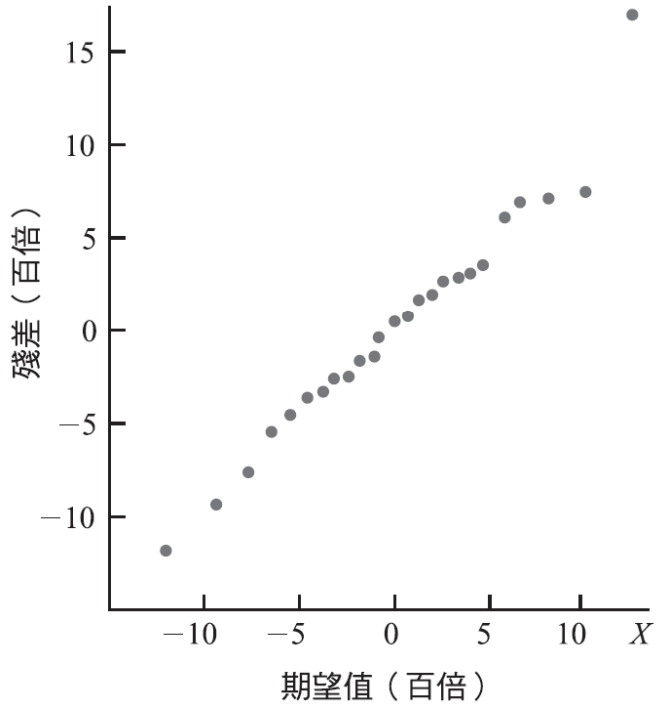
$$\hat{Y}' = 1.135 - .1023X$$

對 X 之殘差圖見圖 3.16c，圖 3.16d 是殘差的常態機率圖，對殘差排序後的結果，與相對應常態分配之期望值，兩者間之相關係數高達 .981，在 $\alpha = .05$ 下，表 B.6 顯示臨界值為 .959，顯示出支持常態分配之合理性，所以簡單線性迴歸模型(2.1)對此轉換後的資料應該是合適的。

(c) 對 X 之殘差圖



(d) 常態機率圖



說明

1. 有時對 Y 轉換時須引入一個常數，例如當 Y 為負數時，以對數轉換為例，可以先將 Y 進行平移，使得所有的觀測值 Y 均為正數，此時 $Y' = \log_{10}(Y + k)$，其中 k 為一個經過選擇後的適當常數。
2. 當誤差項之變異數不是常數，但是迴歸關係顯示為線性時，只有對 Y 轉換可能還不夠，因為會使變異數穩定的轉換有可能會將直線關係轉為曲線，此時可以考慮對 X 轉換，同時利用第 11 章說明的加權最小平方法來處理。 ■

- **Box-Cox 轉換**

  *Box-Cox*程序自動由一組 Y 的**乘冪轉換族**找出一種轉換，其形式如下：

$$Y' = Y^{\lambda} \tag{3.33}$$

其中參數由資料決定，此一轉換族包含下列轉換：

$$\begin{aligned}
\lambda &= 2 && Y' = Y^2 \\
\lambda &= .5 && Y' = \sqrt{Y} \\
\lambda &= 0 && Y' = \log_e Y \quad \text{（根據定義）} \\
\lambda &= -.5 && Y' = \frac{1}{\sqrt{Y}} \\
\lambda &= -1.0 && Y' = \frac{1}{Y}
\end{aligned} \tag{3.34}$$

反應變數透過(3.33)的轉換後，常態誤差迴歸模型變成如下之形式：

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3.35}$$

*Box-Cox*程序透過最大概似法來估計λ，並將所求得之λ作為乘冪轉換，對每一個λ值，首先將Y_i^{λ}標準化以確保誤差平方和的大小不會因值而產生變化：

$$W_i = \begin{cases} K_1\left(Y_i^{\lambda} - 1\right) \\ K_2\left(\log_e Y_i\right) \end{cases} \tag{3.36}$$

其中，

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} \tag{3.36a}$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}} \tag{3.36b}$$

而K_2正式觀測值Y_i的幾何平均數

- 在給定值並計算出標準化觀測值之後，對預測變數進行迴歸程序，算出誤差平方和SSE。
- 在任何情況下透過 Box-Cox 程序得到的轉換仍然需要畫出散佈圖與殘差圖來檢驗其適切性

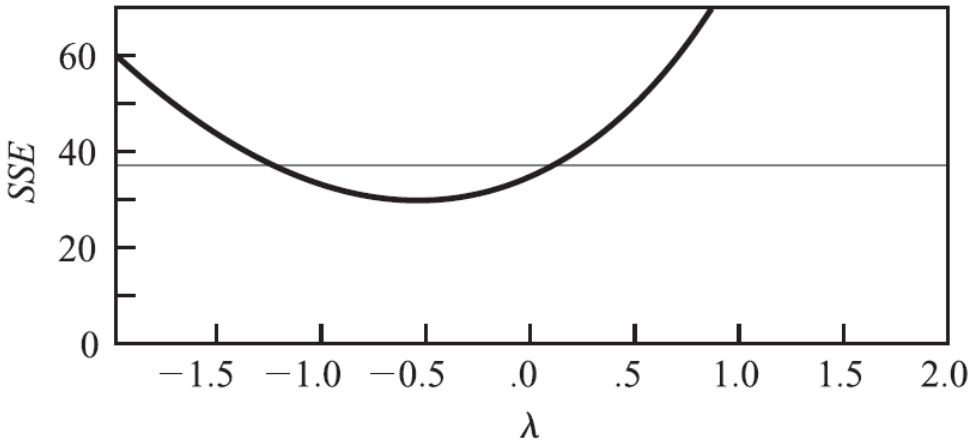
例題

表3.9是血漿多胺水準案例中透過Box-Cox程序的結果，λ值的範圍選擇從 −1.0 到 1.0 ，對每一個所選定的λ值，做(3.36)的轉換，並配適 W 對 X 的線性迴歸，以λ = .5 為例，首先計算 $W_i = K_1\left(\sqrt{Y_i} - 1\right)$ ，然後配適 W 對 X 的線性迴歸，並計算誤差平方和 $SSE = 48.4$ ，透過 Box-Cox 程序找到的乘冪轉換參數λ值接近 −.5 ，此時有最小的 $SSE = 30.6$ 。

表 3.9
Box-Cox 程序結果－血漿多胺案例。

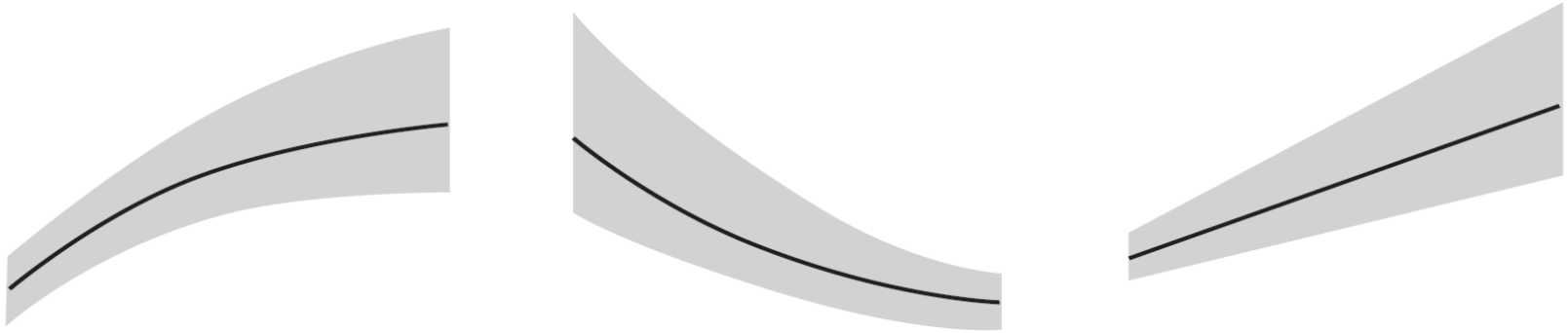
λ	*SSE*	λ	*SSE*
1.0	78.0	−.1	33.1
.9	70.4	−.3	31.2
.7	57.8	−.4	30.7
.5	48.4	−.5	30.6
.3	41.4	−.6	30.7
.1	36.4	−.7	31.1
0	34.5	−.9	32.7
		−1.0	33.9

圖 3.17
SAS-JMP 的 Box-Cox 程序結果－血漿多胺案例。



■圖 3.17 之結果很清楚指出適當的轉換在 λ 值接近 -0.5 附近

非線性迴歸型態之雛型



(a)　(b)　(c)

3.10 迴歸函數型態之探索

Exploration of Shape of Regression Function

- 散佈圖經常能夠清楚地指出迴歸函數之性質，例如圖1.3即能清楚說明類固醇與年齡間迴歸關係之曲線特性，不過有時散佈圖會過於複雜而難以顯示迴歸關係，此時，在不對迴歸函數做出任何限制下而進行平滑線的配適，將有助於迴歸函數的探索，這樣的平滑曲線我們稱之為無母數迴歸曲線(nonparametric regression curves)。
- 已經有許多平滑方法可供應用於獲得時間數列資料之平滑曲線，例如以X_i代表等時間間距，
  - **移動平均法(method of moving averages)**透過相鄰數期的Y觀測值平均數得到平滑曲線，例如最初第三期觀測值平均數為此三時期的中間點，也就第二期的平滑值，而第二、三、四期觀測值平均數則為第三期的平滑值，以此類推。

- **流動中位數法(method of running medians)** 類似移動平均法，不過它是透過中位數當作平均測度以減少離群觀測值對平滑效果之影響，不管是移動平均法或是流動中位數法，均可以對平滑值再做一次平滑程序，或是透過其他更為精細的方法以提供時間數列更適當的平滑曲線。

■ 對於 X_i 不是等時間間距之資料時，我們可以用**波段迴歸(band regression)**來進行平滑程序，首先將資料分成許多由相鄰的X_i水準所組成的波段(band)，對於每一波段分別計算X與Y的分配對中位數，然後對於所計算出中位數用值線連接起來即可，舉例說明，設下列資料我們將其分成三各波段如圖： 然後畫出三個中位數的點並以直線連接起來，便成為簡單的無母數回歸曲線

X	Y	中位數 X	中位數 Y
2.0	13.1	2.7	14.4
3.4	15.7		
3.7	14.9		
4.5	16.8	4.5	16.8
5.0	17.1		
5.2	16.9		
5.9	17.8	5.55	17.35

Lowess 法

- 基本原理是透過連續而局部的線性迴歸來得到平滑曲線，相似於移動平均法或流動中位法， Lowess法代表的意義為局部加權迴歸散佈平滑法(locally weighed regression scatter smoothing)
- 利用每一個X水準之鄰域而計算出相對應的Y值，根據各X鄰域之資料來配適線性迴歸，然後以該X的適配值做平滑值。 舉例而言，(X_1, Y_1)表示最小的X水準之資料若 Lowess 法中採用三個X水準之鄰域，則下料資料

  $$(X_1, Y_1) \quad (X_2, Y_2) \quad (X_3, Y_3)$$

  則可配適一個線性迴歸所適配出的X_2 值即為對應於 X_2 之平滑值，對下列資料配適另外一個線性迴歸

  $$(X_2, Y_2) \quad (X_3, Y_3) \quad (X_4, Y_4)$$

  所適配出的X_3 值即為對應於 X_3 之平滑值，

■Lowess法在計算出最後平滑值時，進行下列步驟

1.採用加權線性迴歸，對於鄰域中離中間X水準較遠處之資料給予較小的權值

2.為此此一程序對離群之觀測值具有穩健性，所適配的線性迴歸程序將會重複，若第一次配適迴歸時，造成較大殘差的個案，於第二次配適迴歸時，將給予較小的權值

3.為更進一步改善此程序之穩健性，可重複步驟二多次，依照最後一次所適配之殘差大小修正權值

■使用Lowess法程序之前必須適當的選擇每一段迴歸線之鄰域大小，同時也必須決定出一個加權函數，給予鄰域中離中間X水準較遠處之資料有較小的權值，並另定一個加權函數，給予殘差大之個案較小的權值，最後決定出能使整體程序穩健的迭帶次數。

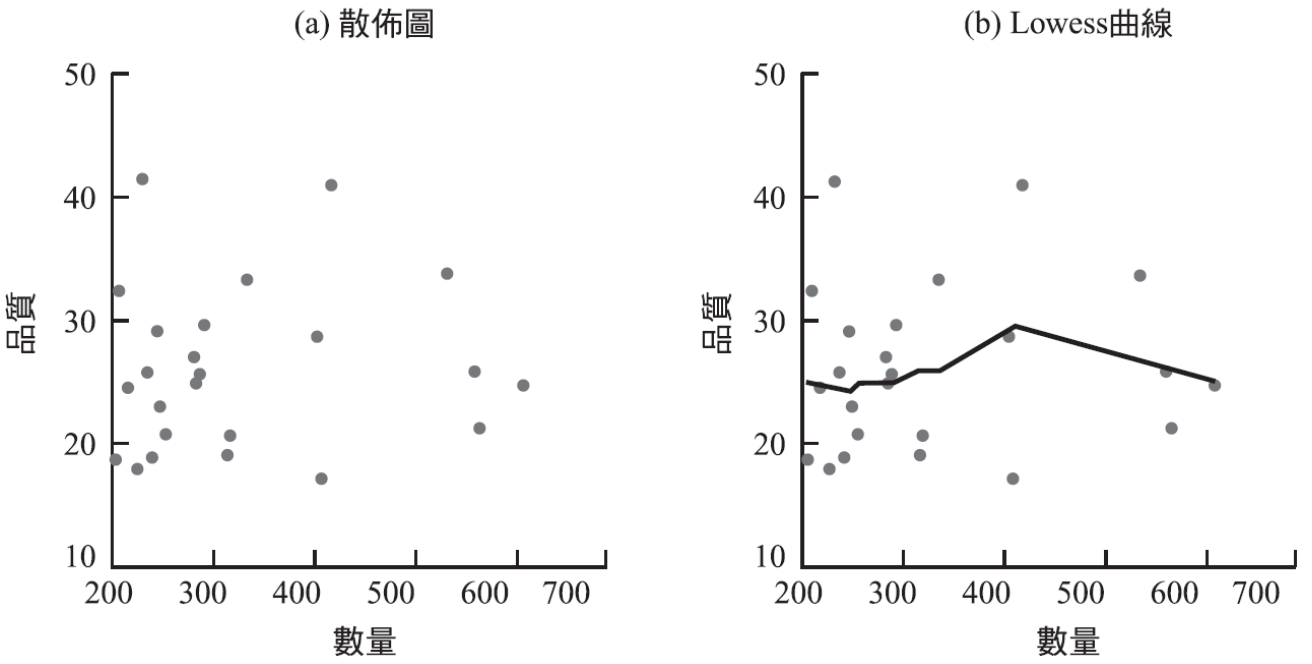


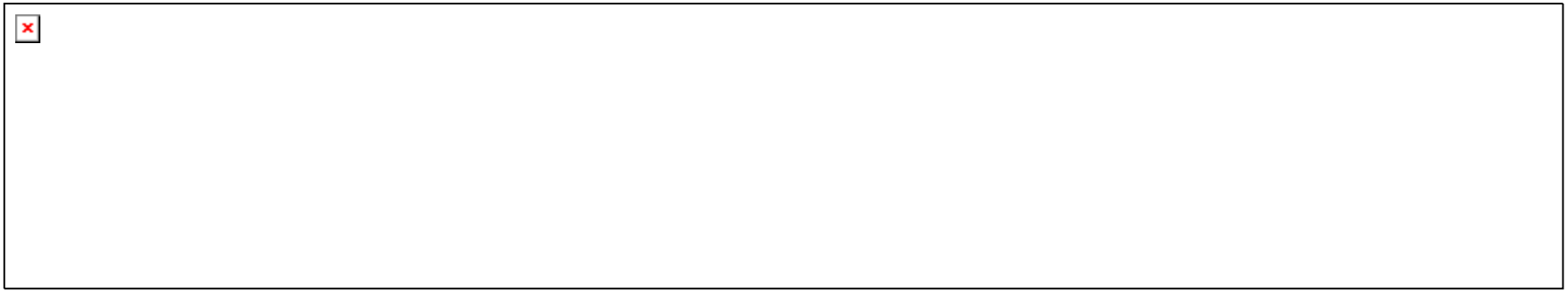
圖 3.18
MINITAB 軟體之散佈圖與 Lowess 平滑曲線－研究所案例。

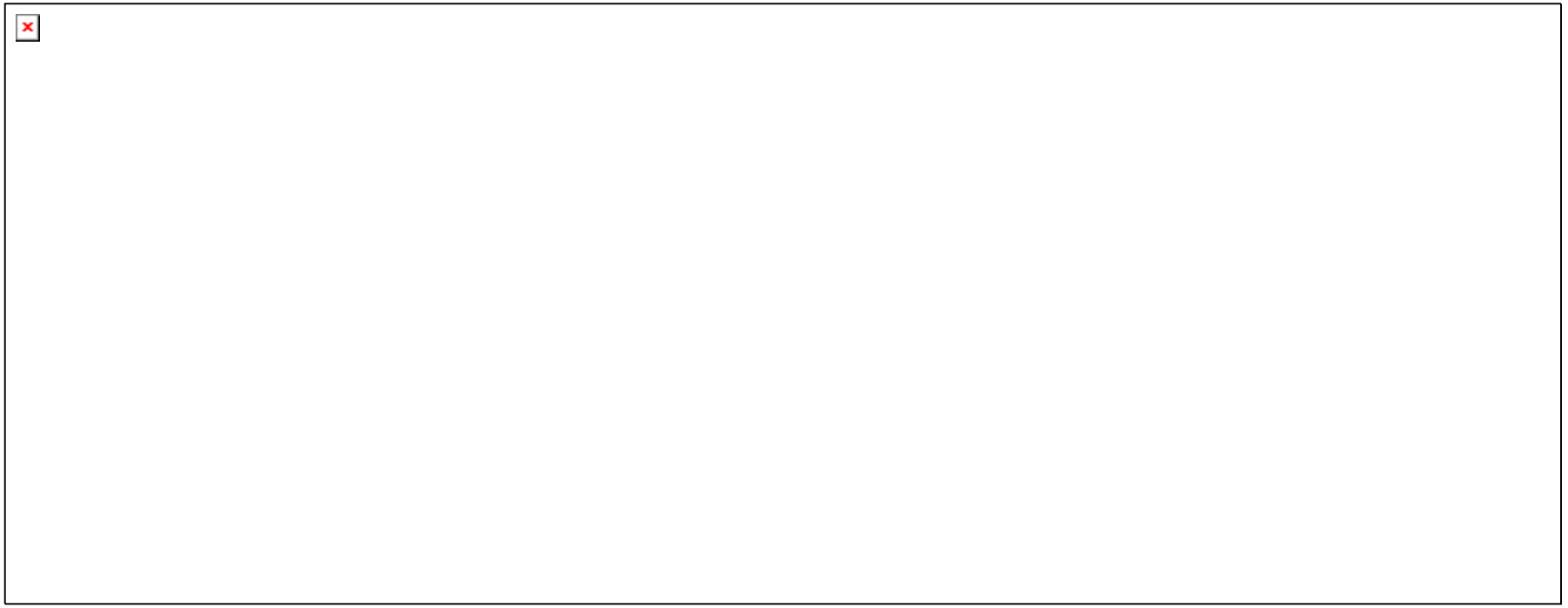
例題

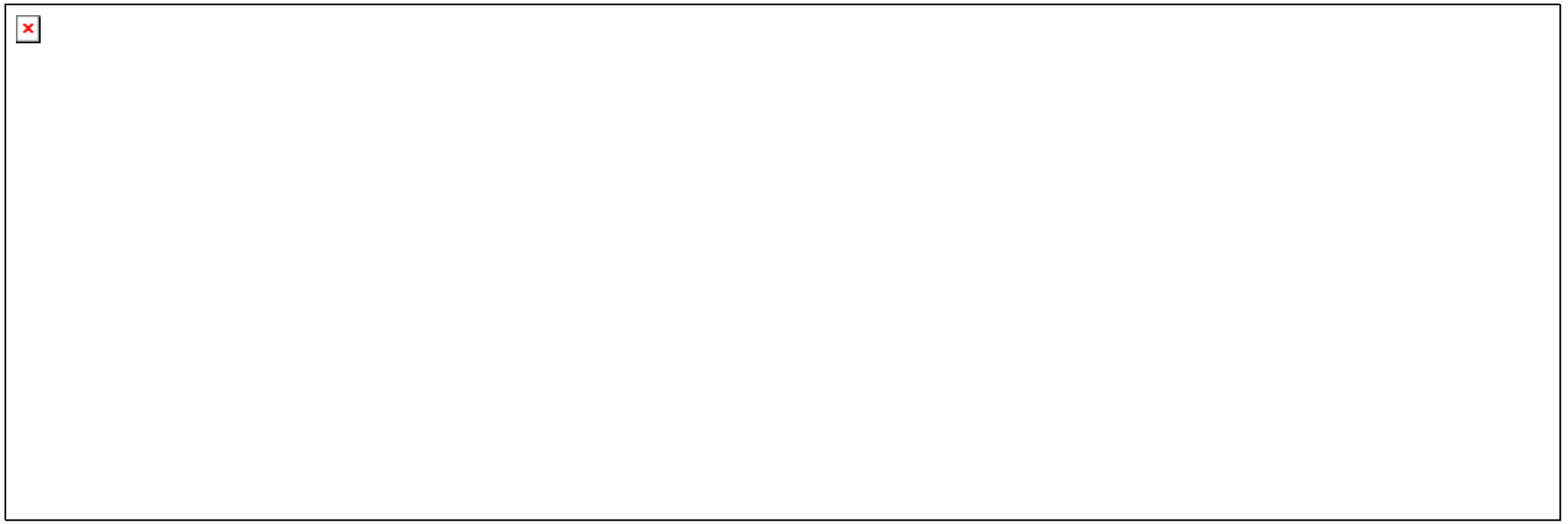
散佈圖3.18a是根據24個研究所教育品質之調查結果的評價資料，反應變數為所完成之研究成果的量化品質，而解釋變數則為各研究所在實驗室所執行的研究數量之測量值，從散佈圖中不容易看出研究品質與研究數量之間的關係，散佈圖3.18b除了畫出原先之觀測點外，再加入Lowess平滑曲線，曲線中顯示研究數量中等程度的研究所，有較高之研究品質，不過因為分散程度過大，Lowess曲線上的關係，僅能說明這種可能性，同時也因為研究品質與研究數量的指標都有其限制性，所以必須另外考慮其他指標以證實圖 3.18b 所顯示之關係。

透過平滑曲線以確認所配適之迴歸函數

- 平滑曲線不僅適用於迴歸模型的探索階段，也有助於判斷所選擇之迴歸函數是否適當，簡單的確認程序可以將平滑曲線與所配適的迴歸函數信賴區間帶畫在一起，如果平滑曲線完全落在信賴區間帶內，則可以成為支持迴歸函數的一項證據。



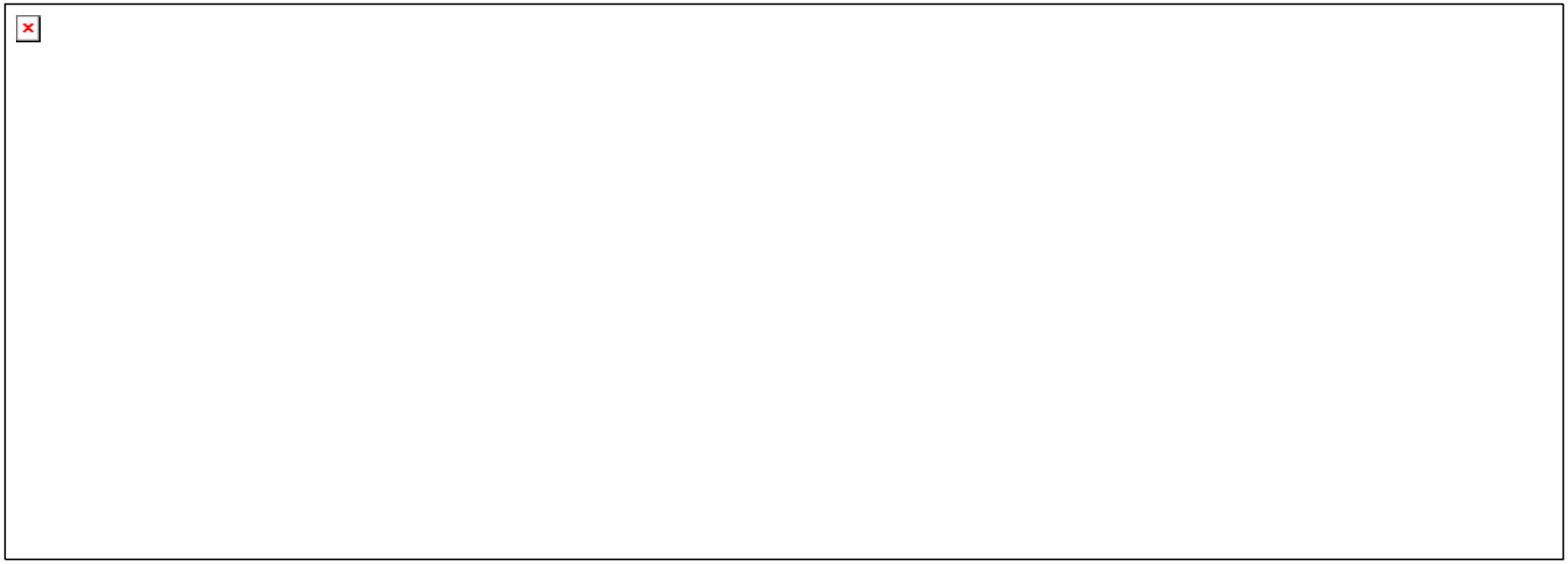




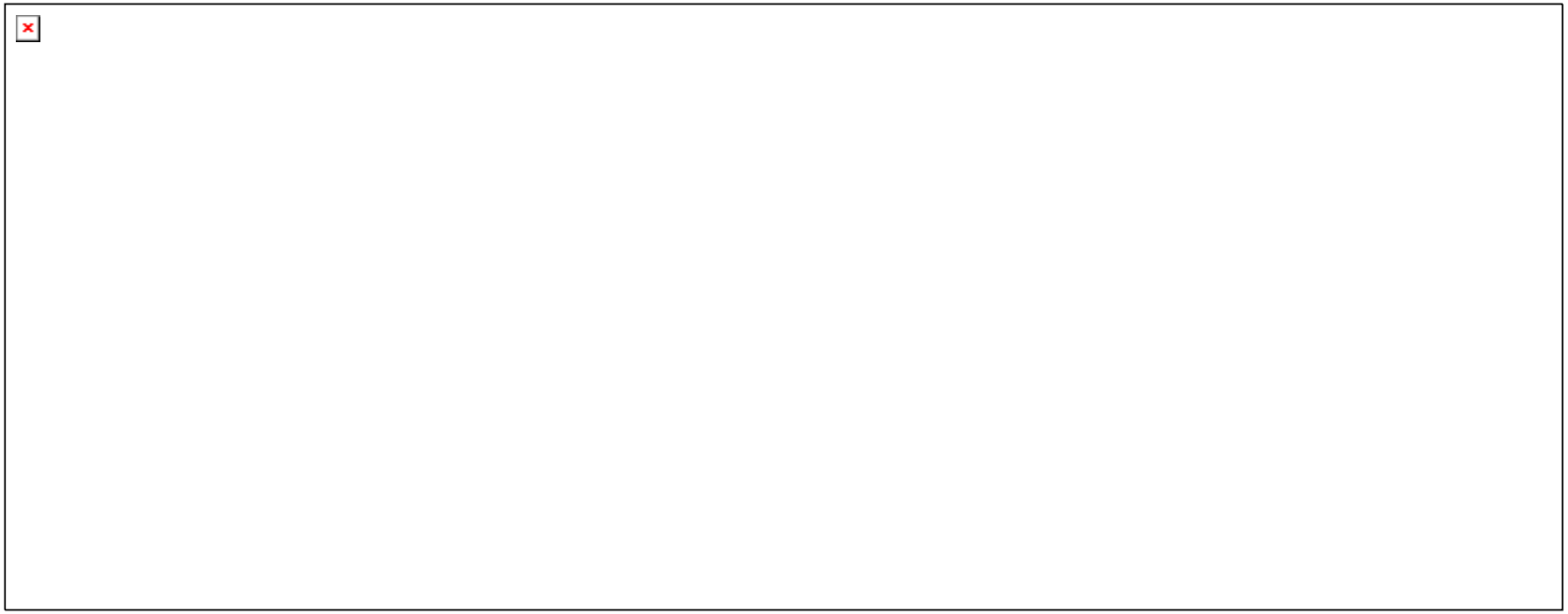
3.11 個案研究－鈽之測量

Case Example-Plutonium Measurement

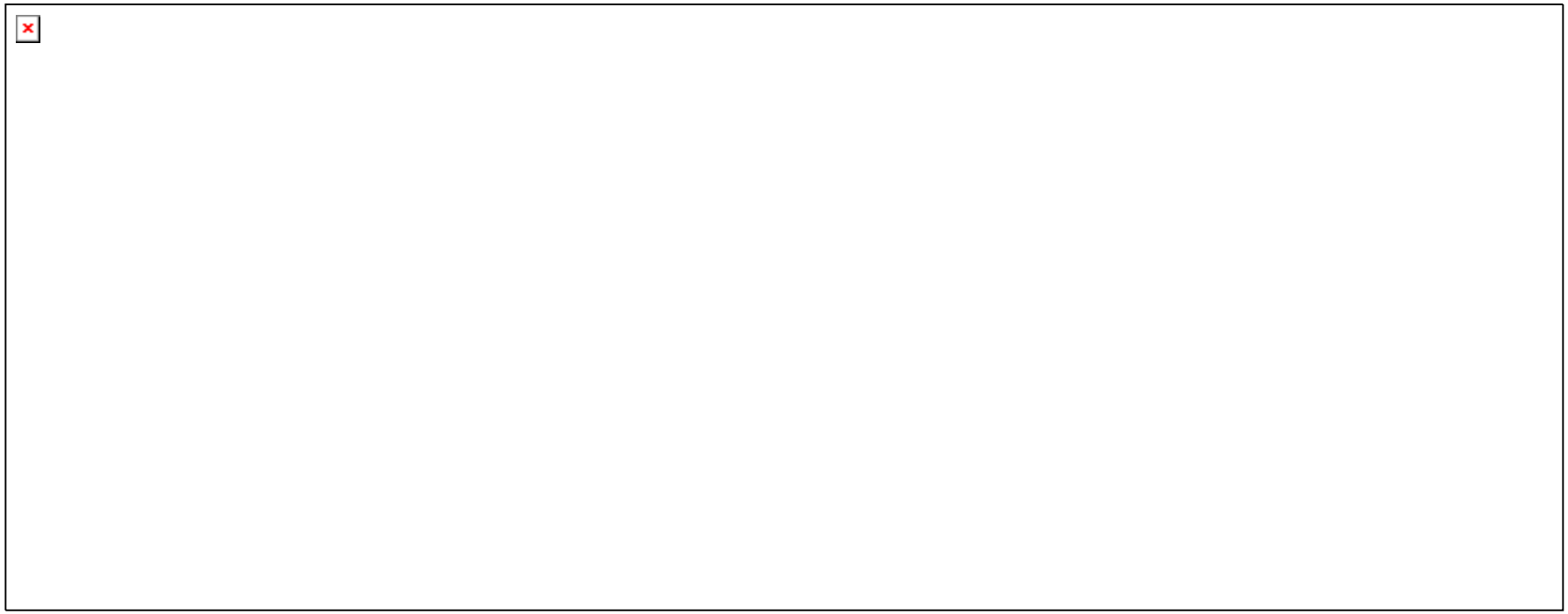
- 在重建區內一些有關於核物料如鈽238的環境清潔工作，由於鈽元素會釋出可被偵測之α粒子，所以仍然有儀器可以探測出α粒子的存在性
- 偵測並紀錄α粒子每一秒鐘內撞擊數的強度，迴歸關係中的反應變數為α粒子的撞擊數，而解釋變數則為鈽元素的活躍性。



- 表3.10為實驗的部分資料，資料之散佈圖繪於途3.20a中，如預期般，α粒子撞擊數隨著鈰元素的活躍性而遞增，不過即使是不含鈰元素的標準材料也有觀測出非零之撞擊數，主要是因為背景輻射之關係，因此我們所配適的迴歸模型須具備有截距項 。



可以先對迴歸關係計算其 Lowess 法平滑曲線並繪於圖 3.20b 中，以進行初步之檢驗，圖中顯示出此一迴歸關係可能為直線或是具有些微彎度之曲線，而在 0.0 pCi/g下所紀錄出的第 24 個個案資料，顯示的結果相當異常，經事後檢查發現該次實驗之條件並非維持在所應當控制的環境下，所以有理由將該筆資料予以剔除，在此同時，我們也注意到 Lowess 平滑過程給予離群值很小的權值而保持其穩健性。



- 接下來再對其他 23 個資料配適直線迴歸函數，SAS-JMP 軟體的迴歸程續結果在圖3.21a中，殘差相對於配適值之結果繪於圖3.21b中，圖3.21c則為常態機率圖，在JMP軟體的輸出結果中，標題“Model”表示變異數分析的迴歸成分，而標題“C Total” 表示修正後之總平方和，在輸出結果中顯示迴歸直線之斜率顯著不等於零(F^*=228.9984,P-值=.0000)，所以迴歸關係存在。

圖 3.21

SAS-JMP 對於轉換資料前之迴歸輸出結果與診斷圖形－鈽之測量。

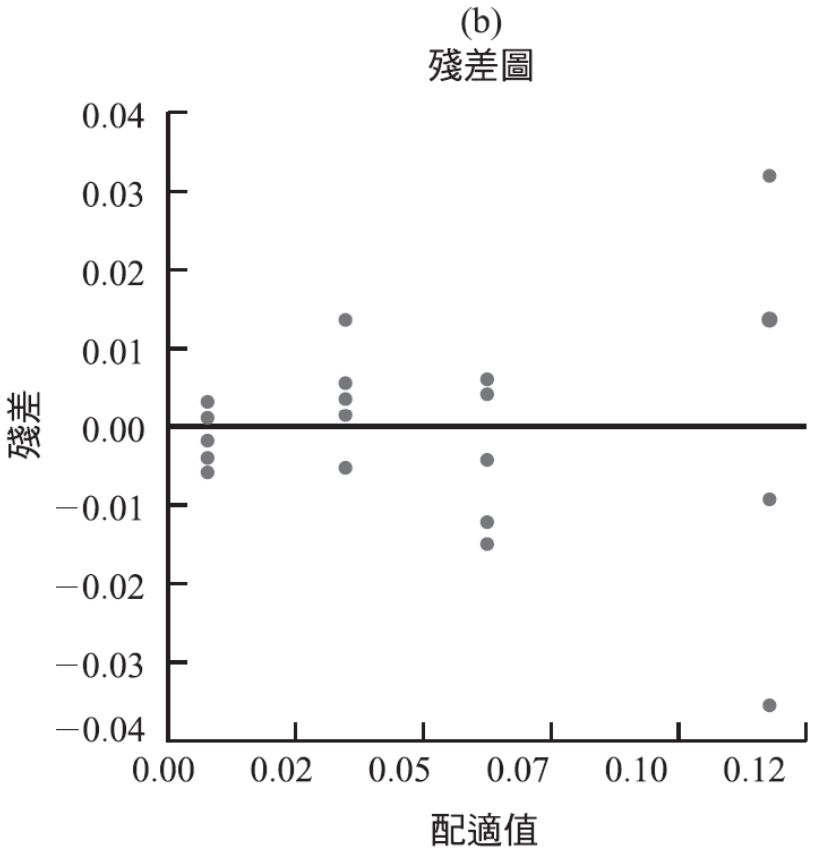
(a) 迴歸輸出

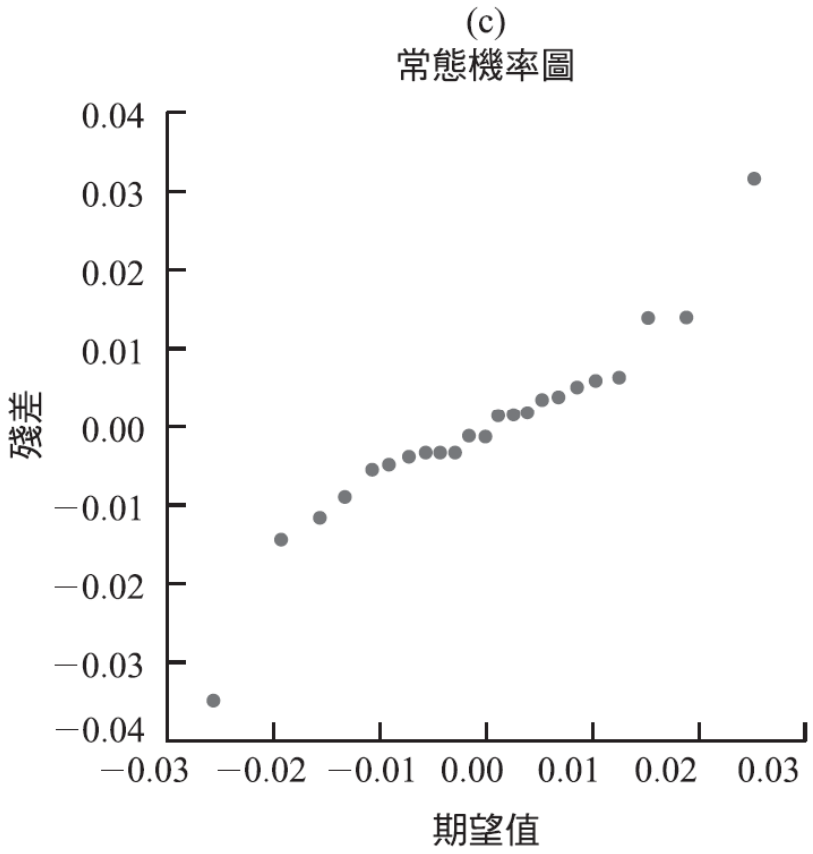
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0070331	0.0036	1.95	0.0641
Plutonium	0.005537	0.00037	15.13	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.03619042	0.036190	228.9984
Error	21	0.00331880	0.000158	Prob>F
C Total	22	0.03950922		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.00016811	0.000084	0.5069
Pure Error	19	0.00315069	0.000166	Prob>F
Total Error	21	0.00331880		0.6103

- 不過殘差則呈現張開支喇叭形狀，這顯示了誤差變異數隨著鈰元素的活躍性而遞增，而常態機率圖也顯示了違反常態分配之假設條件(厚尾)，有時非直線的常態機率圖示受到不是常數的誤差變異數之影響





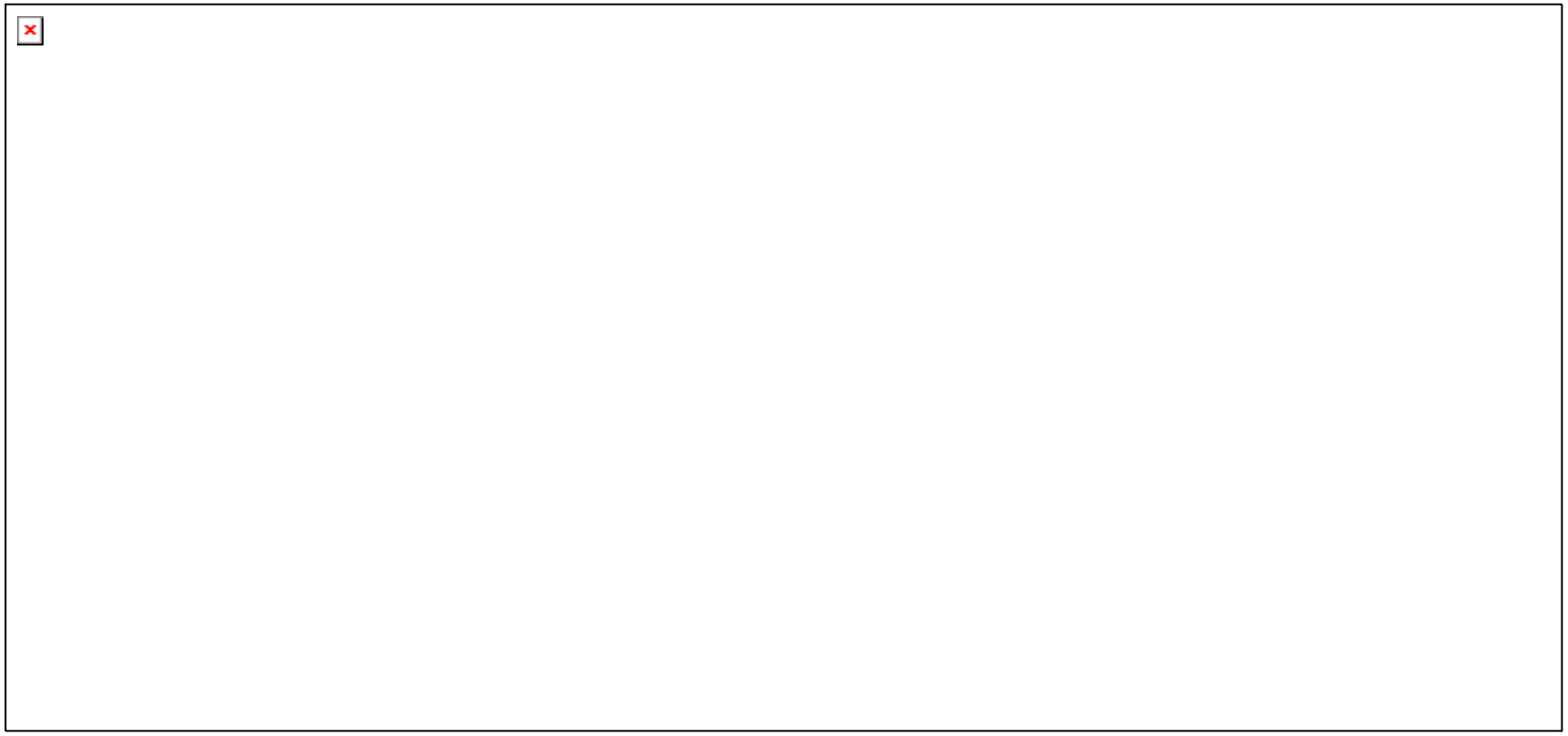
- 關於常數誤差變異數之檢定問題，我們可以透過Breusch-Pagan檢定統計量(3.11)來確認：

$$=23.29 > \chi^2(.95;1)=3.84$$

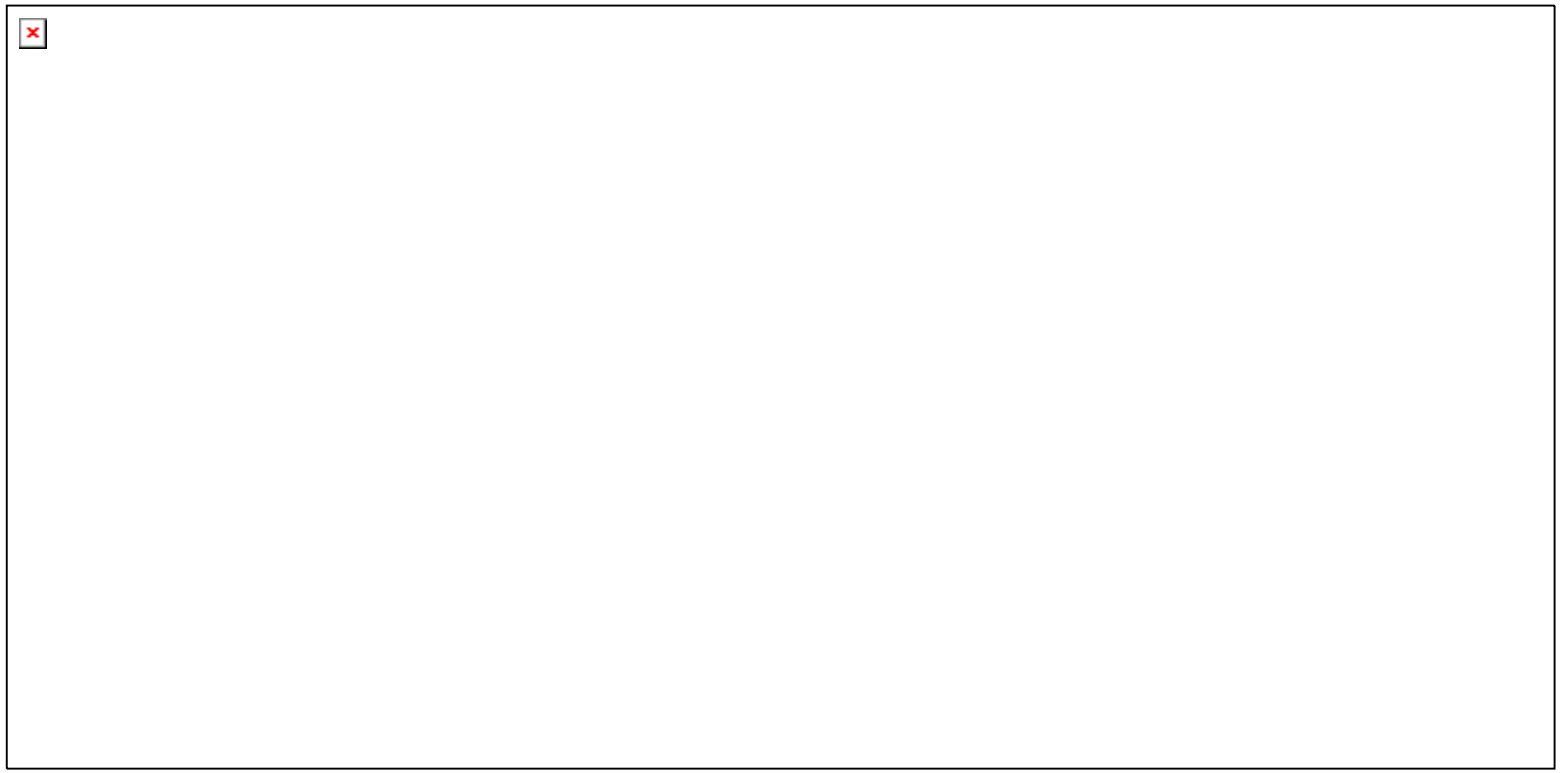
所以我們必須矯正此研究中關於非常數的誤差變異數之問題，第11章中的加權最小平方法是許多解決此一問題的方法之一，有時資料屬於計數性質時，可以藉由對反應變數進行平方方根之轉換，以穩定誤差變異數，

- 由於平方根之轉換僅僅只是成幕轉換下的一個特例，所以可由藉由Box-Cox程序提供我們適當轉換之建議，使用(3.36)的標準化變數，可以得到 λ 的最大概似估計量為 $\lambda = 0.65$，而概似函數 $\lambda = 0.65$附近相當平坦，所以 Box-Cox程序可以也將支持我們進行平方根之轉換(即 $\lambda=0.5$)，當反應變數為 $Y' = (Y)^{1/2}$時，直線函數之配適結果繪於圖3.22a。

- 不過現在有一個新的問題發生了，雖然圖3.22b顯示了誤差變異數較為穩定，而圖3.22c的常態機率圖也大致上呈現一條直線，不過在殘差圖上卻顯示了Ý對X的關係不是直線的，我們可以透過(3.25)的配適不良統計量得到確認(F*=10.1364,P-值=.0010)，事實上Y對X的關係的確是直線，所以轉換資料之後有此一結果，是可以預期的。



- 爲了恢復轉換資料之後Y對X的直線關係，我們可以考慮對X也進行平方根轉換，取$Y'=(Y)^{1/2}$對$X'=(X)^{1/2}$進行迴歸程序，其結果在圖3.23中，在殘差圖3.23b中，已經可以看出配是不良的問題得到解決，而圖3.23c的常態機率圖也有不錯的表現，從相關性檢定(γ=.986)來看也支持誤差像福從常態分配之假設



- (利用表B.6插補臨界值在α=.05，n=23時為.9555)。雖然殘差圖顯示的仍然是誤差變異數的不穩定，不過並不嚴重，透過 Breusch-Pagan 檢定統計量(3.11)得到 =3.85，對應的P-值=.05，同樣支持誤差變異數非常數之問題並不嚴重的結論。

圖 3.23

SAS-JMP 對於同時轉換反應變數與預測變數之資料後的迴歸輸出結果與診斷圖形－鈽之測量案例。

(a) 迴歸輸出

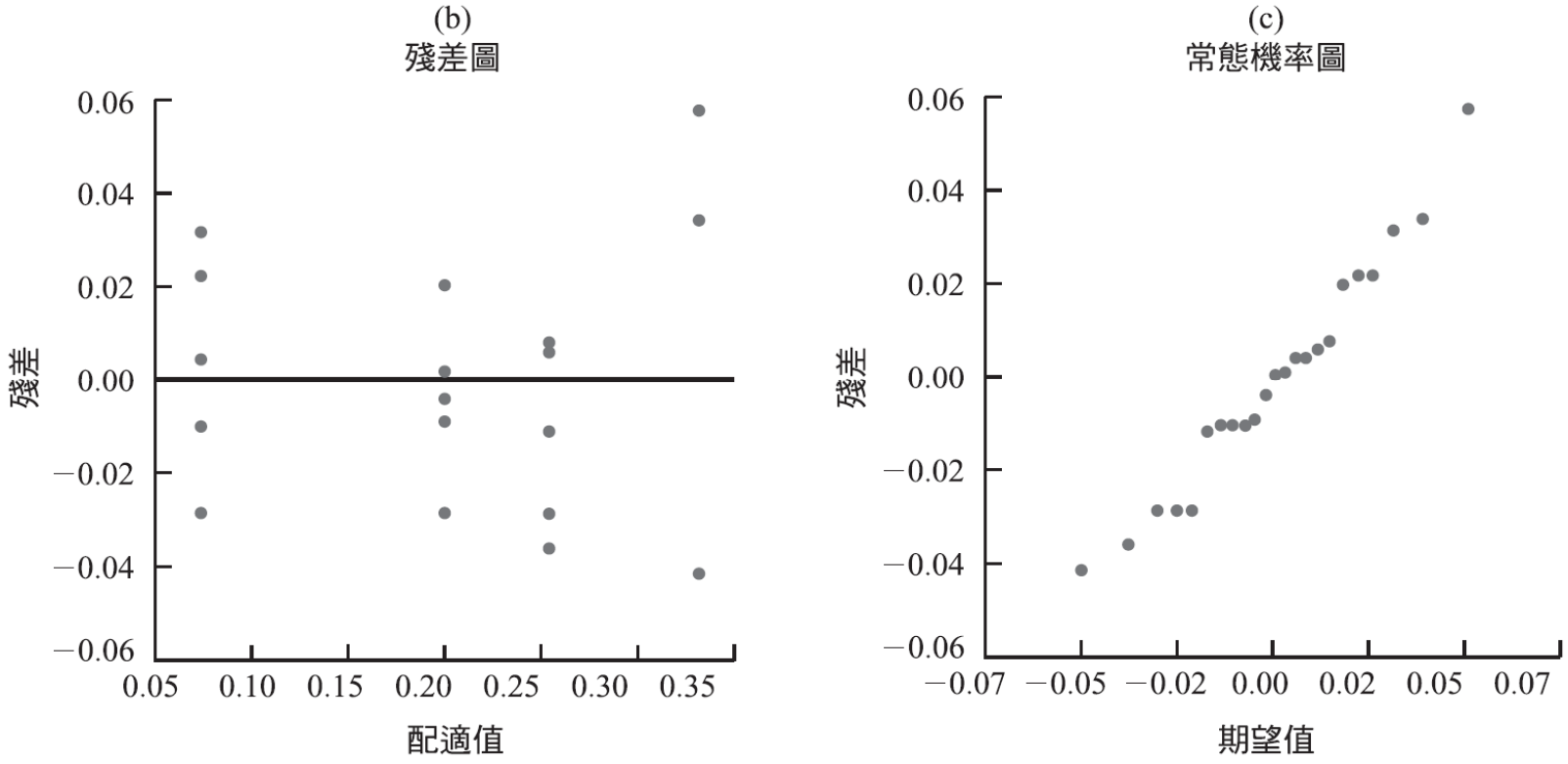
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0730056	0.00783	9.32	0.0000
Sqrt Plutonium	0.0573055	0.00302	19.00	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.22141612	0.221416	360.9166
Error	21	0.01288314	0.000613	Prob>F
C Total	22	0.23429926		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.00153683	0.000768	1.2868
Pure Error	19	0.01134631	0.000597	Prob>F
Total Error	21	0.01288314		0.2992

- 圖3.23d爲利用統計軟體SYSTAT所畫出之配適迴歸Y'=.0730＋.0573X'之信賴區間代(2.40)，由此信賴區間帶可知

$$\acute{Y}' = .0370 + .0573X'$$



- 迴歸線之估計相當精確，而Lowess平滑曲線完全落在此信賴區帶中，表示Ý與X'被假設成直線關係是合理的，配是不良的統計量(3.25)為F*=1.2868(P-值=.2992)，同樣支持$Y'=(Y)^{1/2}$與對$X'=(X)^{1/2}$的直線關係