簡單線性迴歸　第 1 部分

第 1 章

單一預測變數線性迴歸

Linear Regression with One Predictor Variable

迴歸分析（Regression analysis）

- 一種研究兩個或多個量化變數間關係之統計方法，作用在於透過某些變數來預測所關心的反應變數。
- 應用實例：廣告支出與營收之關係來預測產品之銷售

1.1 變數間之關係

- 兩變數間的關係區分**函數關係**（functional relation）與**統計關係**（statistical relation）兩種

- **兩變數間之函數關係**

通常兩變數間之函數關係可以透過數學方程式來表達，如果利用 X 表示**獨立變數**，Y 表示**相依變數**，則兩變數間之函數關係形式為

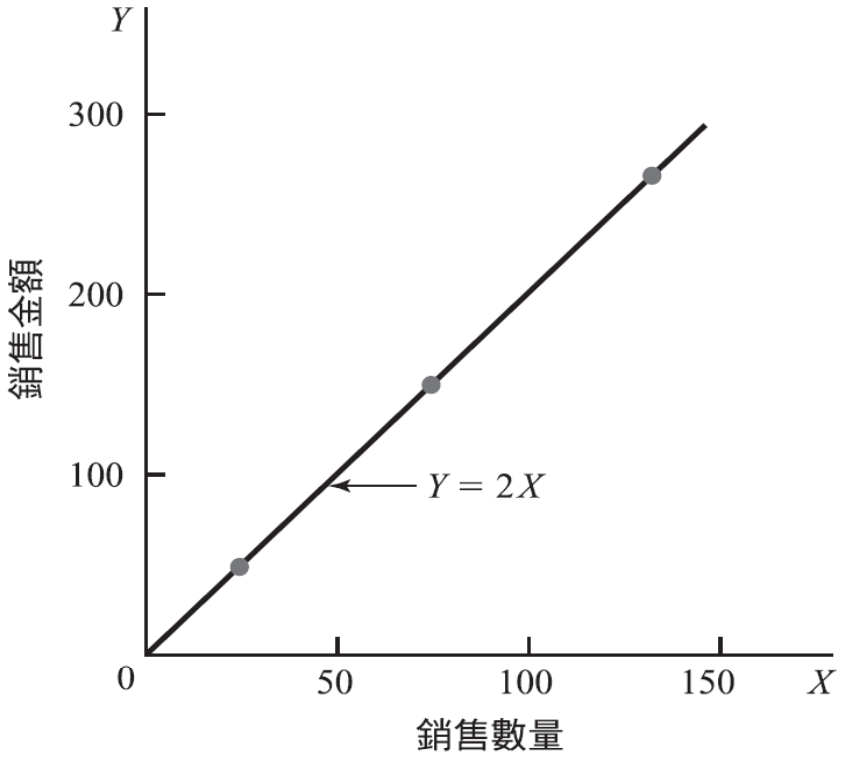
$$Y = f(X)$$

例題

產品在固定價格 2 美元之下，若用 X 表示產品之銷售數量，Y 表示銷售金額，則銷售金額與銷售數量之函數關係透過數學方程式可以表示為：

$$Y = 2X$$

這樣的函數關係可以透過圖 1.1 來說明。假設最近三期銷售金額與銷售數量分別如下：



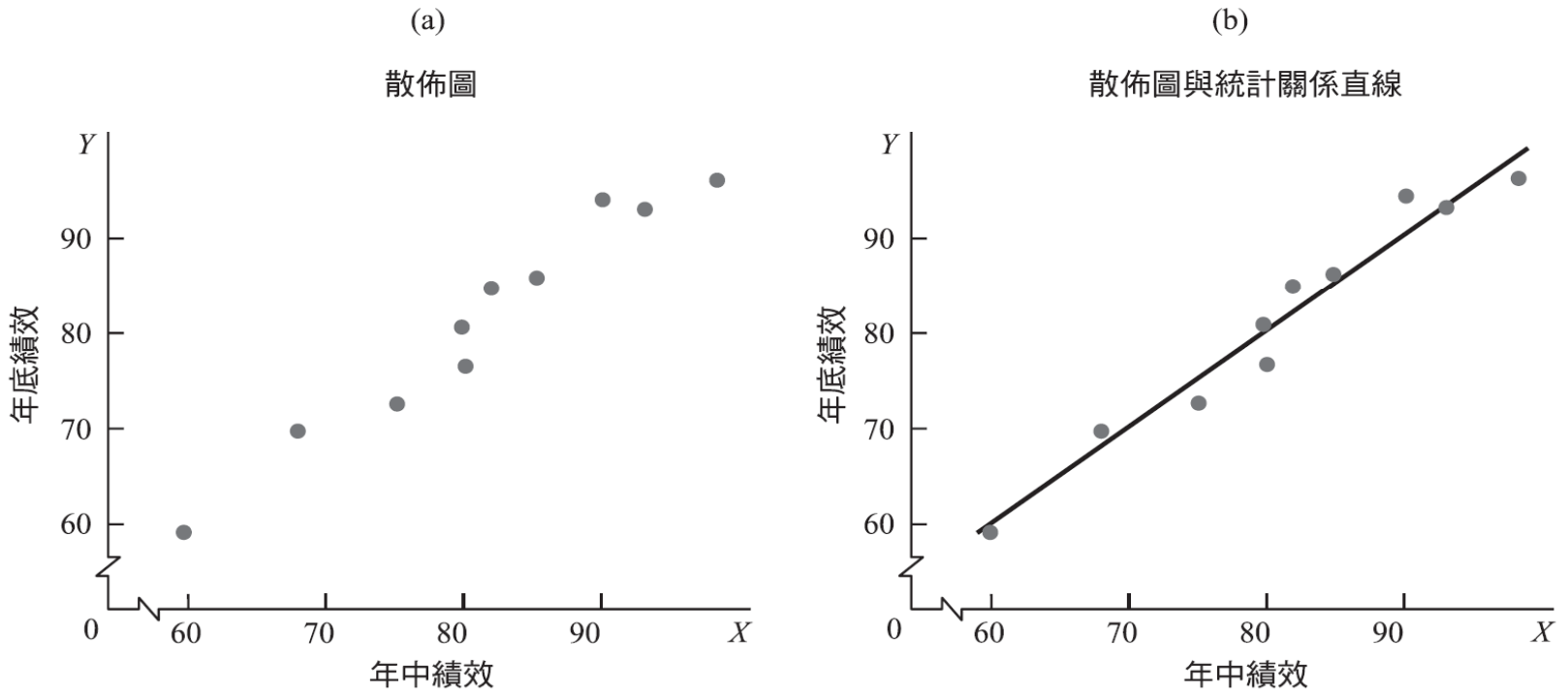
期別	銷售數量	銷售金額
1	75	$150
2	25	50
3	130	260

圖 1.1 中分別將上述三個觀測值畫出，同時需注意的是，在函數關係的特性中，觀測值應該會出現在代表函數關係的直線上，而所有的函數關係均具有此種特性。

- **兩變數間之統計關係**

一般而言，具備統計關係的觀察值並不會完美地出現在關係線上。

圖 1.2
年中績效與年底績效之統計關係。



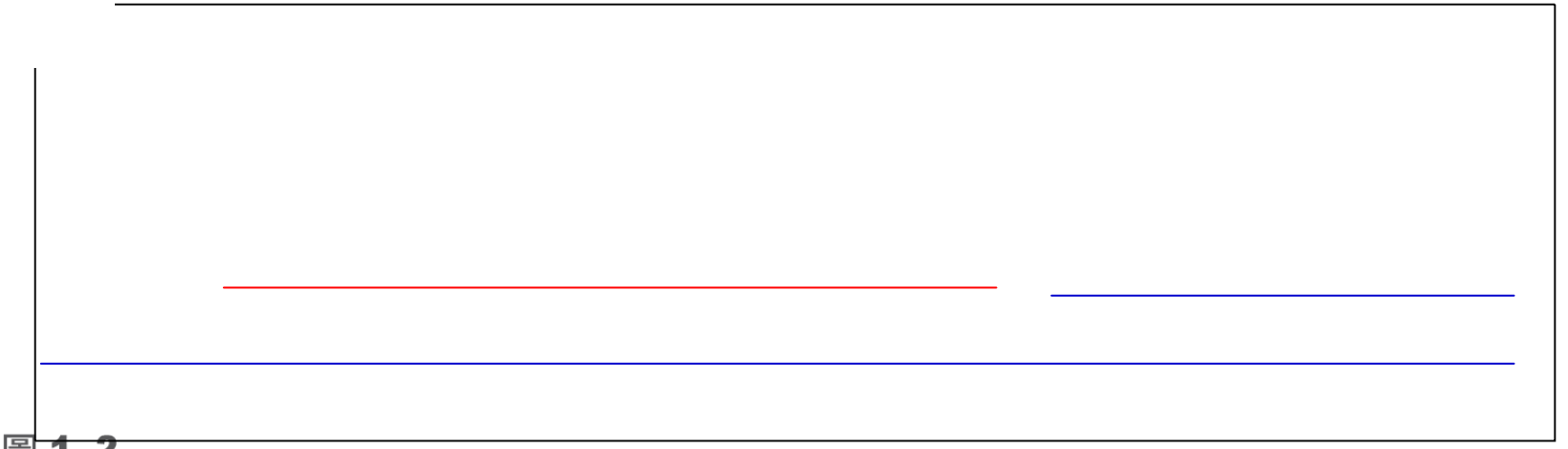


圖 1.2
年中績效與年底績效之統計關係。

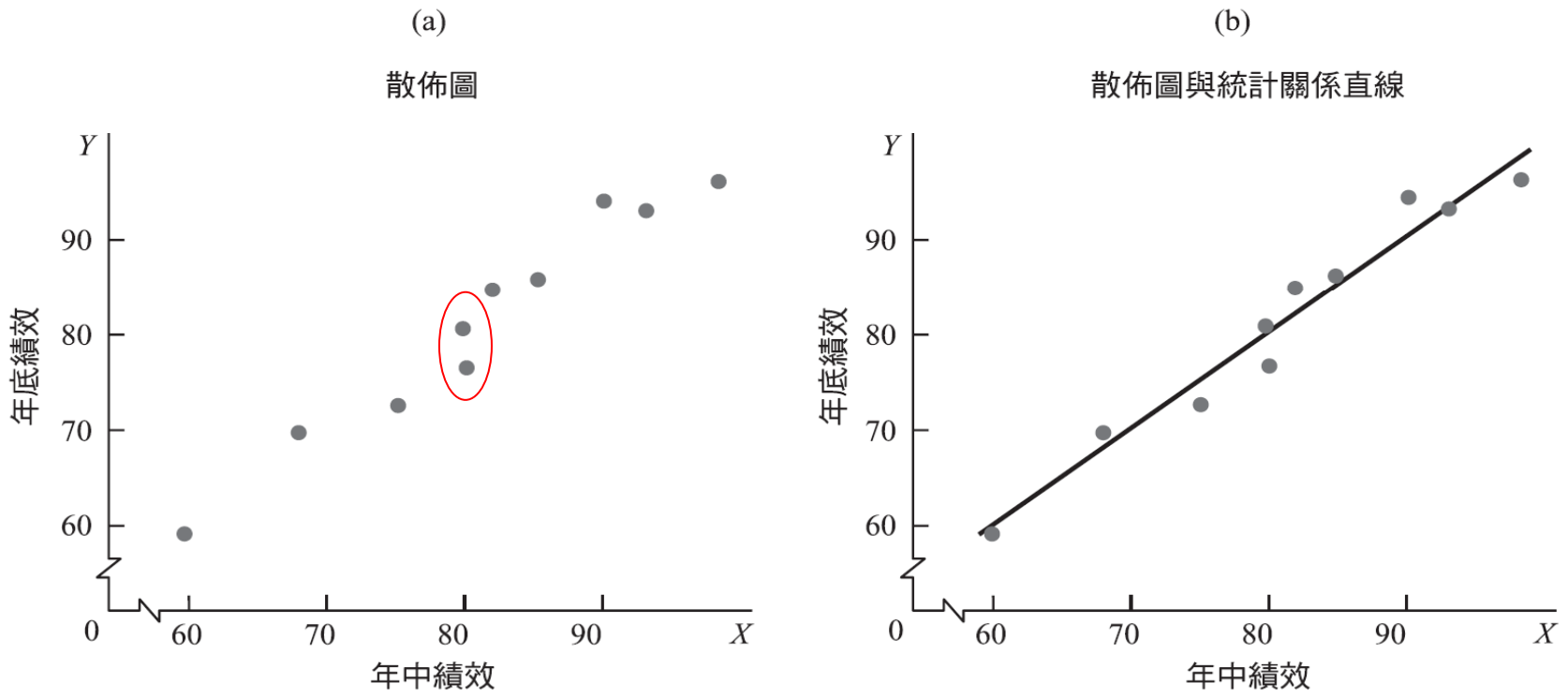
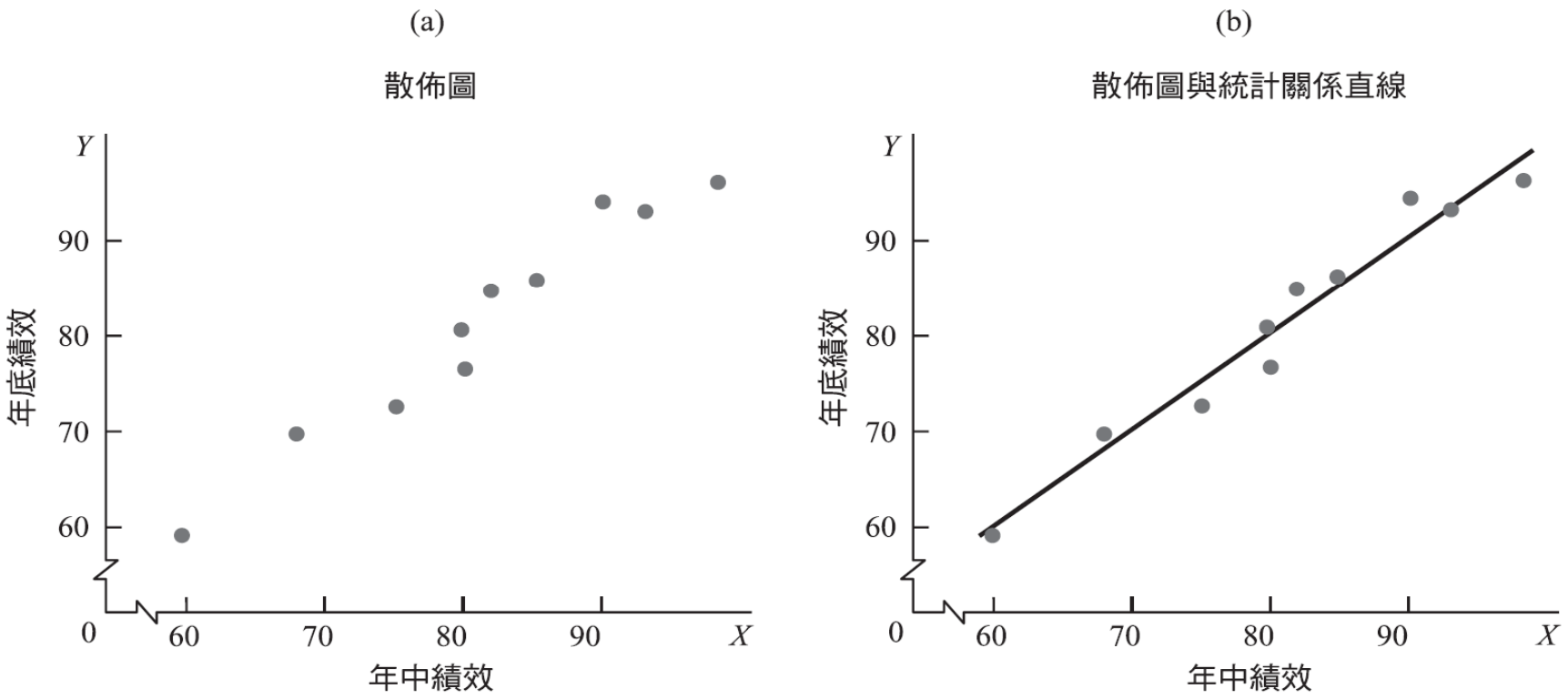


圖 1.2
年中績效與年底績效之統計關係。



在圖1.2b中，我們多畫出了一條代表年中績效與年底績效統計關係的直線，該直線所表現出的大體趨勢為年底績效受到年中績效之高低而有所改變，需注意的是，絕大多數的點並未剛好落在此統計關係之直線上，事實上圖上的點散佈於直線附近，代表了年中績效與年底績效之間另外具有某些不相關的成分存在，此種成分被視為隨機現象，雖然統計關係不如函數關係如此完美，但它仍具有高度的實用性。

例題 2

圖 1.3 的資料是來自 27 位年齡介於 8 至 25 歲的健康女性其血液中的類固醇含量，圖中明確顯示年齡與類固醇含量在統計上之關係為曲線（非直線），在圖 1.3 中之曲線代表了隨著年齡的增加，類固醇含量亦隨之增加，直到某一高點後，類固醇含量便隨著年齡的增加而降低了。同前一例子，絕大多數的點散佈在此統計關係之曲線附近，而這正是典型的統計關係圖形。

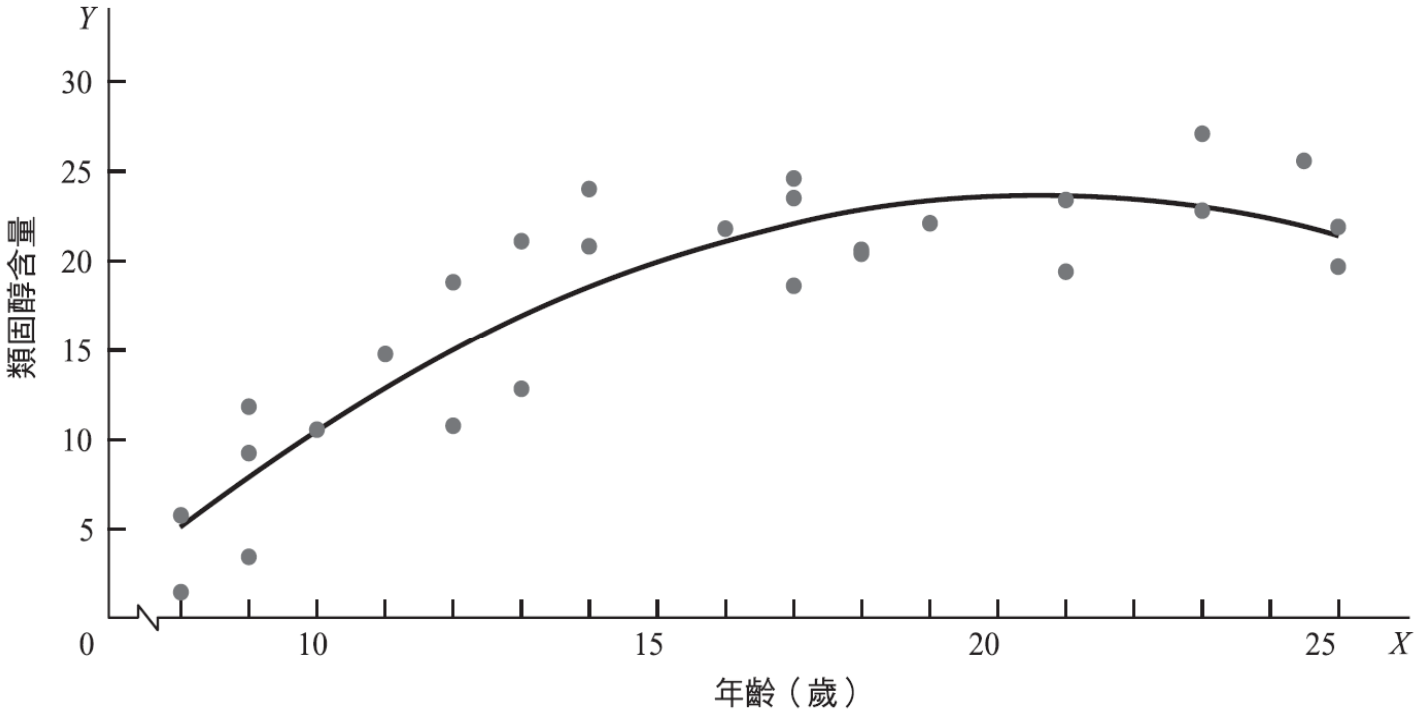


圖 1.3
8 至 25 歲的健康女性其年齡與類固醇含量之統計曲線關係。

1.2 迴歸模型與用途
（Regression Models and Their Uses）

- **起源**
- Francis Galton 先生於十九世紀末，用於研究父母身高與子女身高之相關性，發現子女身高有「迴歸於群體之平均水準之傾向」
- **基本概念**

■ 迴歸模型表示下料兩種統計關聯

- 反應變數Y受到不同的預測變數X之影響，所表現的系統性規律
- 資料散佈點於統計關聯線附近的情形

■ 透過下列之假設可將迴歸模型具體化

- 對於不同的 X，Y 具有其特定之機率分配
- 具有特定之機率分配之 Y，其平均數將隨 X 表現出系統性之規律

迴歸曲線來表示此種迴歸函數。此外，圖 1.4 中的迴歸函數並不是一條直線，而是帶有些微弧度的曲線，這隱含了年底績效*Y*之平均水準，將隨著年中績效*X*之提高，所增加之幅度較為平緩。

- 具有多個預測變數下的迴歸模型

  迴歸模型可能超過一個預測變數，例如

- 一個消費性金融環境中，對 67 家分支機構所進行的調查研究；反應變數 Y 為全年之營運成本，而影響 Y 的預測變數 X 可能有四個：每筆貸款之平均金額 X1、貸款之平均件數 X2、處理中之新貸款總件數 X3及辦公室員工支薪資指數 X4。
- 農場曳引機購買案；反應變數Y為曳引機總購買量（單位馬力），預測變數X則多達有九個，例如曳引機之平均使用年齡、農場數目、農作物產量指數、、、、、。

- 當迴歸模型包含超過一個預測變數時，圖1.4 之模型必須透過更高維度之空間表示，如 X1 及 X2，將同時影響反應變數 Y 之機率分配與平均值，此關係可以一個迴歸曲面來表示。

迴歸模型架構

- **預測變數的選取**

✓ 考慮有限的幾個預測變數來建立所需的迴歸模型

✓ 權衡取捨哪些預測變數值得加入模型中

✓ 增加的預測變數與反應變數是否有因果關係

- **迴歸關係的函數形式**

✓ 透過相關理論找到適當的函數形式

✓ 一次線性或二次平方迴歸函數常是初步的迴歸模型。

- **模型的範圍**

✓ 適用範圍或區段決定於研究調查之設計或現有資料限制（不可外插）

迴歸分析用途

- （1）敘述（2）控制（3）預測
- 用途相互重疊

迴歸與因果關係

- ✓ 反應變數Y與預測變數X之間存在統計關係，並不意味Y與X必然存在有因果關係，如學童識字數量X與書寫速度Y，可能存在其他變數如學童年齡或受教育量。
- ✓ 即便存在強烈統計關係推論出當中因果關係，此因果關係卻未必是X影響Y，如進行溫度計的校正時，利用迴歸關係來評估溫度計的讀數X與實際溫度Y，事實上是Y影響預測變數X。
- ✓ 迴歸分析不能提供任何有關因果型態之訊息，必須輔以其他分析條件來完成因果關係條件之建立

電腦操作

- ✓ 由於常伴隨許多冗長而繁雜的計算，藉助電腦計算能力完成迴歸分析，大多數的統計套裝軟體均提供迴歸程序，MINTAB，SAS，SPSS，MATLAB（EXCEL）

1.3 誤差分配未知之簡單線性迴歸模型

- **模型的標準敘述**
- 最基本的單一預測變數之簡單線性迴歸模型

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1.1)$$

上式 Y_i 為反應變數在第i次實驗下之結果，

β_0與 β_1均為參數

X_i為已知之常數，代表了預測變數在第 i 次實驗時之值

ε_i為隨機誤差項，期望值$E\{\varepsilon_i\}=0$，且變異數$\sigma^2\{\varepsilon_i\}=\sigma^2$，$\varepsilon_i$與 ε_j 無相關性，$i = 1,2,\ldots,n$。

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1.1)$$

- ❑ 上式（1.1）稱為單一、參數線性又預測變數亦為線性之迴歸模型
  - ✓ 僅存在一個預測變數，『單一』
  - ✓ 因模型中參數不具備有指數、相乘項或互除項之型態，所以稱為『參數線性』
  - ✓ 預測變數之型態為一次式，『預測變數為線性』
  - ✓ 當迴歸模型的參數線性且預測變數亦為線性時，此迴歸模型「一階模型」
- ❑ **模型的重要特性**

1. 在第 i 次實驗下之反應量 Y_i 是由兩個部分相加所共同組成：1.常數項 ： $\beta_0 + \beta_1 X_i$ 2.隨機項： ε_i，故 Y_i 也是一個隨機變數

2. 由於 E{ε_i}=0 ，當 X 在第 i 次實驗時之值 X_i 時，反應量 Y_i 來自平均數 (1.2)

之機率分配，因此模型(1.1)之迴歸函數為

(1.3)

3. 當第 i 次實驗下反應量Y_i與迴歸函數兩者之差 ε_i 為誤差項

4. 誤差項被假設為具有常數變異數 σ^2 ，故反應量 Y_i 也具有相同的常數變異數 (1.4)

透過式子 (A.16a)：

5.模型假設誤差項之間無相關性，亦即， ε_i 與 ε_j 無相關性，故任何一次的實驗結果 Y_i 均不會影響實驗結果 Y_j

6. 對於(1.1)迴歸模式，對不同的 X 均隱含了假設反應量 Y_j 來自平均數 與變異數 σ^2 之機率分配，且實驗結果Y_i與實驗結果Y_j彼此間無相關

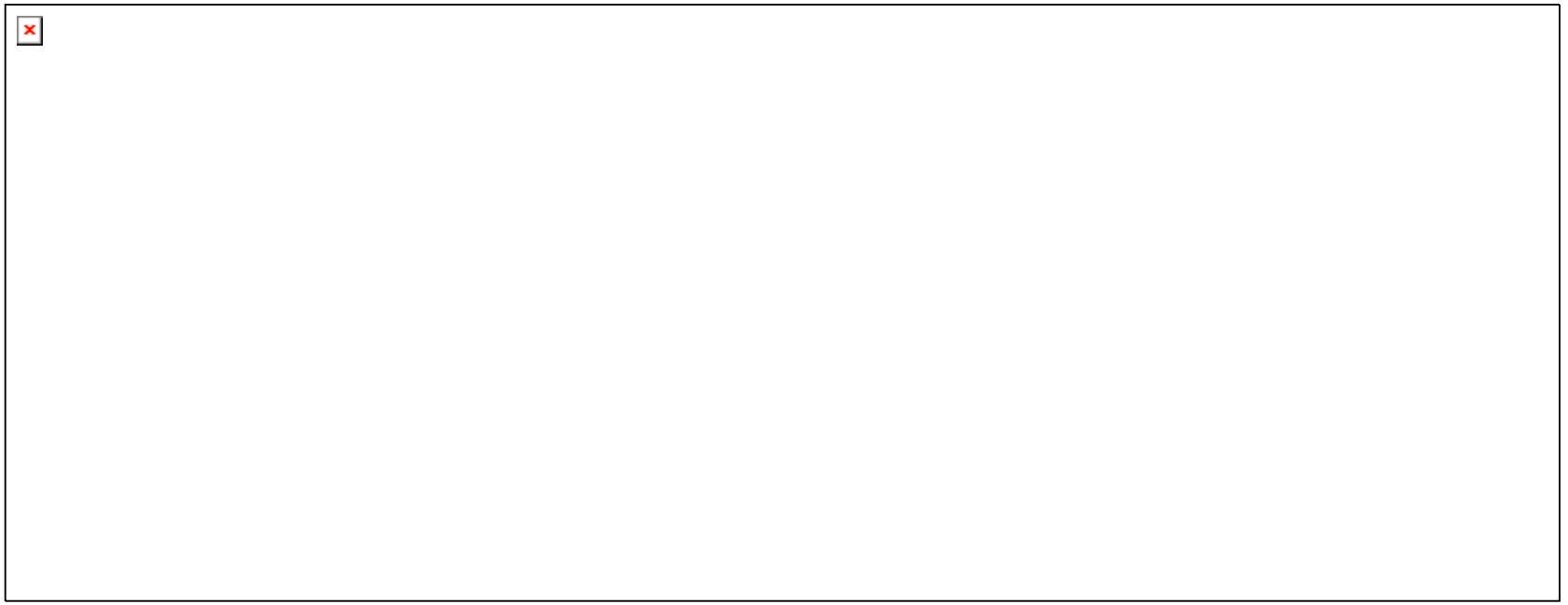
例題

配電公司顧問想研究有關營造承包商在一週內申請照明用電之件數與處理申請所需時間，兩者間所存在之關係，假設迴歸模型(1.1)適用，並有如下之結果：

$$Y_i = 9.5 + 2.1X_i + \varepsilon_i$$

其中，X 表示一週內申請照明用電之件數，而 Y 表示處理申請所需時間（小時），在圖 1.6 中畫出了如下之迴歸函數：

$$E\{Y\} = 9.5 + 2.1X$$



假設在第 i 週中，$X_i = 45$ 下之實際 $Y_i = 108$，在本例中誤差項 $\varepsilon_i = 4$，於是我們有

$$E\{Y_i\} = 9.5 + 2.1(45) = 104$$

且

$$Y_i = 108 = 104 + 4$$

圖 1.6 中除了標示出 $X = 45$ 下 Y 之機率分配，同時也從該分配中取出觀測值 $Y_i = 108$，而誤差項 ε_i 可以視為是 Y_i 與其平均數 $E\{Y_i\}$ 之差距。

另外在圖 1.6 中也標示出了 $X = 25$ 下 Y 之機率分配，在圖中可以看出此分配之變異與 $X = 45$ 下 Y 之機率分配相同，這一點與迴歸模型(1.1)之假設要求相符合。

迴歸參數之意義

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1.1)$$

- 上式迴歸模式(1.1)中參數β_1及β_0稱為迴歸係數(regression coefficient)，β_1表示此迴歸直線的斜率，表示 X 每增加一單位，Y 之機率分配的期望值所改變之量，β_0表示此迴歸直線通過Y軸之截距。
- 如果模型範圍不包含 X=0，則截距β_0本身並不具備任何特殊意義

有關配電公司之舉例，在圖 1.7 中之迴歸函數：

$$E\{Y\} = 9.5 + 2.1X$$

此迴歸直線之斜率 $\beta_1 = 2.1$ 表示一週內每增加一個申請件數，所增加的處理時數其機率分配之期望值為 2.1 小時。

另外，截距 $\beta_0 = 9.5$ 是當 $X = 0$ 時，迴歸函數所反映之值，但是本例中一週內之申請件數介於 20 至 80 之間，此一線性迴歸模型之範圍並不包含 $X = 0$，因此 β_0 本身並不具備實質意義。

❑ 不同版本的迴歸模型

- 令 [ ] 表示恆為1之虛擬變數，則迴歸模型(1.1)可以用如下之表示方式：

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i \quad \text{當} X_0 \equiv 1 \qquad (1.5)$$

此種形式之模型將讓每一個迴歸係數均與 X 相結合。

- 另一種迴歸模型的形式是利用預測變數之離差代替 X_i，則迴歸模型(1.1)可以表示如下：

[ ]

- 所以此一版本之迴歸模型可以表示為：

$$Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \varepsilon_i \qquad (1.6)$$

其中，[ ] (1.6a)

1.4 迴歸分析資料

正常情況下迴歸模式(1.1)中迴歸係數β_1及β_0之值，必須經由資料的分析來發展適當的迴歸模式。迴歸分析之資料區分為非實驗性資料及實驗資料

- **觀察資料**

  非實驗性資料，主要限制在於此類資料無法提供合適的因果關係訊息，如員工之年齡與請病假的天數成正比的關係，其實未必可直接推論出兩者是年齡康健之因果關係，可能年輕員工為內勤人員，而高齡員工為外勤人員或工作地點

- **實驗資料**

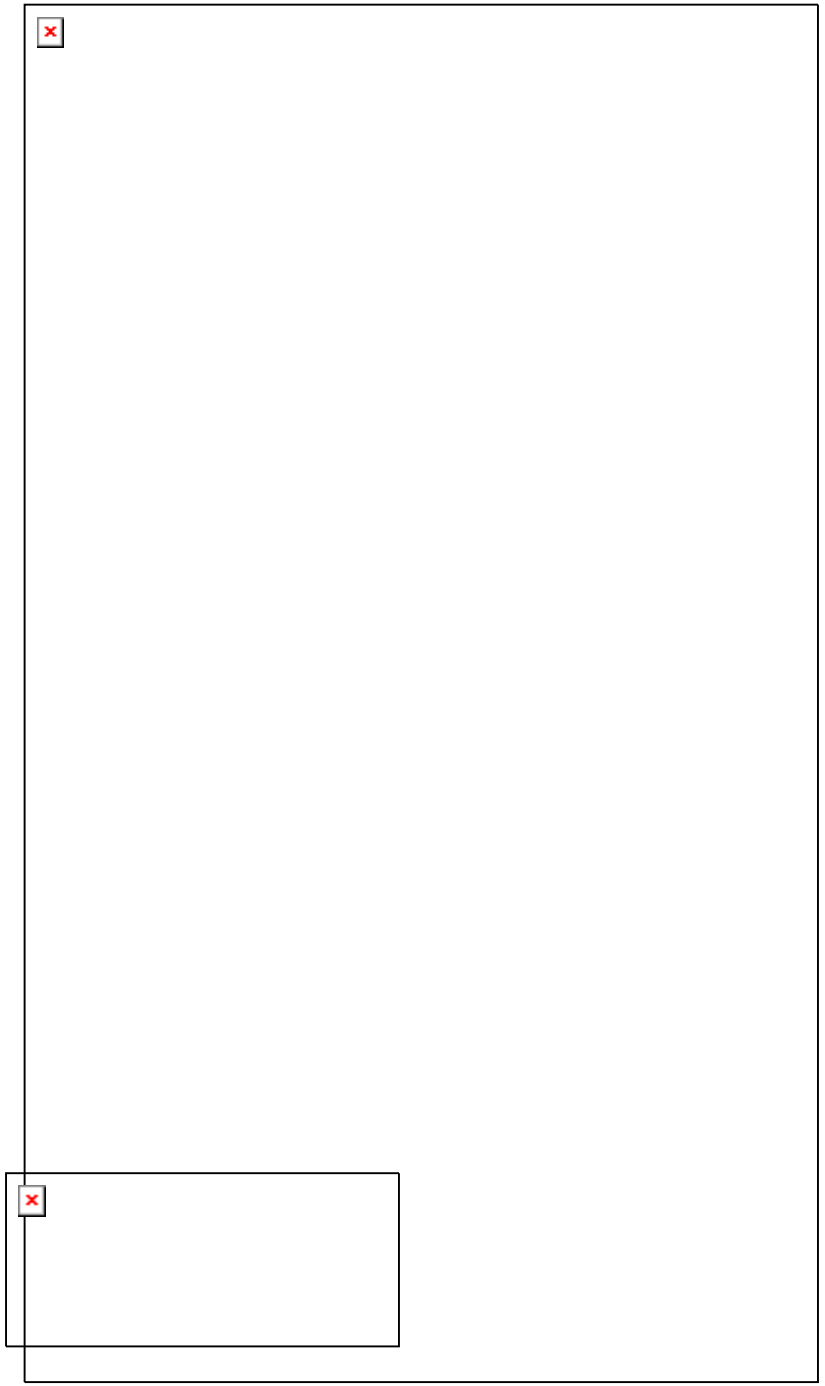
  採用可控制的實驗來收集資料，以提供迴歸參數估計用，由於透過解釋變數進行隨機指派，以取得迴歸分析所需之資料，這樣的實驗結果具有較強烈的因果關係訊息，其理由是因為透過隨機化的安排，會使得可能影響反應變數的其他可能變數，其影響效果將趨於平衡

完全隨機化設計

- 控制解釋變數如 X 大小，在實驗設計中稱為處理(treatment)，而被列為研究的對象稱為實驗單位(experiment unit)，控制解釋變數的方法則是透過隨機指派給每一實驗單位
- 透過隨機指派處理給每一個實驗單位是一種最基本的統計方式稱為完全隨機化處理(completely randomized design)，是指所有處理指派工作給實驗單位的各種不同的組合，發生的機率都一樣大
- 完全隨機化處理可適用於多種不同處理時，進行樣本大小不一樣之實驗。

1.5 迴歸分析步驟概述

- 往後所提之迴歸模式均適用於觀察資料與完全隨機化設計下的實驗資料
  - 基本要求：滿足迴歸模型所要求的基本假設
- 迴歸分析步驟概述
  - 模型的推論為迴歸分析的第一步，一般情形其實無法事先確知適當的迴歸模型，故對資料進行探索性分析是首要工作，如圖 1.8



1.6 迴歸函數估計

- 可以透過觀察資料或實驗資料，這些資料是透過解釋變數或預測變數X，以及與其對應的反應變數Y所構成的。
- 每一次的實驗結果均將產生一個 X 與一個 Y，以符號(X_1,Y_1)表示(X,Y)在第一次實驗下的值，而符號(X_i, Y_i)表示(X,Y)在第 i 次實驗下的值，其中 $i = 1,\ldots,n$ 。

例題

在一個小規模的耐性實驗中，實驗人員提供一項非常困難之任務給三個研究對象，資料中有關研究對象之年齡(*X*)與研究對象在放棄該任務前所嘗試之次數(*Y*)如下表所列：

最小平方法 (method of least squares)

為了求得迴歸參數 β_1 及 β_0 之估計量，是透過最小平方法(method of least squares)

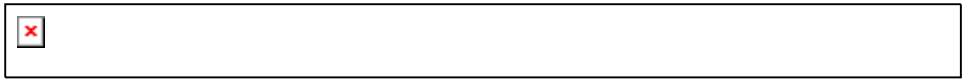
- 對於每一觀察值(X_i, Y_i)，最小平方法考慮了 Y_i 與本身期望值之離差：

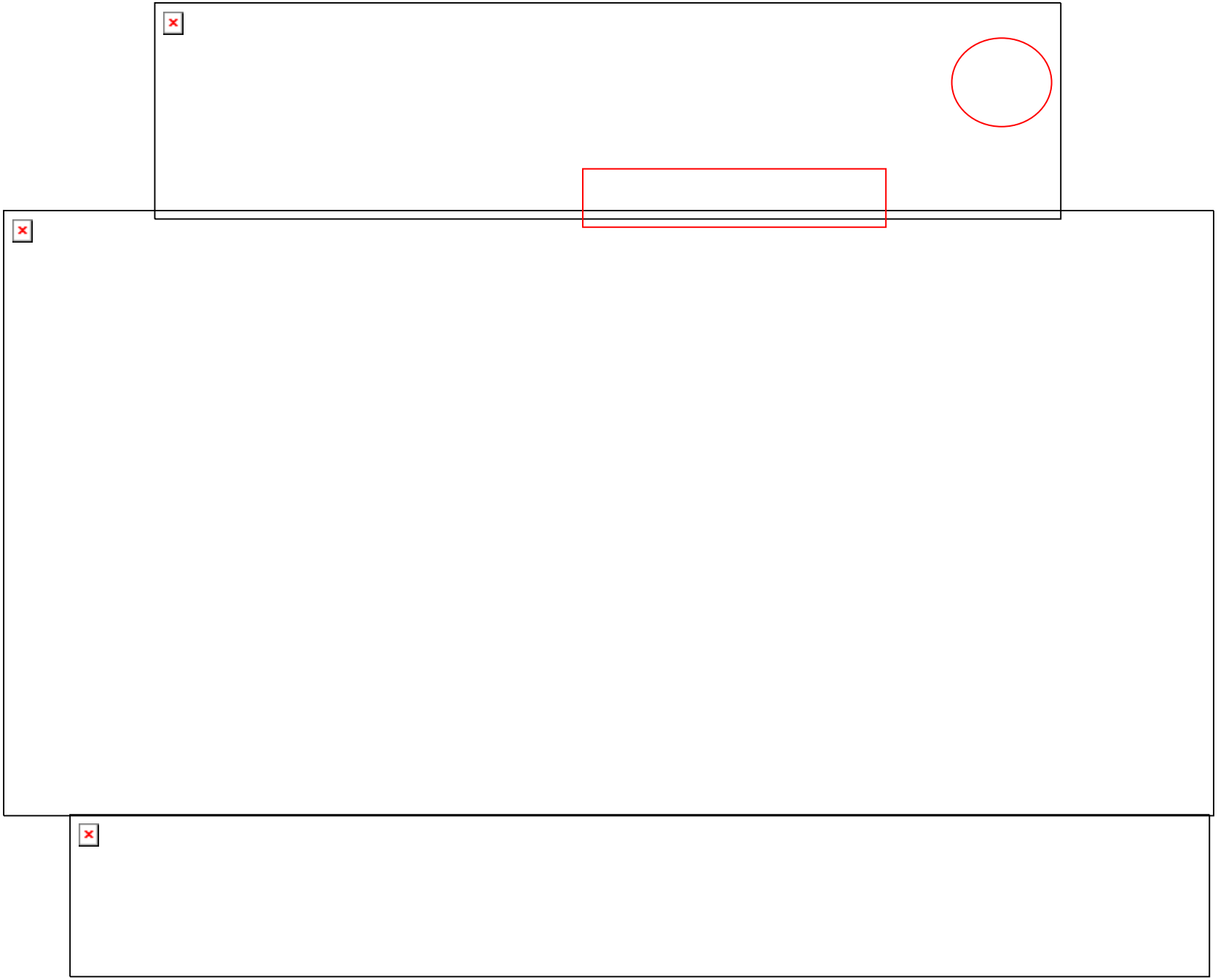
$$Y_i - (\beta_0 + \beta_1 X_i) \qquad (1.7)$$

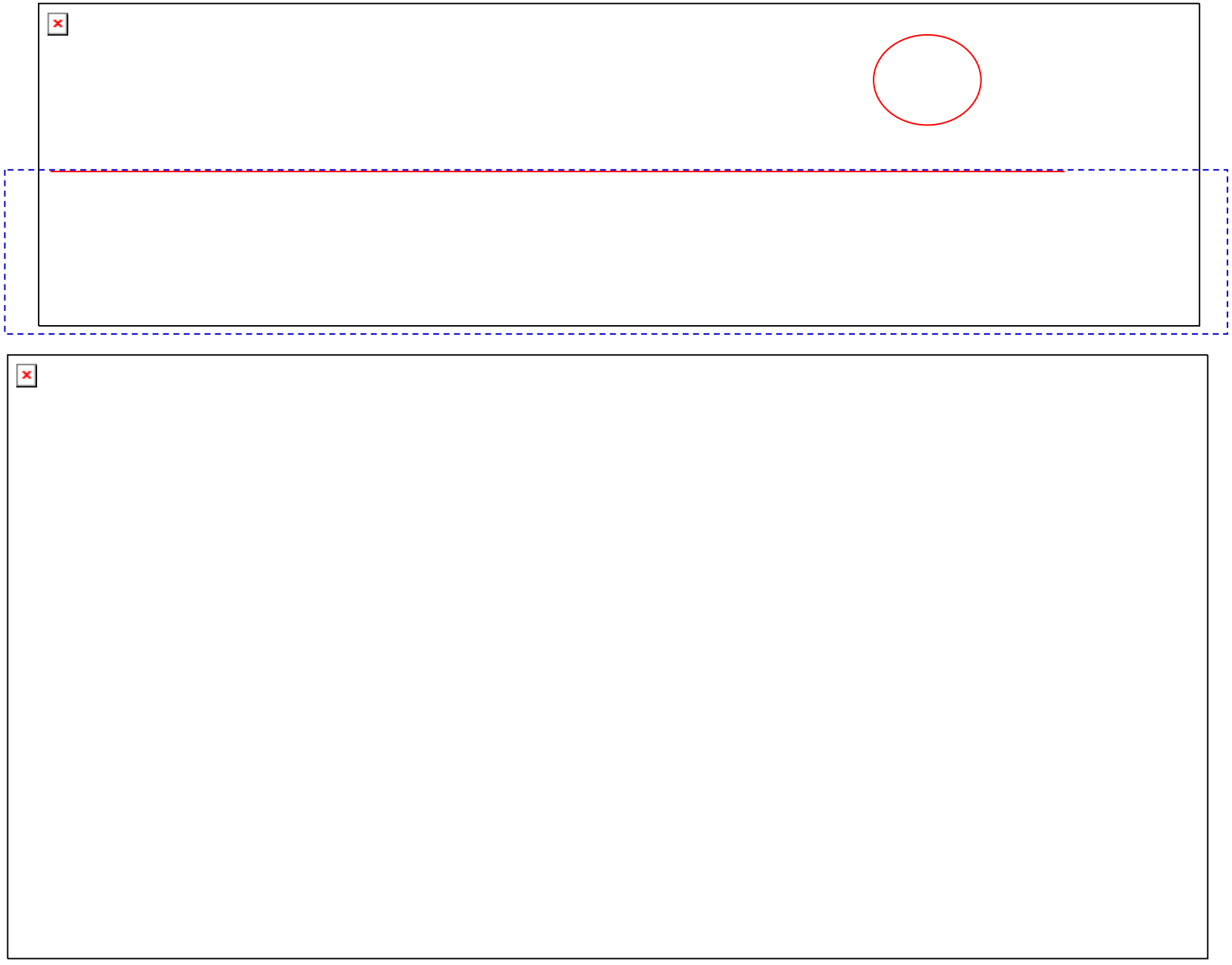
- 而最小平方法便是將上述 n 個離差平方後取總合，用符號 Q 來表示：

(1.8)

- 根據最小平方法之原理，計算出 β_1及β_0 之的估計量為 b_0 與 b_1， 計算 Q 為最小之數







最小平方估計量

- 採用**解析的方法**找出滿足迴歸模型(1.1)，可以證明出最小的 Q 值所對應之 b_0 與 b_1 需同時滿足下列聯立方程式：

$$\sum Y_i = nb_0 + b_1 \sum X_i \qquad (1.9a)$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2 \qquad (1.9b)$$

- 由 P1-17 說明可得出聯立方程式 (1.9a) 與 (1.9b) 稱為標準方程式，b_0 與 b_1 分別稱為 β_0 與 β_1 的點估計量。
- 透過標準方程式(1.9)可以同時解出 b_0 與 b_1 如下：

(1.10a)

(1.10b)

$$\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \ldots, n$$

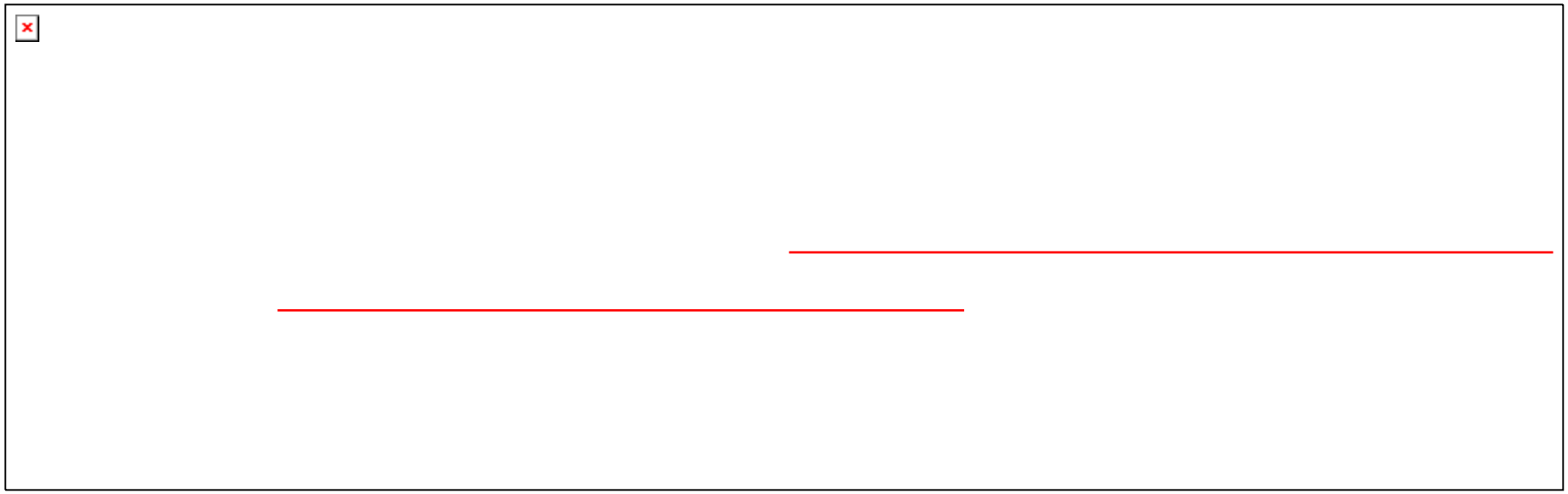
最小平方估計量之性質

■ 有一重要定理稱為「Gauss-Markov定理」，陳述如下：

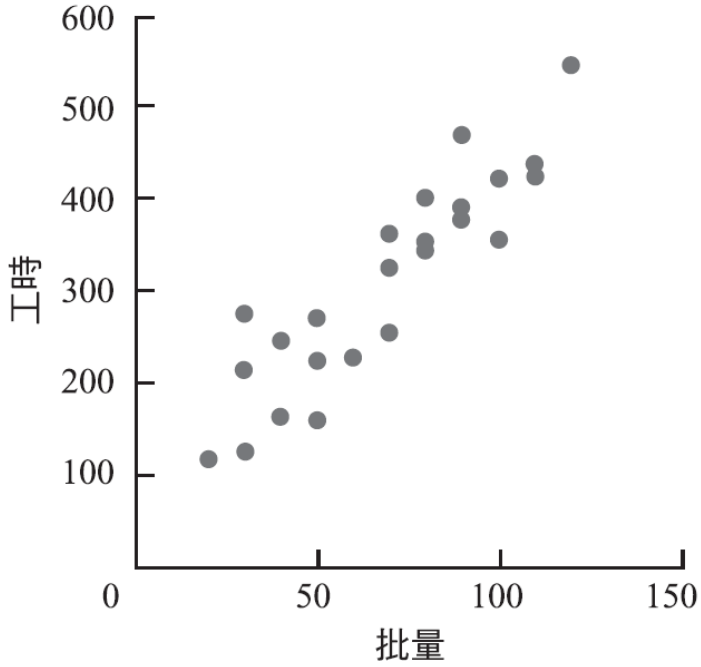
在迴歸模型(1.1) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ 之假設條件下，透過最小平方法得到之估計量 b_0 與 b_1 (1.10)式，是一組**不偏** (*unbiased*) 之估計量，同時在所有不偏線性之估計量中，該組估計量之變異數為最小。 (1.11)

- b_0 與 b_1是一組不偏估計量；$\mathrm{E}\{b_0\} = \beta_0$ 與 $\mathrm{E}\{b_1\} = \beta_1$
- b_0 與 b_1在所有不偏之之估計量，具有最小變異數

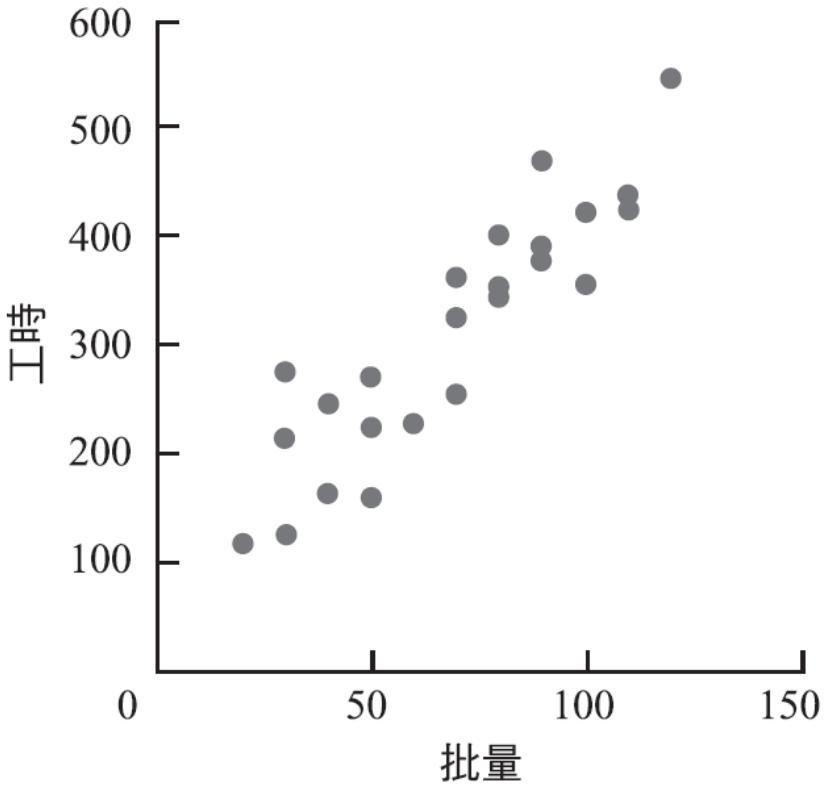
$$\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \ldots, n$$

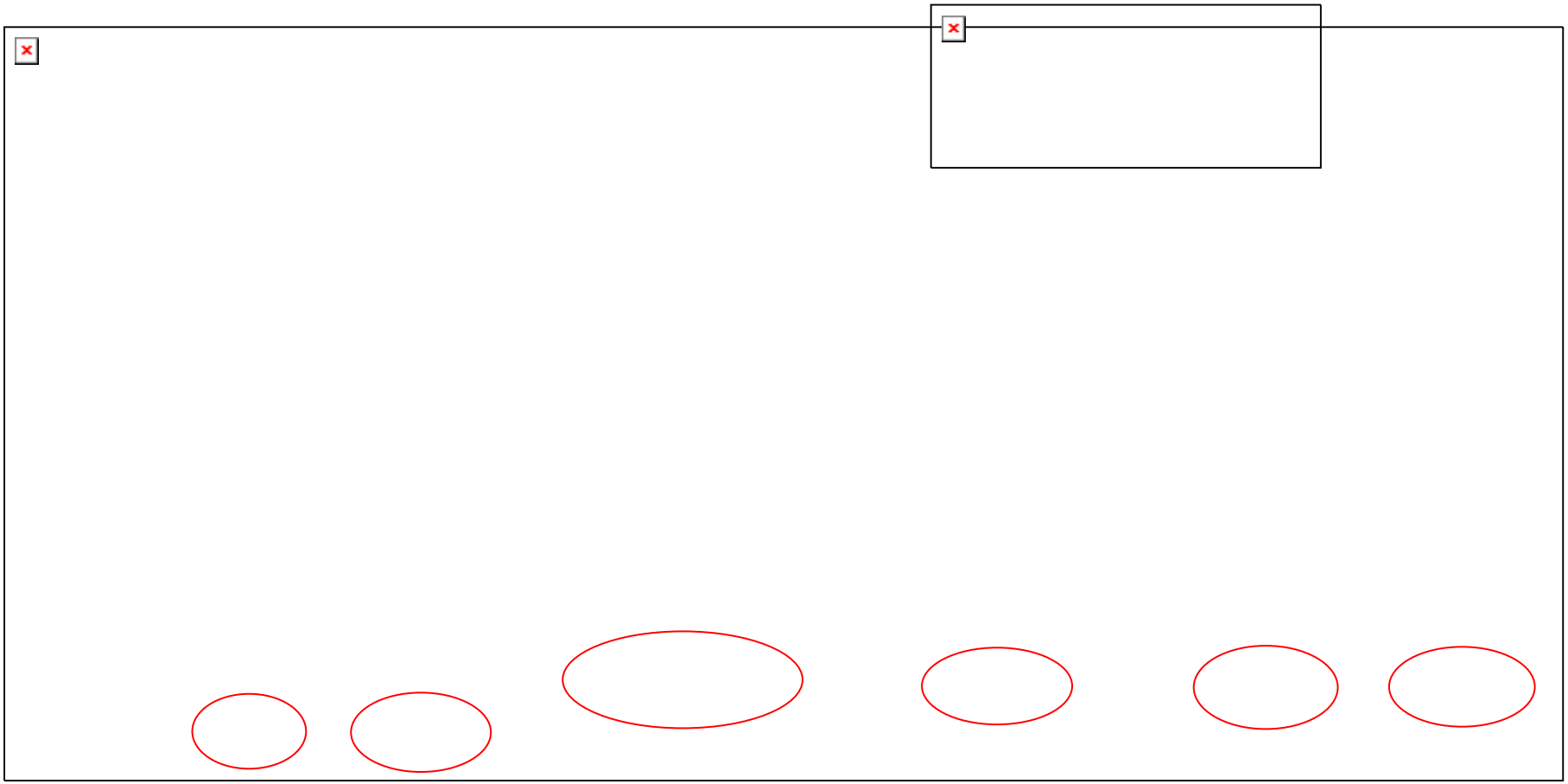


(a) 散佈圖



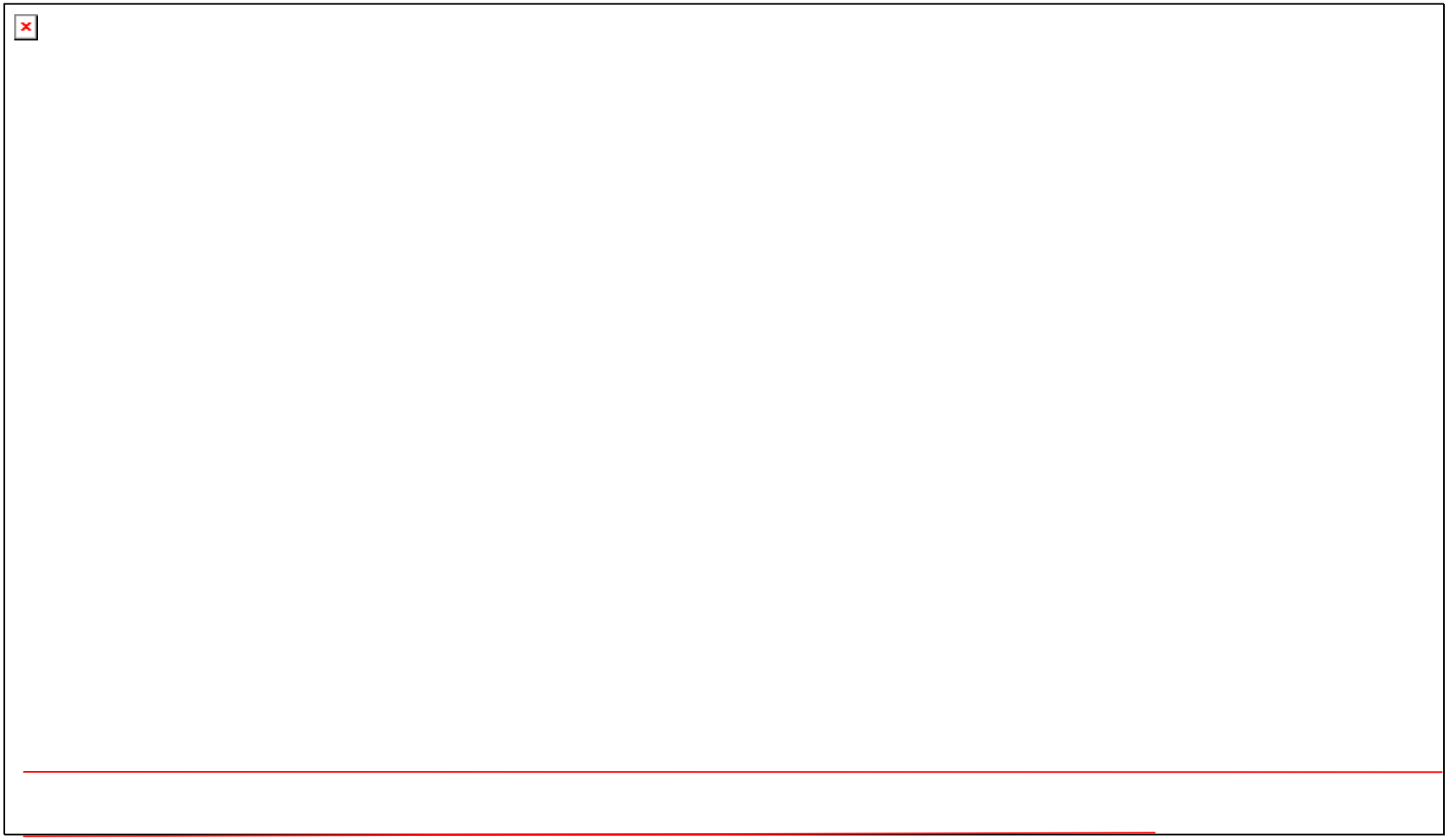
(a) 散佈圖





計算 b_0 與 b_1

- 先計算 X_i-$\overline{X}$ 與 Y_i-$\overline{Y}$這兩個離差數量，結果在表1.1中的第三欄與第四欄
- 兩者之交叉乘積項 $(X_i-\overline{X})(Y_i-\overline{Y})$ 與離差平方 $(X_i-\overline{X})^2$ 分別在表1.1中第五欄與第六欄，第七個欄位計算的離差平方$(Y_i-\overline{Y})^2$



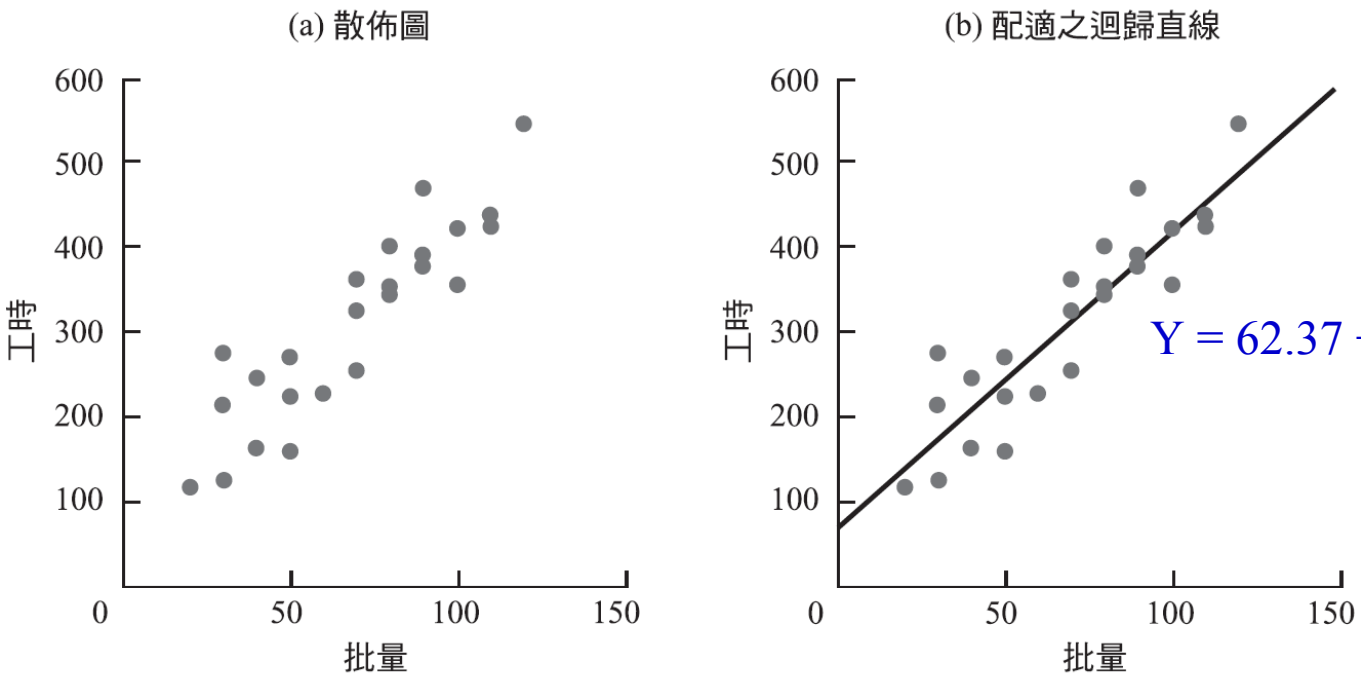


圖 1.10
SYSTAT 所繪製之散佈圖與配適之迴歸直線— Toluca 公司案例。

```
The regression equation is
Y = 62.4 + 3.57 X

Predictor       Coef      Stdev     t-ratio         p
Constant       62.37      26.18        2.38     0.026
X             3.5702     0.3470       10.29     0.000

s = 48.82       R-sq = 82.2%     R-sq(adj) = 81.4%
```

圖 1.11
輸出之部分結果— Toluca 公司案例。

平均反應值之點估計

- **估計迴歸函數**　**迴歸函數** (1.3)：

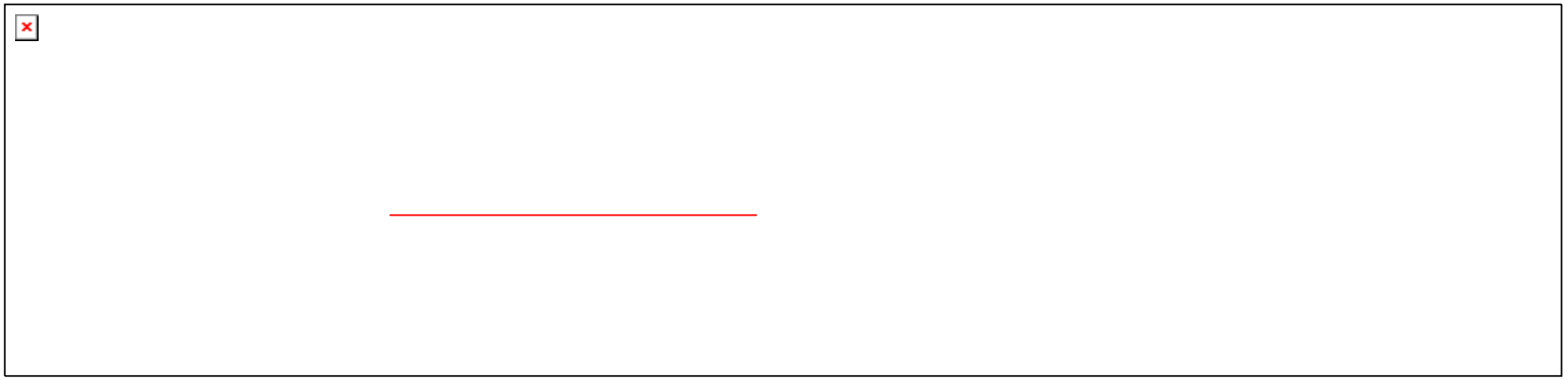
若其參數分別為**樣本估計量** b_0 與 b_1 ，可以下式**估計迴歸函數**

$$\hat{Y} = b_0 + b_1 X \tag{1.12}$$

其中 $\hat{Y}$ (讀成Y-hat) 指當預測變數 X 在該水準下之估計迴歸函數值(平均反應值的點估計)。反應變數所表現之數值大小稱為反應值，E{Y} 稱為平均反應值，(mean response)，平均反應值 E{Y} 就是當預測變數 X 在該水準下 Y 之機率分配的平均數，而 $\hat{Y}$ 是指當預測變數 X 在該水準下 Y 之平均反應的點估計，而 $\hat{Y}_i$ ：

$$\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \ldots, n \tag{1.13}$$

是指第 i 個**配適值**，並且必須要區分出配適值 $\hat{Y}_i$ 與觀測值 Y_i 兩者之不同。



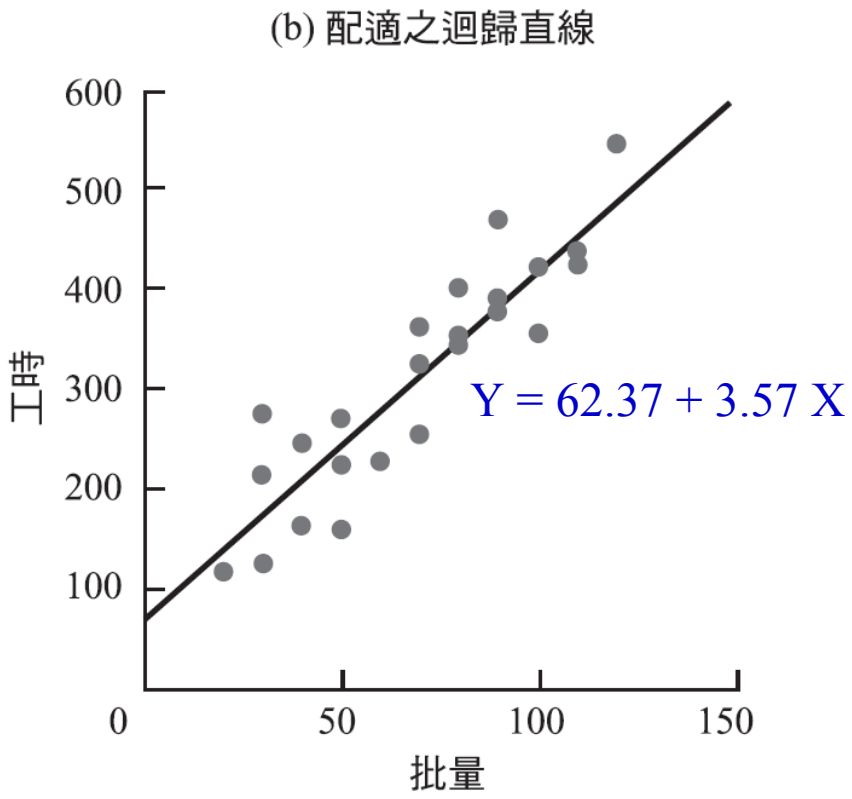
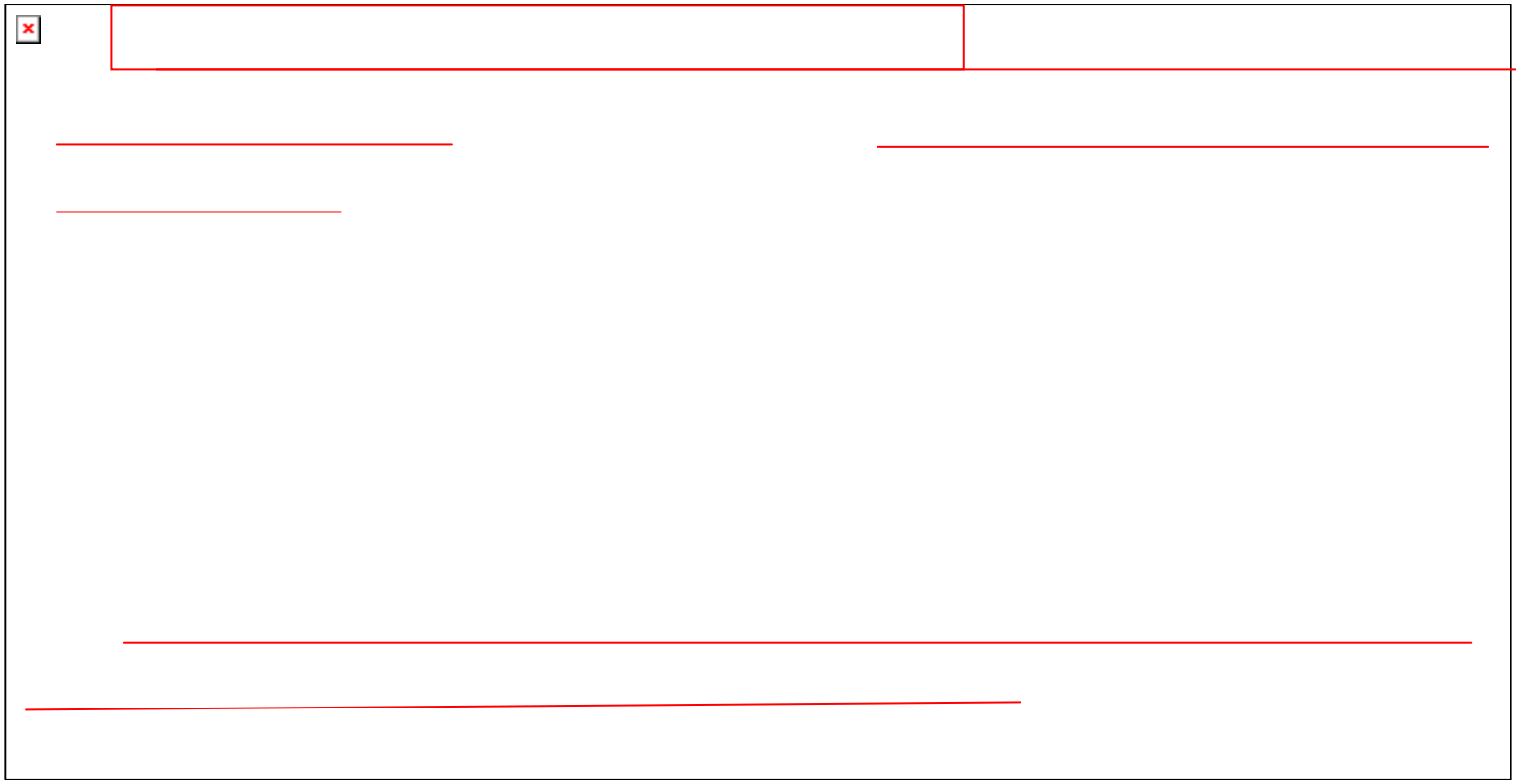
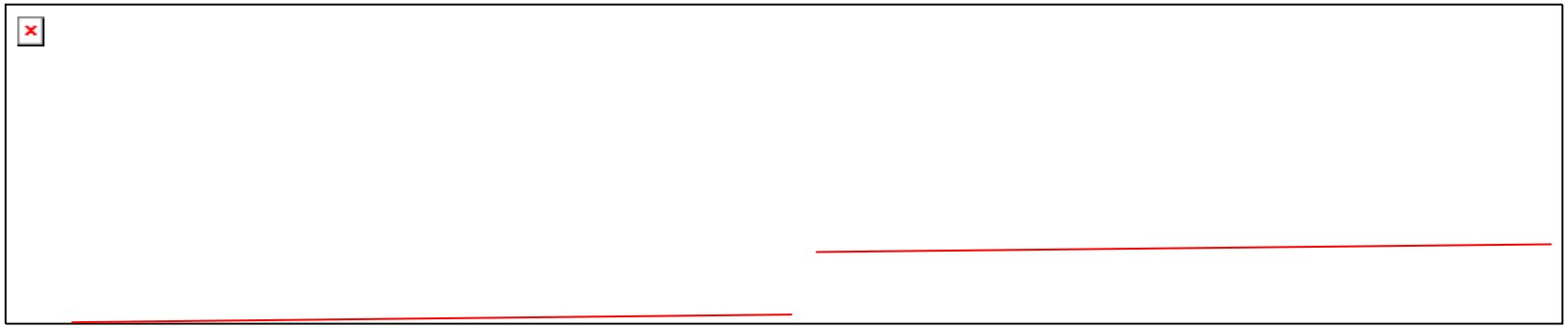
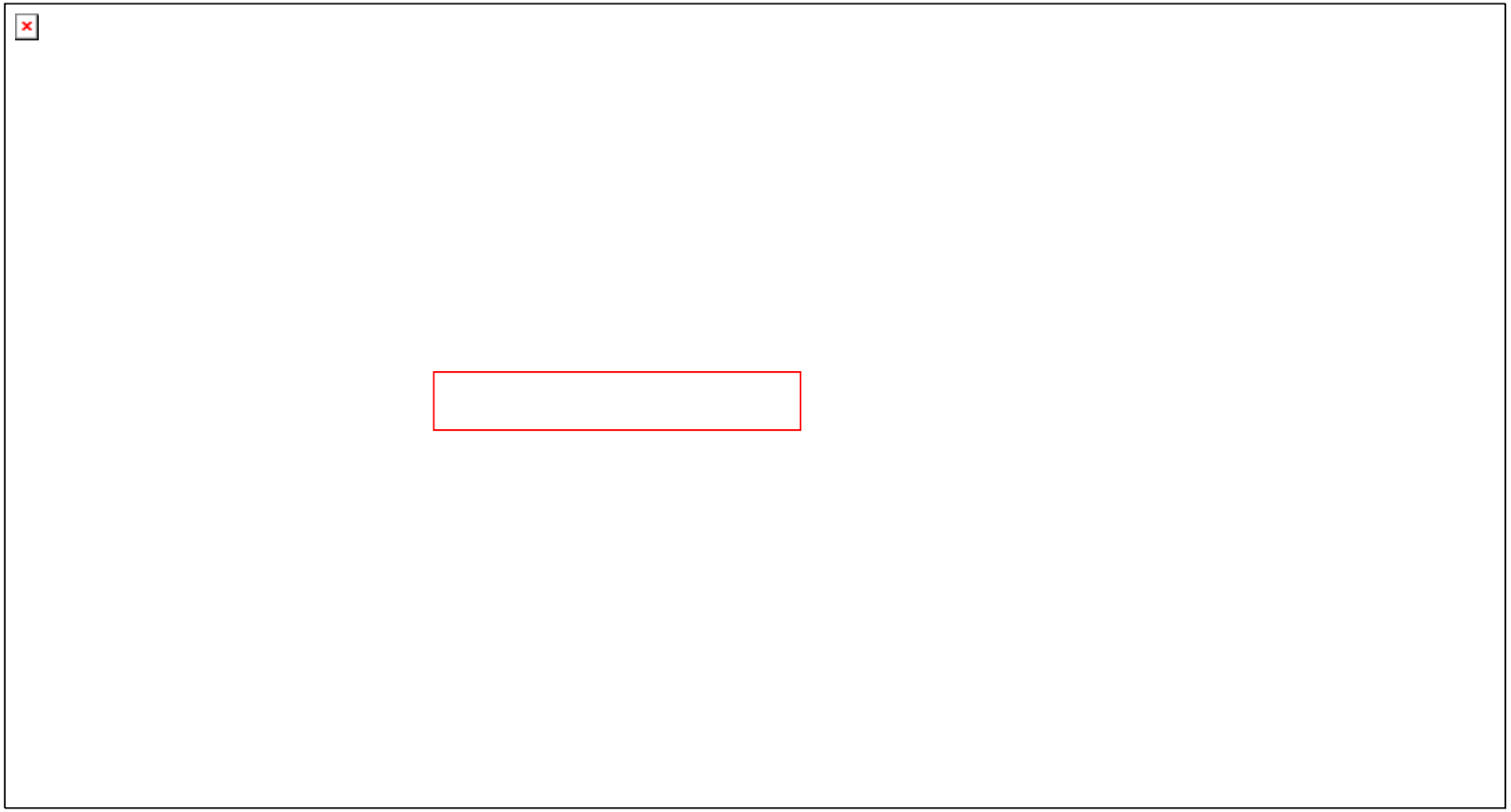


圖 1.10
SYSTAT 所繪製之散佈圖與配適之迴歸直線— Toluca 公司案例。







- **替代模型(1.6)** 如果考慮模型(1.6)：

則 β_1 之最小平方估計量仍然是 b_1 ，而 $\beta_0^* = \beta_0 + \beta_1 \bar{X}$ ，所以根據(1.10b)：

(1.14)

所以**替代模型**(1.6)所估計之迴歸函數為：

(1.15)

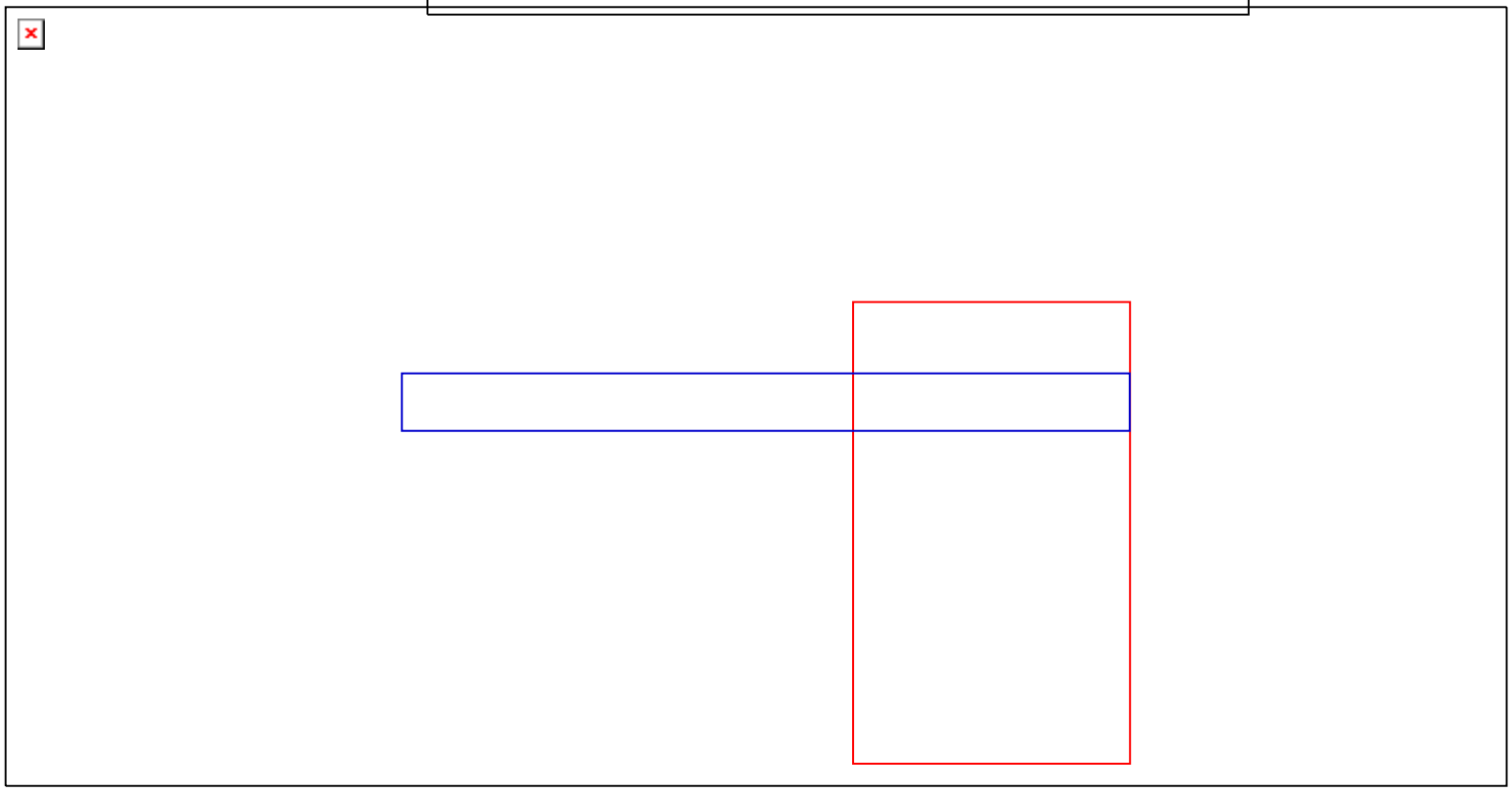
- 殘差

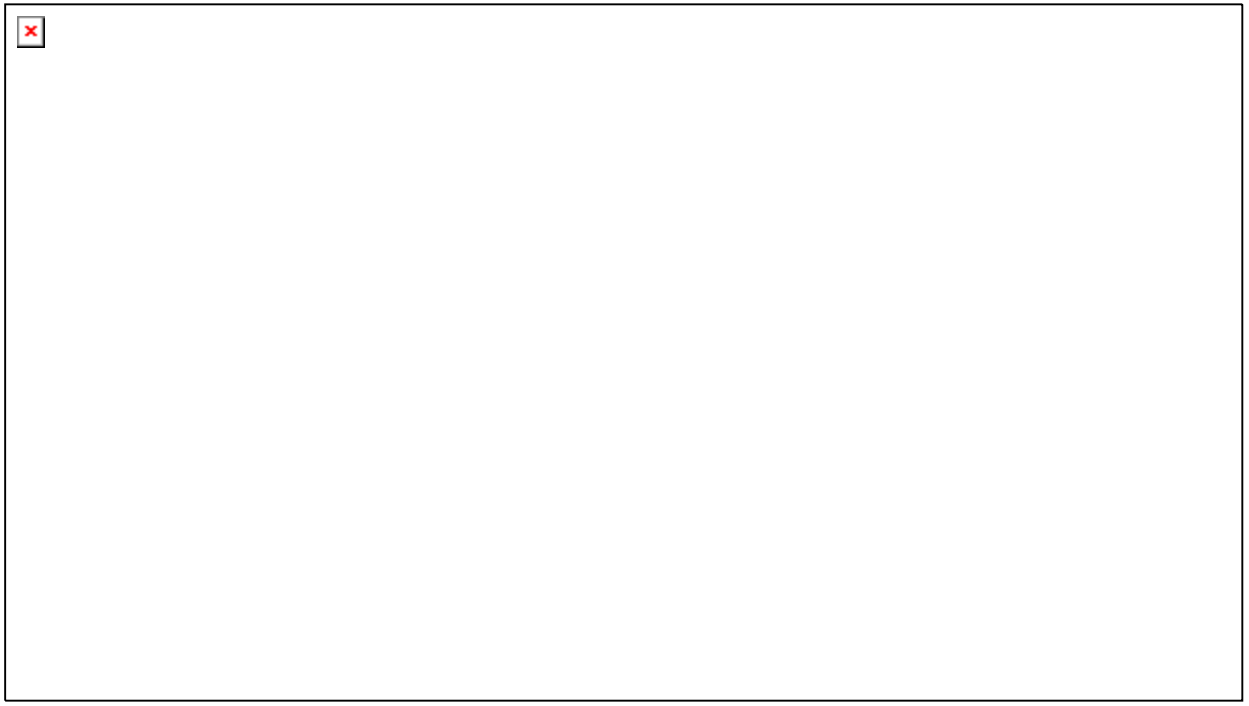
第 i 個**殘差**（residual）是指觀測值 Y_i 與配適值 $\hat{Y}_i$ 之間的差，我們用符號 e_i 表示，並做如下之定義：

(1.16)

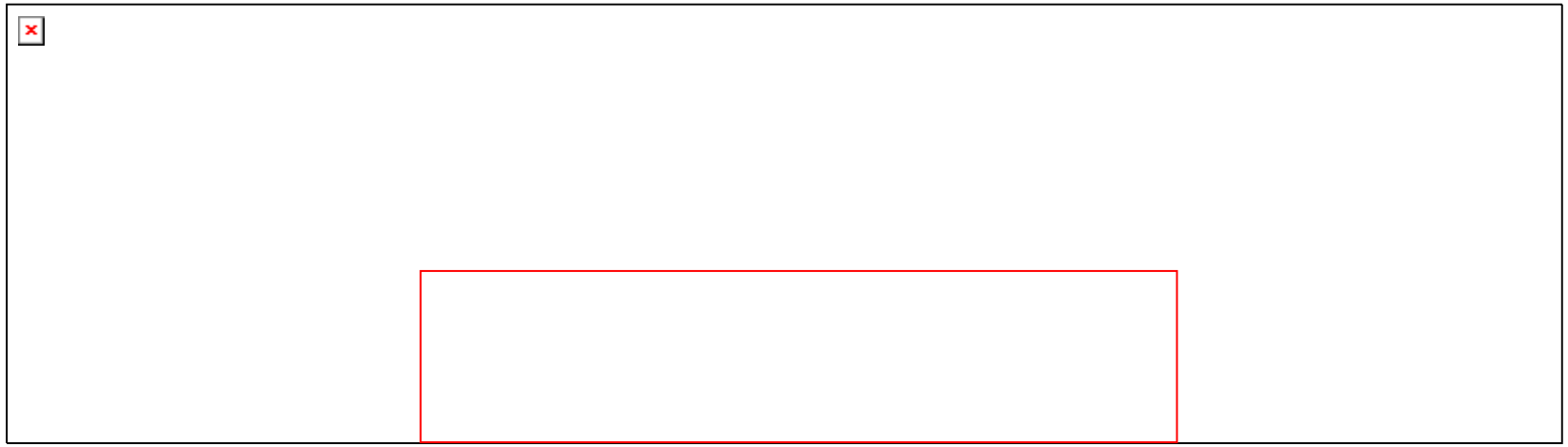
對迴歸模型(1.1)，殘差e_i變成：

(1.16a)

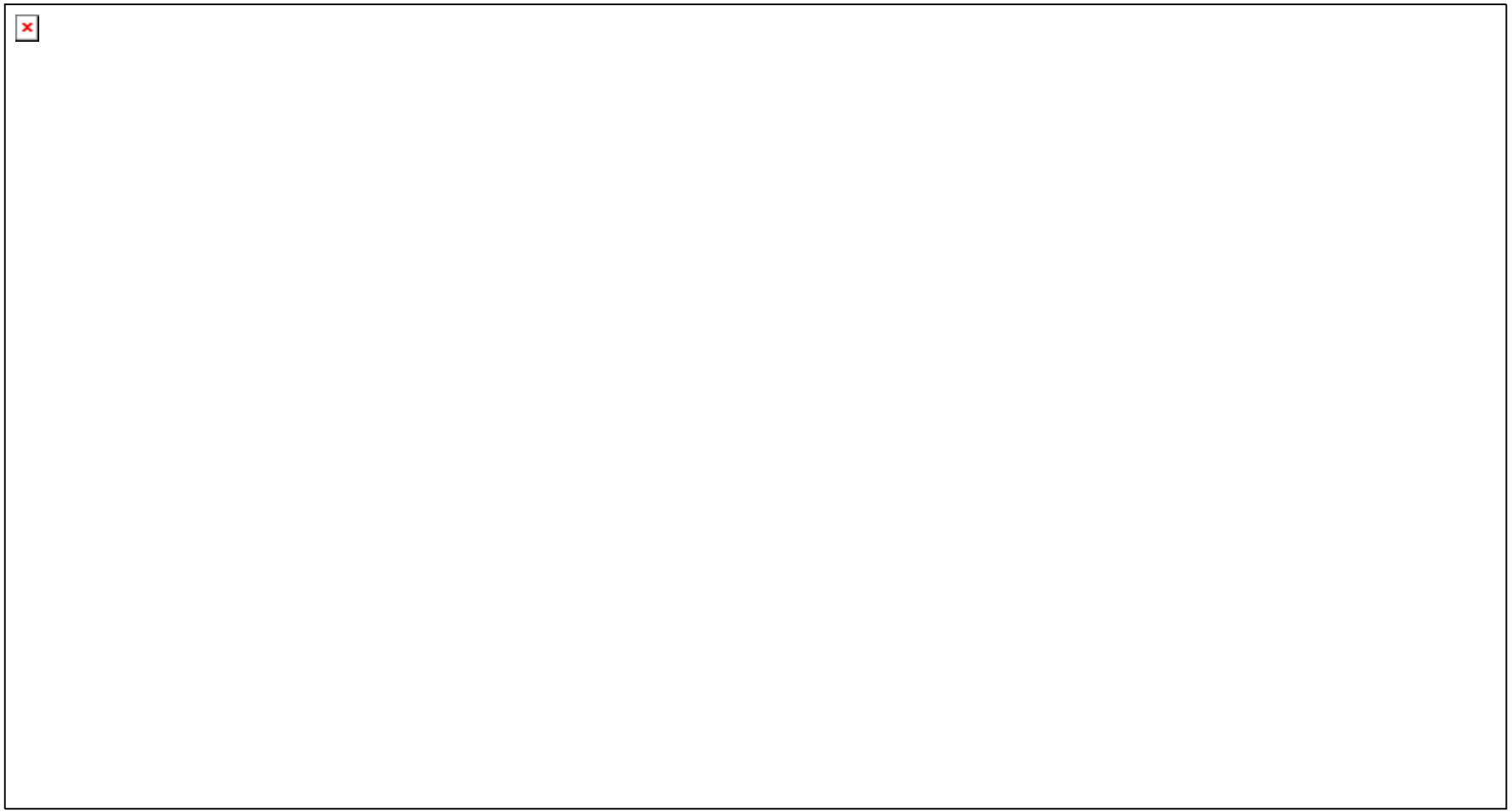




- $e_i = Y_1 - \hat{Y}_1 = 399 - 347.98 = 51.02$，圖中可以看出殘差大小根據 Y_i 與所估計迴歸函數之**垂直離差**而定(即對$\hat{Y}_i$之離差)。



- 特別需區分的是誤差 $\varepsilon_i = Y_i - E\{Y_i\}$，而殘差 $e_i = Y_i - \hat{Y}_i$，前者誤差是Y_i對於真實歸迴線之離差，所以是處於未知之狀態，而殘差是指觀測值Y_i與所估計出之歸迴線上配適值 $\hat{Y}_i$之間的垂直離差，所以是已知。



• 配適迴歸線之性質

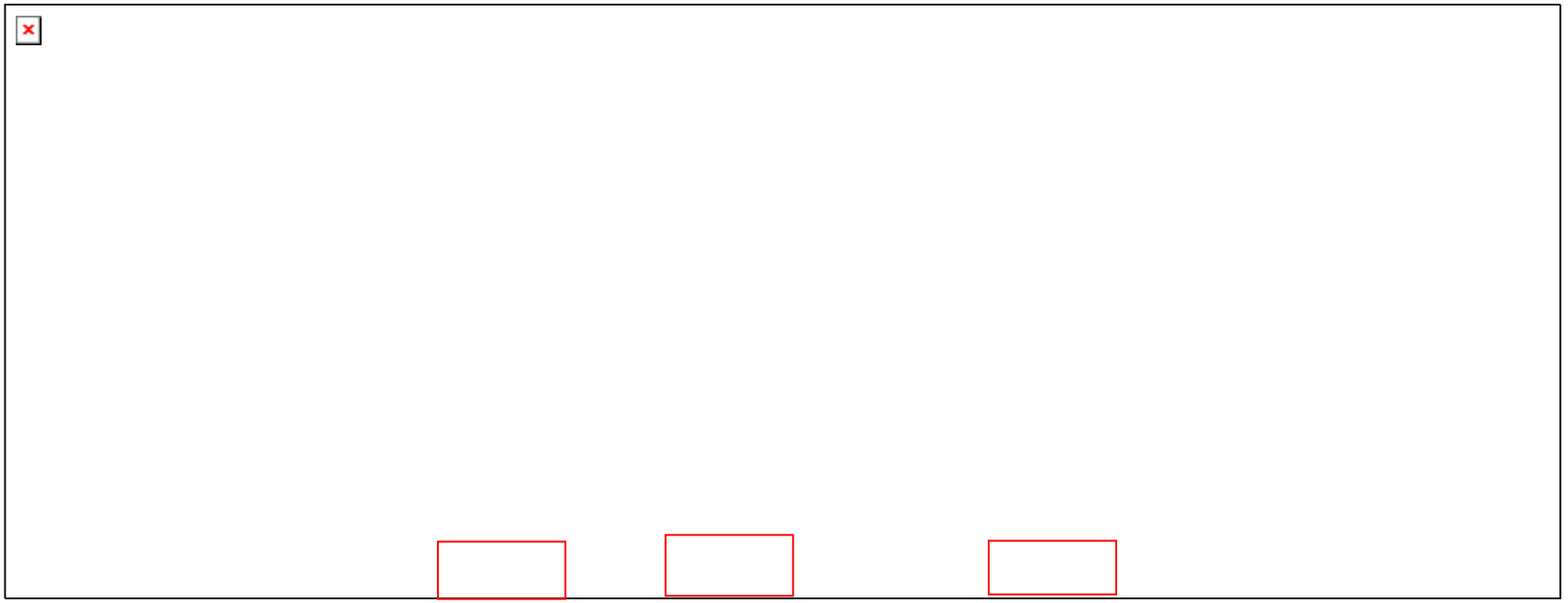
以最小平方法所估計之歸迴直線有一些之性質

1.殘差總合為零：

(1.17)

2.殘差平方和 為最小值，而這正是最小平方法所要求的條件，在(1.8)中要求極小化 Q。

3.觀測值 Y_i 總合與配適值 $\hat{Y}_i$ 總合相等：$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$ (1.18)



4.以X_i作為殘差e_i之權重，並且計算總合時，其結果亦為零：

$$\sum X_i e_i = 0 \qquad (1.19)$$

5.利用(1.17)與(1.19)之結果，以 $\hat{Y}_i$ 作為殘差 e_i 之權重時，並且計算其總合，結果亦為零：

$$\sum \hat{Y}_i e_i = 0 \qquad (1.20)$$

6.迴歸直線必定通過點（ $\bar{X}$， $\bar{Y}$ ）。

1.7　誤差項變異數σ^2之估計

- 迴歸模型（1.1）中的誤差項 ε 之變異數 σ^2 可以用來當作 Y 之機率分配變異性的指標

σ^2 之點估計

■ 單一母體

- 單一母體變異數σ^2是透過樣本變異數 s^2 來進行估計的，而計算樣本變異數之方法，是考慮觀測值 Y_i 與估計的離差經過平方後取總合：

$$\sum_{i=1}^{n}(Y_i-\hat{\overline{Y}}_i)^2$$

- 上述過程我們稱為**平方和**（*sum of squares*）

- 最後是除上它的**自由度**（*degress of freedom*），在此處自由度為 $n-1$，因為透過樣本平均數 來估計未知的母體平均數 μ，因此失去了一個自由度，所以常用的**樣本變異數** s^2 如下：

$$s^2 = \sum_{i=1}^{n} (Y_i - \overline{\hat{Y}_i})^2 / (n-1)$$

樣本變異數常被稱為**均方 (mean square)**

■ 迴歸模型

- 迴歸模型計算平方和，是以觀測值 Y_i 與平均數之離差，是根據平均數之估計值 $\hat{Y}_i$ 而進行，所以此處之離差即為前述之**殘差**：

而**平方和**為：

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \qquad (1.21)$$

其中*SSE*表示 **誤差平方和** (error sum of squares)或 **殘差平方和** (residual sum of squares) 。

- 平方和 SSE 自由度為 $n-2$，所損失的兩個自由度是發生在進行平均數之估計值 $\hat{Y}_i$ 時，須先估計 β_0 與 β_1 這兩個參數，因此較為適合的均方或 s^2 為：

(1.22)

其中 MSE 表示**誤差均方**(error mean square)或**殘差均方**(residual mean squares)。

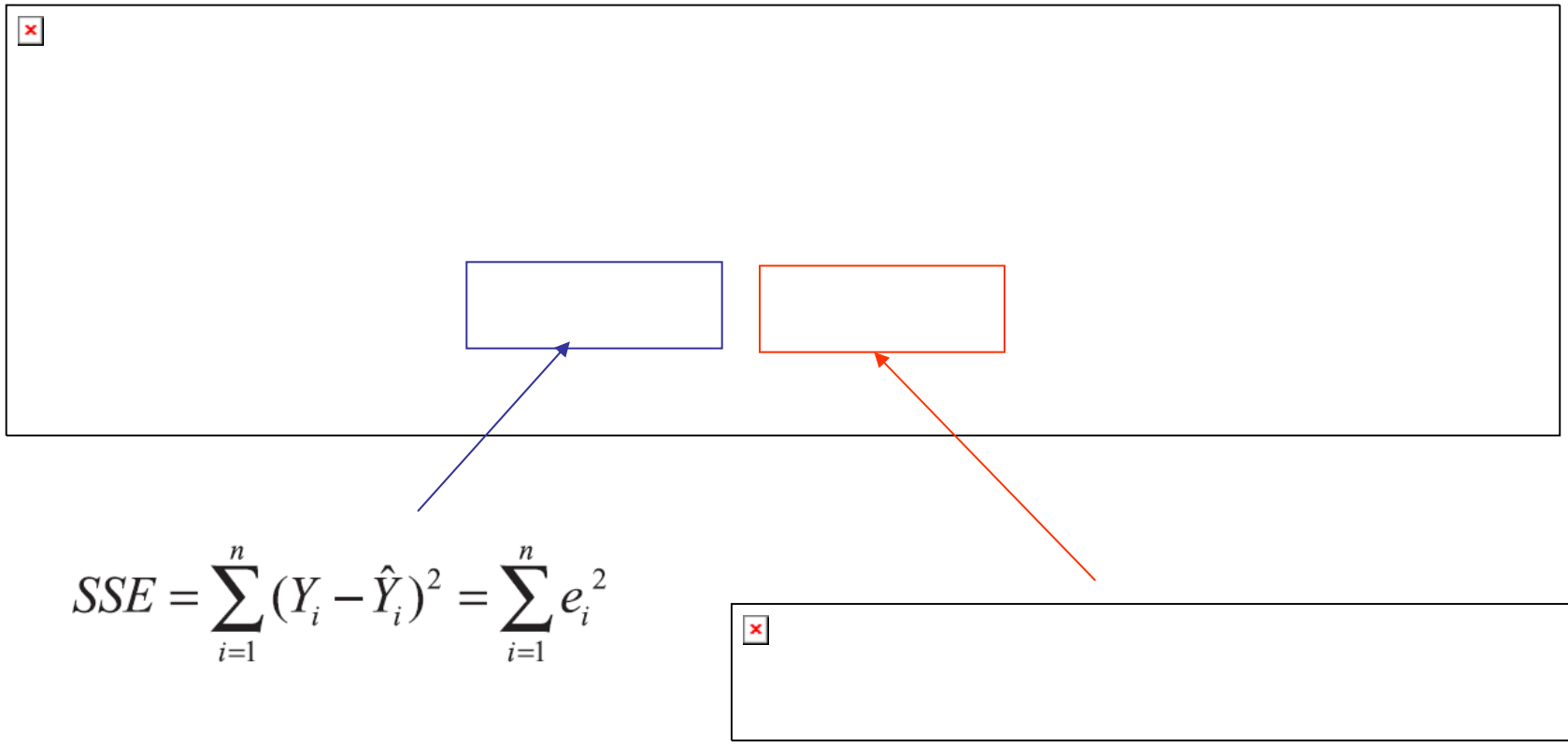
- 在迴歸模型(1.1)中，可以證明出 MSE 為 σ^2 的一個不偏估計量，亦即：

(1.23)

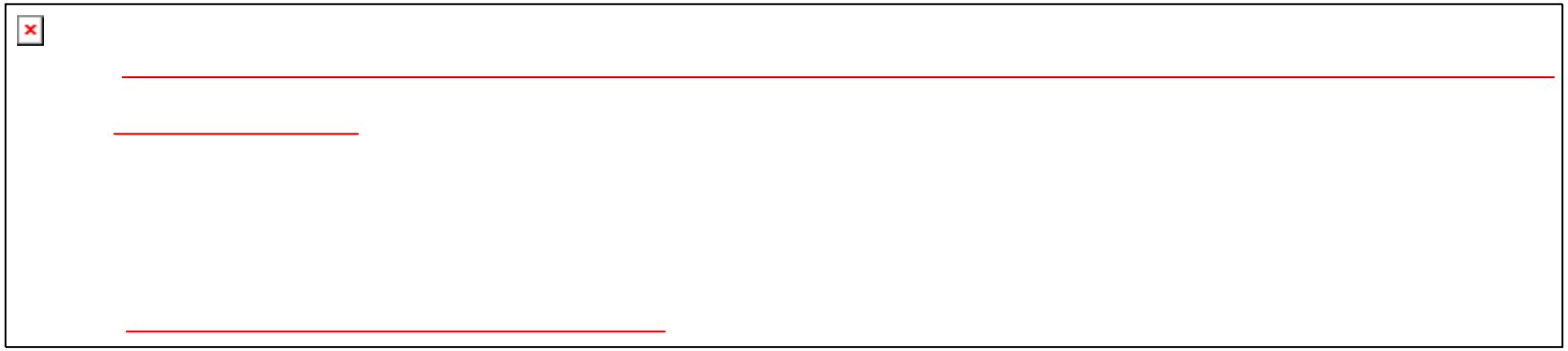
而標準差 σ 之估計量為***MSE*之平方根**，或  。

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

EXCEL檔 - CH01TA01



$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$



The regression equation is
Y = 62.4 + 3.57 X

Predictor	Coef	Stdev	t-ratio	p
Constant	62.37	26.18	2.38	0.026
X	3.5702	0.3470	10.29	0.000

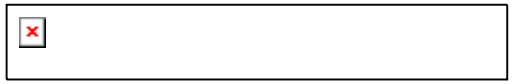
s = 48.82 R-sq = 82.2% R-sq(adj) = 81.4%

1.8 常態誤差迴歸模型

- 不論誤差項 ε_i（或是觀測值Y_i）之機率分配是何種形式，最小平方法提供了 β_0 與 β_1 之不偏點估計量，同時在所有線性不偏之估計中，該組估計量之變異數為最小。
- 然而如果想要進行**區間估計或檢定**，就必須對於誤差項 ε_i 進行某些形式的機率分配之假設，通常我們會假設誤差值服從**常態分配**，這樣的假設將有助於簡化迴歸分析之理論，並且在許多應用迴歸分析之問題上，也被認同。

- **模型**

常態誤差迴歸模型如下：



(1.24)

其中， Y_i 為反應變數在第i次實驗下之結果

X_i 為已知之常數，代表了預測變數在第i次實驗時之值

β_0 與 β_1 均為參數，

 獨立且服從**常態分配** $N(0，\sigma^2)$，$i = 1, \ldots, n$

說明

1.符號 $N(0, \sigma^2)$ 表示常態分配，平均數為 0，變異數 σ^2 。

2.迴歸模型（1.24）與迴歸模型（1.1）除了前者假設誤差項 ε_i 服從**常態分配 $N(0, \sigma^2)$** 外，兩者為相同之模型。

說明

3.因為迴歸模型（1.24）假設誤差項 ε_i 服從常態分配，在模型（1.1）中假設 ε_i 無相關性，就是常態誤差模型獨立性。

4.迴歸模型（1.24）也說明了 Y_i 是互相獨立之常態隨機變數，其平均數與變異數分別為 $E\{Y_i\}=\beta_o+\beta_1 X_1$ 與 σ^2 。

5.誤差項的代表涵義常是一些不被模型考慮且與X無關效果，在許多情形下對誤差項進行常態性的假設是合理。

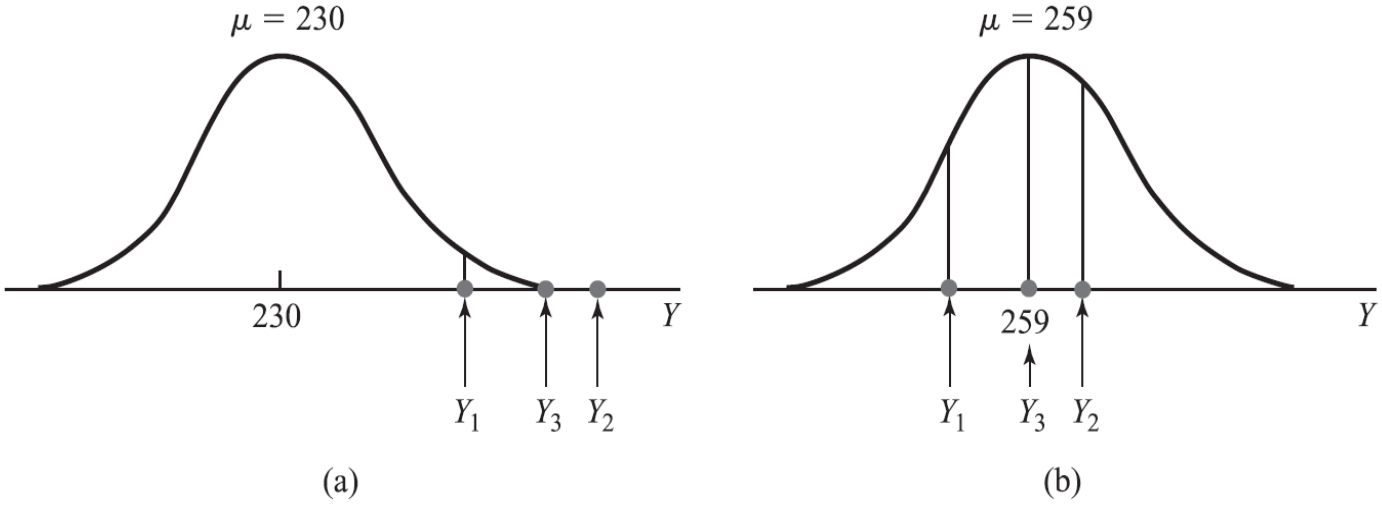
最大概似估計法下之參數估計

- 當誤差項的機率分配函數形式被指定後(常態分佈)，參數 β_0、β_1 與 σ^2 之估計量將可以透過**最大概似估計法**（*method of maximum likelihood estimation*）得到；基本上是在找出一組與樣本資料最為一致的參數值。
- **單一母體**

  考慮一個標準差 $\sigma = 10$，但是平均數未知是常態分配，從該分配中抽樣 $n = 3$ 個隨機樣本，觀測值 $Y_1 = 250$，$Y_2 = 265$，$Y_3 = 259$，想知道平均數 μ 是多少時最有可能得到上面三樣本觀測值 。

- 圖 1.3a 為如果 $\mu = 230$、$\sigma = 10$ 的常態分配圖形以及三個樣本觀測值得相對位置，三個樣本觀測值都是在圖中右尾的位置，所以 $\mu = 230$與樣本資料似乎不一致。
- 圖 1.13b 為如果 $\mu = 259$、$\sigma = 10$ 常態分配圖形以及三個樣本觀測值得相當位置，此時三個樣本觀測值都是在圖形中間附近的位置，因此 $\mu = 259$相對於 $\mu = 230$而言，與樣本資料較為一致。

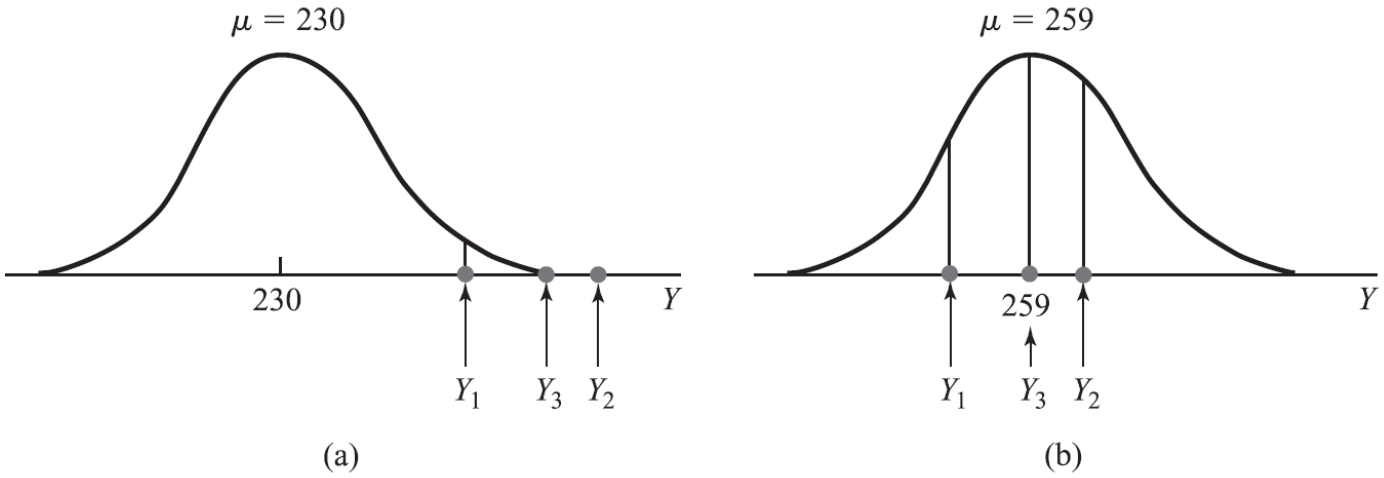
圖 1.13
兩個可能的 μ 值其樣本觀測值($Y_1 = 250$、$Y_2 = 265$、$Y_3 = 259$)之密度。



觀測值 $Y_1 = 250$，$Y_2 = 265$，$Y_3 = 259$

- **最大概似估計法**是利用不同觀測值 Y_i 在其機率分配中分別的**密度**（亦即密度曲線在 Y_i 之高度）作為觀測值 一致性的衡量。
- 考慮例中的觀測值Y_1，若Y_1在尾部（如圖1.13a），則**曲線高度**相當小；若Y_1在較靠近中間的位置（如圖1.13b），高度會比較高。

圖 1.13
兩個可能的μ值其樣本觀測值(Y_1 = 250、Y_2 = 265、Y_3 = 259)之密度。



- 利用附錄 A 所提之常態分配的密度函數(A.34)[A.4-PA7]，可以得到Y_1的密度，以符號f_1表示，圖1.13中兩個情況的μ值其f_1分別為：常態分佈之隨機變數$Y_1 = 250$ (P1-27)

$\mu = 230$:

$\mu = 259$:

- 兩個情況的μ值其三個樣本觀測值如下： f_2、f_3 [P1-27]

	$\mu = 230$	$\mu = 259$
$Y_1 = 250$， f_1	0.005399	0.026609
$Y_2 = 265$， f_2	0.000087	0.33322
$Y_3 = 259$，f_3	0.000595	0.39894

觀測值 $Y_1 = 250$，$Y_2 = 265$， $Y_3 = 259$

- 最大概似估計法是利用**密度乘積**（亦即此處高度的乘積），作為參數值與樣本資料一致性的衡量，此**積**稱為參數值 μ 之概似值（likelihood value），以符號 L（μ）表示，**如果 μ 值與樣本資料越一致，則密度將越高，於是相乘積（即概似值）也將相對地高**
- 在上例中，兩個不同的 μ 值其概似值分別如下：

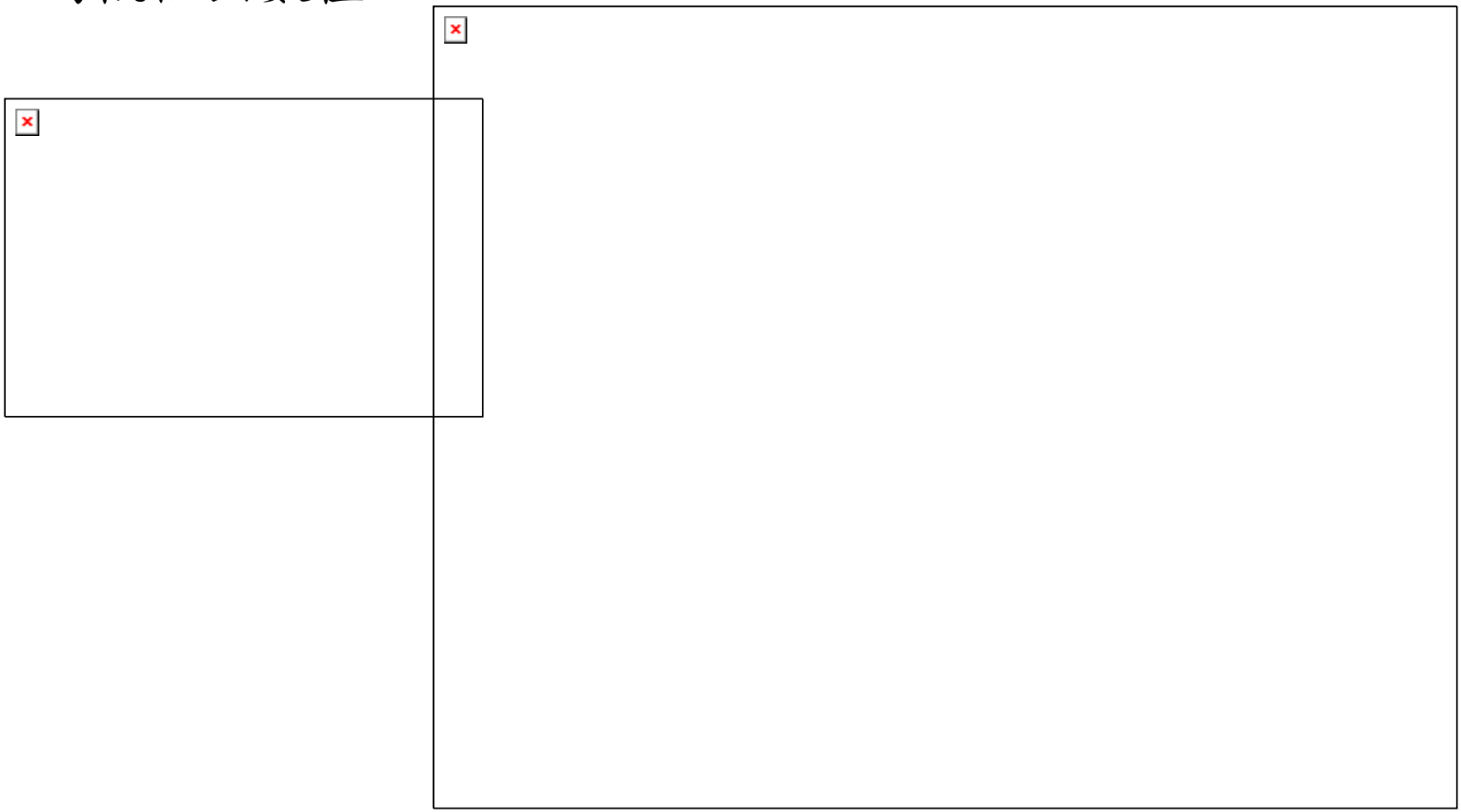
  $L(\mu=230) = (0.005399)(0.000087)(0.000595) = 0.279\times10^{-9}$

  $L(\mu=259) = (0.026609)(0.33322)(0.039894) = 0.0000354$
- 本例中平均數 258，母體平均數平均數 μ 的得最大概似估計量為 258，事實上，L（$\mu=258$）$=0.0000359$的確比L（$\mu=259$）$=0.0000354$略高。

- 最大概似估計法選擇能讓**概似值最大的μ值**，做為**最大概似估計值**，如同最小平方法，它也有兩個方法可以求得，分別是透過數值方法有系統地搜尋以及透過解析方法求解。
- 在上例中，我們若是透過解析的方法求得最大概似估計量，可以證明常態母體平均數μ的最大概似估計量為樣本平均數Y，本例中樣本平均數 Y ＝ 258，所以母體平均數平均數μ的得最大概似估計量為 258，L（μ＝258）＝ 0.0000359 較高
- 如果將密度的相乘積視為未知參數的函數，則該函數稱為**概似函數**（likelihood value），在上例中已知σ＝10，所以概似函數為：

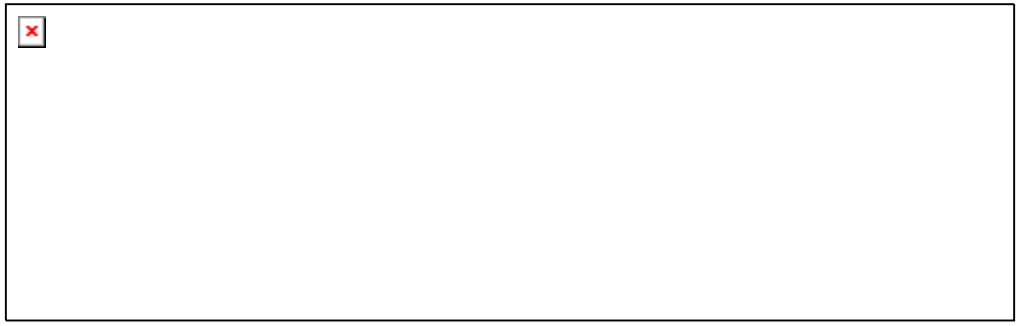
$$L(\mu)=[\frac{1}{\sqrt{2\pi}(10)}]^3 \exp[-\frac{1}{2}(\frac{250-\mu}{10})^2]\exp[-\frac{1}{2}(\frac{265-\mu}{10})^2]\exp[-\frac{1}{2}(\frac{259-\mu}{10})^2]$$

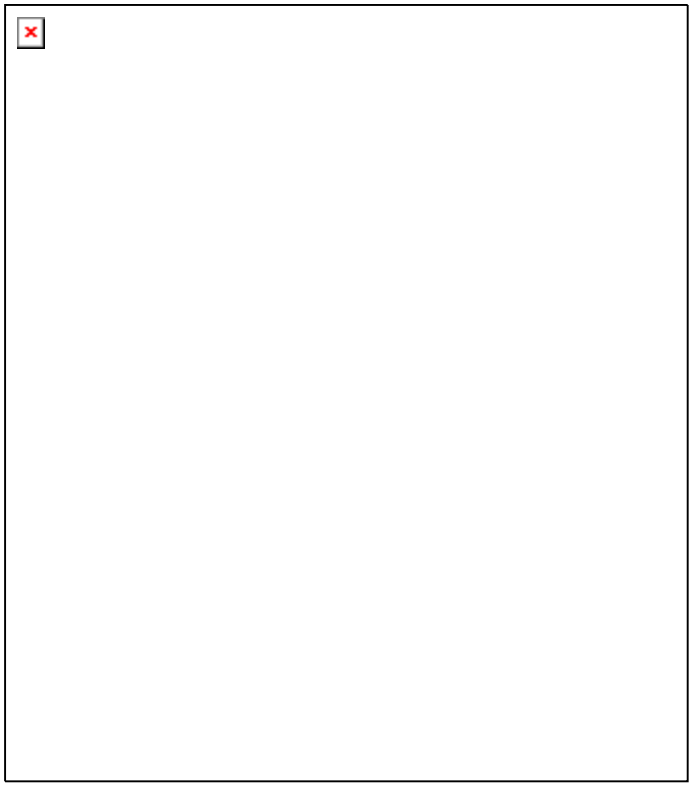
- 圖1.14為利用電腦針對本例，透過許多個不同的μ計算分別的概似函數值



迴歸模型

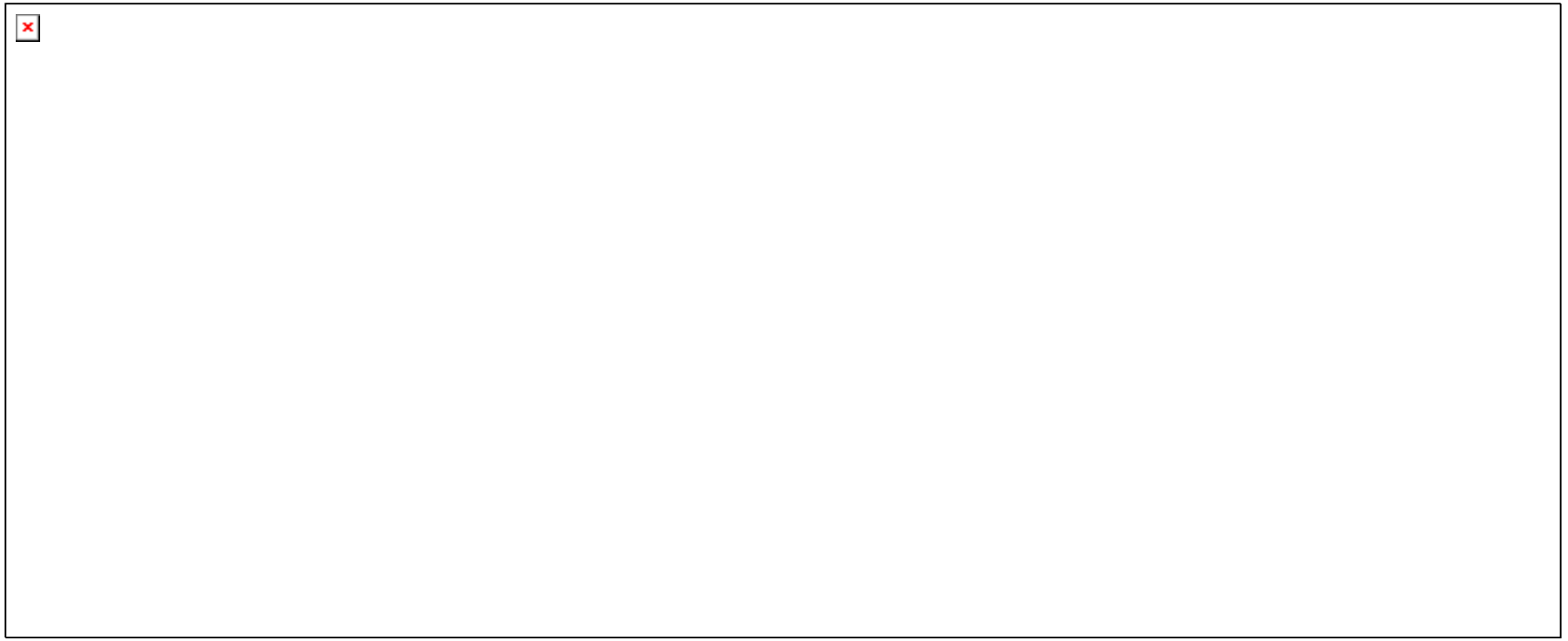
- 單一母體平均數的得最大概似估計觀念，可以直接用在常態誤差迴歸模型（1.24）的參數估計上。在此模型下每一個Y_i均服從平均數 $E\{Y_i\} = \beta_0 + \beta_1 X$、標準差 σ 的常態分配。利用前面有關耐性實驗例子說明迴歸模型下的最大概似估計法，直接假設 $\sigma = 2.5$，想計算出當參數 $\beta_0=0$、$\beta_1=0.5$ 下之概似值



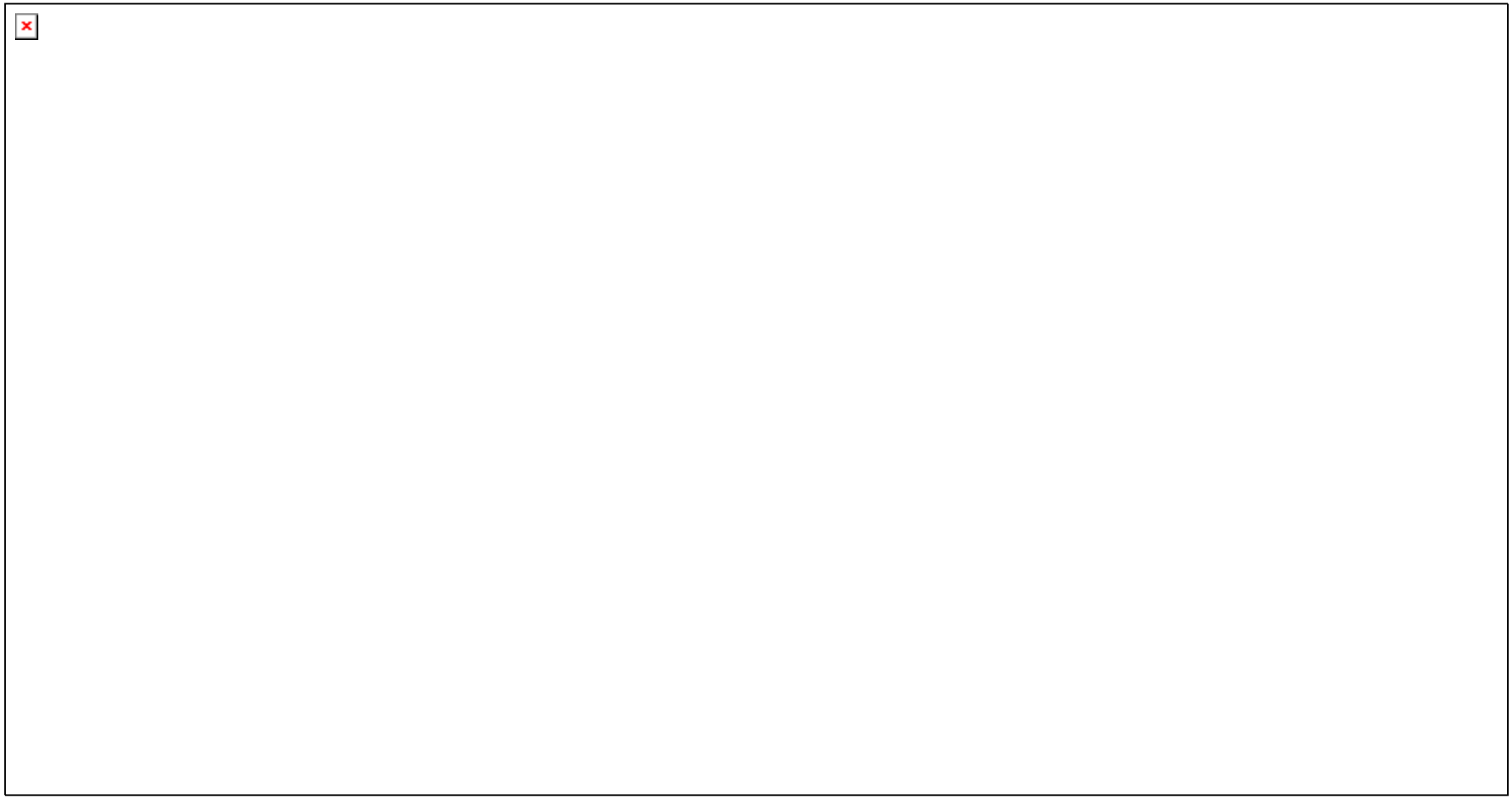


- 對第一位研究對象，$X_1 = 20$，機率分配之平均數為 $\beta_0+\beta_1 X_1 = 0+0.5(20) =10.0$，如圖 1.15a，平均數為10.0，標準差 $\sigma = 2.5$之常態分配，觀測值 $Y_1 = 5$ 是位於此分配的左尾，所以密度相對較小

- 第二位研究對象，$X_2 = 55$ ，如圖 1.15b，機率分配之平均數為 $ß_o + ß_1 X_1 = 0 + 0.5(55) = 27.5$ ，在圖1.15b中為平均數為 27.5 ，標準差 $\sigma = 2.5$ 之常態分配，此時 $Y_2 = 12$，因其**密度相當低**，完全不像是來自此分配之樣本。
- 第三位研究對象，$X_3 = 30$ ，如圖 1.15c，機率分配之平均數為 $ß_o + ß_1 X_1 = 0 + 0.5(30) = 15.0$ ，此時 $Y_2 = 10$，因其**密度相當低**，也非常不像是來自此分配之樣本。



- 圖 1.15d 綜合上述訊息，畫出迴歸函數 $E\{Y\} = 0 + 0.5X$，三個樣本以及三個相對的常態分配，顯然此一迴歸直線的配適非常差，同時其密度值也很低，所以 $\beta_0 = 0$ 與 $\beta_1 = 0.5$ 與樣本資料並不一致。



- 同樣計算個別的密度，對於 $Y_1 = 5$、 $X_1 = 20$，當 $ß_o = 0$ 與 $ß_1 = 0.5$ 時，平均數為 $ß_o + ß_1 X_1 = 0 + 0.5(20) = 10.0$：(P1-28)

第一個樣本之密度 $f_1 = \frac{1}{\sqrt{2\pi}(2.5)} \exp\left[-\frac{1}{2}\left(\frac{5-10.0}{2.5}\right)^2\right] = 0.021596$

第二個樣本之密度 $f_2 =$ …… $= 0.7175 \times 10^{-9}$

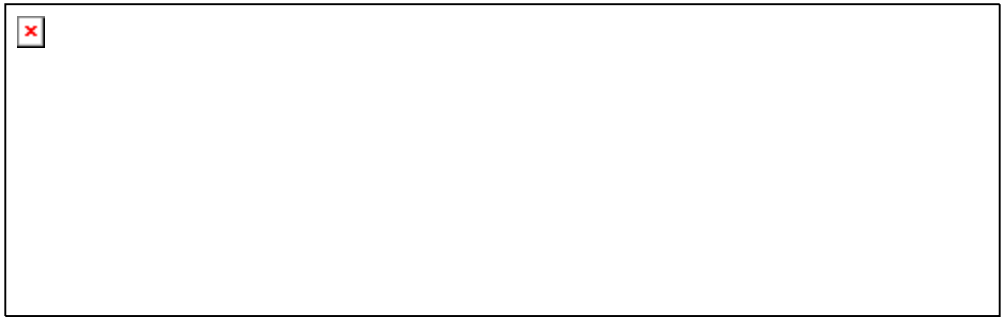
第三個樣本之密度 $f_3 =$ …… $= 0.021596$

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

- **概似值為**:

$L(ß_o = 0 \text{ 與 } ß_1 = 0.5) = (0.021596)\ (0.7175 \times 10^{-9})\ (0.021596)$

$= 0.3346 \times 10^{-12}$



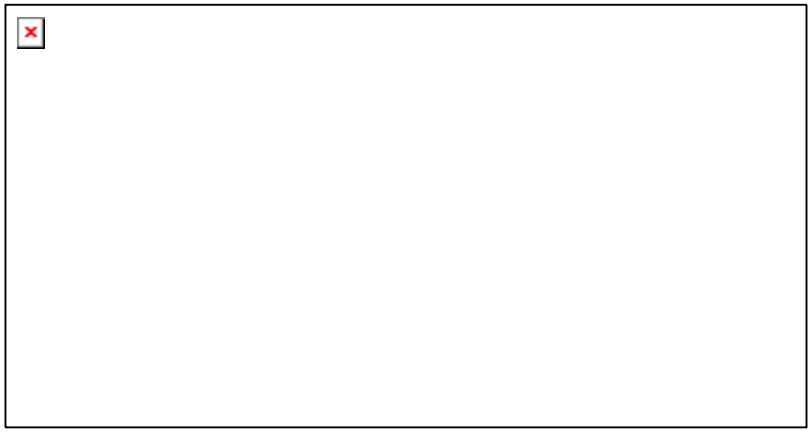
- 一般而言，在常態誤差迴歸模型(1.24)中，利用$E\{Y_i\} = \beta_0 + \beta_1 X_i$與$\sigma^2\{Y_i\} = \sigma^2$，可以計算出觀測值$Y_i$之密度：

$$f_i = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right] \qquad (1.25)$$

對於n個觀測值$Y_1, Y_2, \ldots, Y_n$，其對應之**概似函數為(1.25)中個別密度值的相乘積**，因為誤差項之變異數σ^2通常未知，所以概似函數為參數β_0、β_1與σ^2之函數：

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \end{aligned} \qquad (1.26)$$

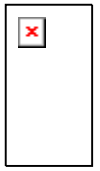
- 能夠**最大化**(1.26)概似函數的一組β_0、β_1與σ^2值即為**最大概似估計值**，而分別以符號 [ ]、$\hat{\beta}_1$ 與 [ ] 表示。這三個估計量解析出後之結果如下：



(1.27)

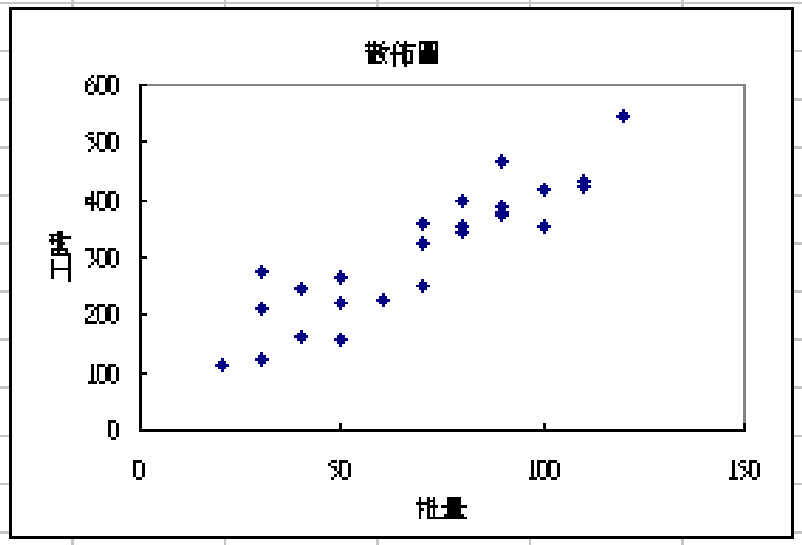
- 故 β_0 與 β_1 的最大概似估計量跟最小平方法相同，但是變異數的最大概似估計量有偏，以不偏估計量 *MSE* 取代且與最大概似估計量 [ ] 差別不大，特別是當 *n* 夠大時：

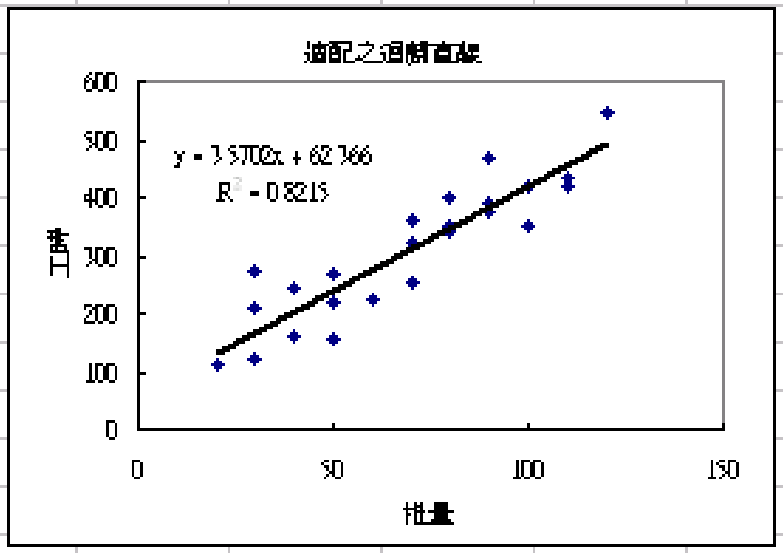
×



(1.28)

	A	B	C	D	E	F	G	H
1	批量	工時						
2	80	399						
3	30	121						
4	50	221						
5	90	376						
6	70	361						
7	60	224						
8	120	546						
9	80	352						
10	100	353						
11	50	157						
12	40	160						
13	70	252						
14	90	389						
15	20	113						
16	110	435						
17	100	420						
18	30	212						
19	50	268						
20	90	377						
21	110	421						
22	30	273						
23	90	468						
24	40	244						
25	80	342						
26	70	323						
27								





	A	B	C	D	E	F	G	H
1		(1)	(2)	(3)	(4)	(5)	(6)	(7)
2	批次	批量	工時					
3	i	Xi	Yi	X_i-X̄	Y_i-Ȳ	(X$_i$-X̄)(Y$_i$-Ȳ)	$(Xi-X)^2$	$(Yi-Y)^2$
4	1	80	399	10	86.72	867.2	100	7520.4
5	2	30	121	-40	-191.28	7651.2	1600	36588.0
6	3	50	221	-20	-91.28	1825.6	400	8332.0
7	4	90	376	20	63.72	1274.4	400	4060.2
8	5	70	361	0	48.72	0.0	0	2373.6
9	6	60	224	-10	-88.28	882.8	100	7793.4
10	7	120	546	50	233.72	11686.0	2500	54625.0
11	8	80	352	10	39.72	397.2	100	1577.7
12	9	100	353	30	40.72	1221.6	900	1658.1
13	10	50	157	-20	-155.28	3105.6	400	24111.9
14	11	40	160	-30	-152.28	4568.4	900	23189.2
15	12	70	252	0	-60.28	0.0	0	3633.7
16	13	90	389	20	76.72	1534.4	400	5886.0
17	14	20	113	-50	-199.28	9964.0	2500	39712.5
18	15	110	435	40	122.72	4908.8	1600	15060.2
19	16	100	420	30	107.72	3231.6	900	11603.6
20	17	30	212	-40	-100.28	4011.2	1600	10056.1
21	18	50	268	-20	-44.28	885.6	400	1960.7
22	19	90	377	20	64.72	1294.4	400	4188.7
23	20	110	421	40	108.72	4348.8	1600	11820.0
24	21	30	273	-40	-39.28	1571.2	1600	1542.9
25	22	90	468	20	155.72	3114.4	400	24248.7
26	23	40	244	-30	-68.28	2048.4	900	4662.2
27	24	80	342	10	29.72	297.2	100	883.3
28	25	70	323	0	10.72	0.0	0	114.9
29	總合	1750	7807	0	0	70690	19800	307203
30	平均	70.0	312.28					
31								

變異數分析Analysis of Variance 或ANOVA- SAS

GTL 或 ANOVA 程序，除了ANOVA之F檢定外，GTL程序或ANOVA程序提供了多重平均數比較的檢定方法，這些比較方法包括有班佛尼氏(Bonferroni) t 檢定、沙菲氏(Sheffe)檢定，土其氏(Tukey)檢定、LSD

單因子變異數分析 ANOVA

某生產線選定四種不同的機台來量測生產產品的產量，所得數據如下，試問四種不同的機台來量測生產產品的產量是否有差異

machine			
1	2	3	4
349.4	347.8	350.8	349.8
348.0	349.2	350.1	351.4
349.8	347.8	348.9	350.9
344.5	348.1	351.2	349.5
353.5	348.1	349.6	346.8
345.8	347.6	348.5	350.9
346.8	348.3	350.9	351.0
344.7	347.1	349.6	349.5
348.3	346.5	351.0	351.6
347.6	347.6	351.2	347.6

雙因子變異數分析 ANOVA

調查男女學生在不同住處（學校宿舍、自宅、校外公寓及分租房間）其每週的飲酒量如下，每住處選 8 位學生男女生各半