CH 12

迴歸分析的統計推論

Modern Engineering Statistics

工 程 統 計

在第4章及第5章中，我們已經介紹了如何利用最小平方法來建立變數間的迴歸關係式。

本章主要將介紹迴歸分析的統計推論，迴歸預測及變異分析。

12-1 線性迴歸分析的假設與性質

◎線性迴歸分析的假設

對任一特定或固定的X值，均有一對應的平均值時，即母體平均值時，以$\mu_{Y,X}$表示。由於$\mu_{Y,X}$為在給予特定X值下Y的平均值，因此稱為已知X下Y的條件平均數（Conditional Mean）。

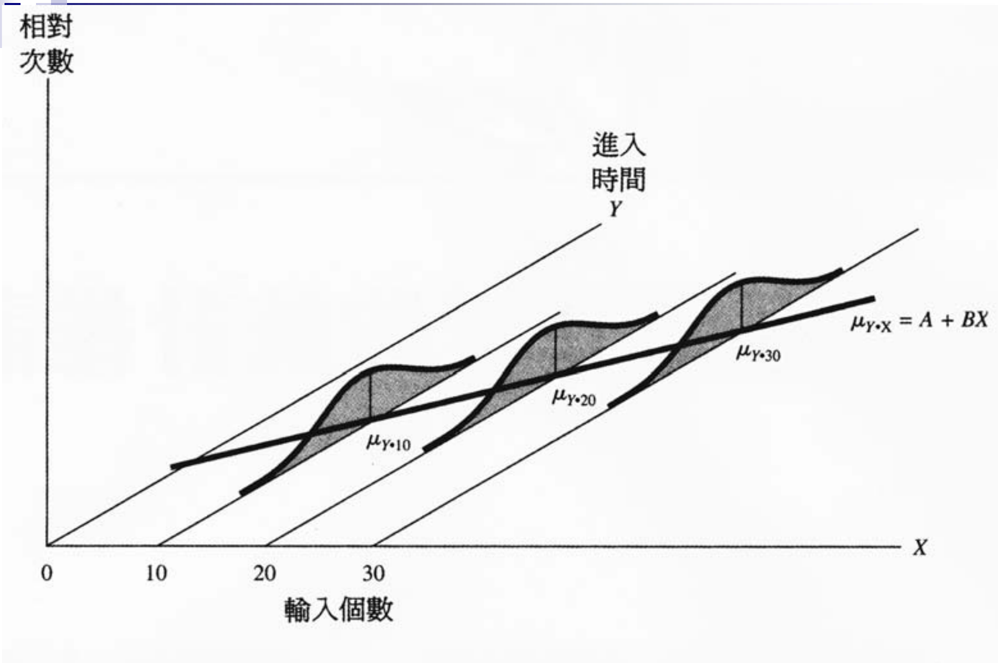


圖 12-1 在特定 X 下母體平均數之真實迴歸線

以下為我們對母體Y所作的四項假設

a.當條件平均數均落於相同的直線上，此直線稱為母體的真實迴歸線（True Regression Line），直線可表為下式。

$$\mu_{Y\cdot X} = A + BX$$

當樣本觀測未完全落在真實迴歸線，即產生隨機誤差（Random Error)第i個樣本觀測點的因數Y值，等於由該觀測點真實值加上隨機誤差項，即$\in_i$假設唯一獨立的隨機變數，其期望值為0，變異數σ^2為一特（固）定值。

$$Y_i = A + BX_i + \in_i$$

b.因此對任一已知X值下，σ^2可視為觀察值偏離母體迴歸直線的衡量，而觀測值Y的期望值等於$\mu_{Y,X}$。

不論X值為何，所有母體均有一相同的標準差，以$\mu_{Y,X}$表示。

$$Var(Y_i) = Var(A + BX_i + \in_i) = Var(\in_i)$$

c.由於連續的樣本觀察值互為獨立，其對應的Y值亦為獨立。

d.雖然X有時可能為一不確定量，然在迴歸分析中係視X為一事先給予或已知的特定值，並用以預測或計算未知的Y值。亦即，計算Y在某一特定的X值下之條件機率。

◎真實迴歸式的估計

由於真實迴歸式$\mu_{Y \cdot X} = A + BX$中之**真實迴歸係數（True Regression Coefficients）**A與B未知，因此通常須以最小平方法求得樣本資料之估計的迴歸式$\hat{Y}(X) = a + bX$的係數a與b，加以推論求得。

◎最小平方法的合理性

由於以最小平方法所求得之估計迴歸係數a與b，是為真實迴歸係數A與B的不偏估計量。

◎殘差分析

個別觀察值Y_i距適值$\hat{Y}_i$之差稱為殘差（Residual）。殘差可視為觀測誤差，可表為下式

$$e_i = Y_i - \hat{Y}(X_i)$$

表 12-1　工作站處理資料需時資料及迴歸計算表列

觀察值 t	處理資料需時 Y_i	資料個數 X_i	迴歸式 $Y(X_i) = 30.04 + 2.854X_i$	殘差 $e_i = Y_i - \hat{Y}(X_i)$
1	66	2	35.748	30.252
2	77	19	84.266	−7.266
3	37	6	47.164	−10.164
4	106	23	95.682	10.318
5	55	10	58.580	−3.580
6	89	23	95.682	−6.682
7	52	9	55.726	−3.726
8	128	30	115.660	12.340
9	63	18	81.412	−18.412
10	104	25	101.390	2.610
11	76	19	84.412	−8.266
12	44	2	35.746	8.252
13	97	27	107.098	−10.098
14	109	28	109.952	−0.952
15	40	8	52.872	−12.872
16	124	29	112.806	11.194
17	98	29	112.806	−14.806
18	63	16	75.704	−12.704
19	131	33	124.222	6.778
20	41	3	38.602	2.398
21	111	34	127.076	−16.076
22	151	32	121.368	29.632
23	76	13	67.142	8.858
24	114	33	124.222	−10.222
25	143	35	129.93	13.070
	$\sum Y = 2,195$	$\sum X = 506$		
	$\bar{X} = 20.24$		$\bar{Y} = 87.80$	
	$\sum X^2 = 13,130$		$\sum Y^2 = 220,445$	$\sum XY = 52,670$
	$\hat{Y} = 30.04 + 2.854X$		$s_{Y \cdot X} = 13.51$	

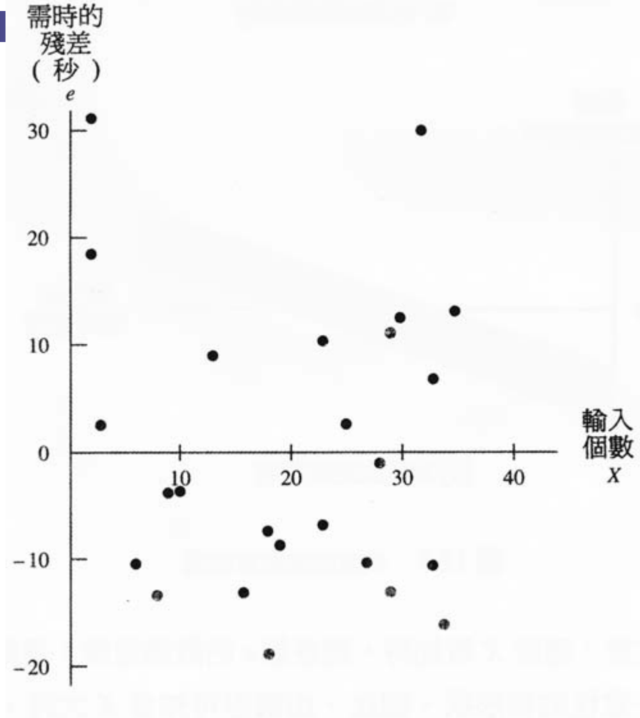
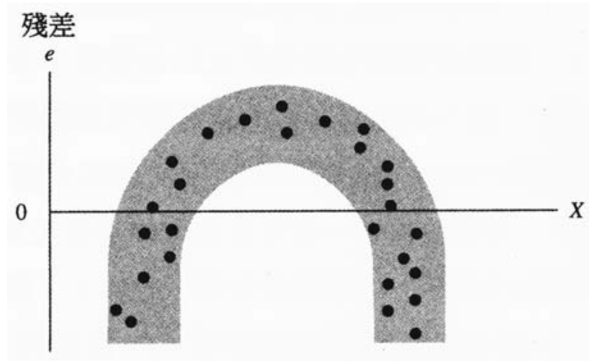
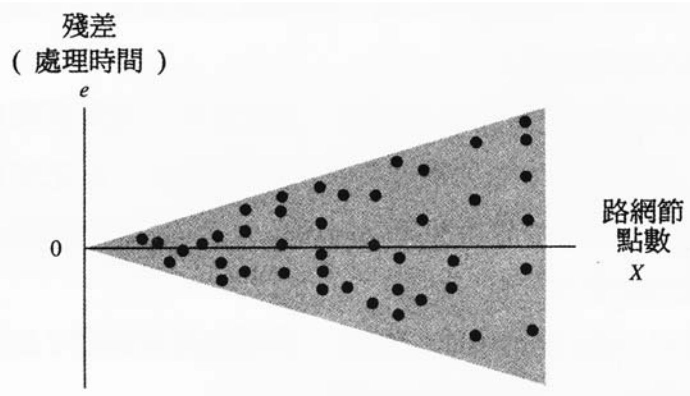


圖 12-2 殘差的散佈圖示



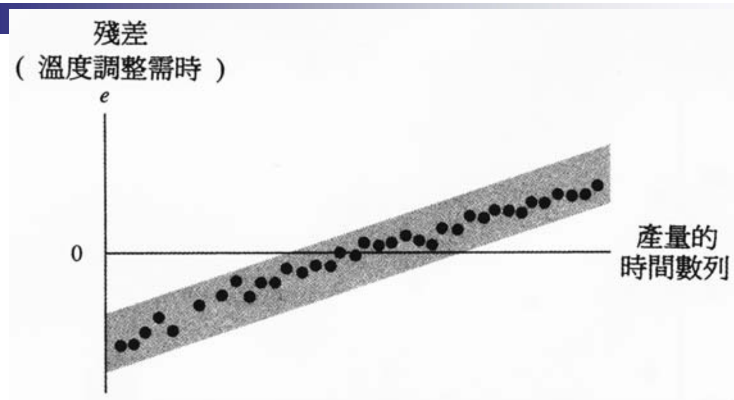
(a) 非線性關係

回歸函數為非線性，利用第五章轉換為線性



(b) 非定值的變異數

誤差隨X增大而增大，違反誤差變異數為一致性之假設。可利用$Y=\sqrt{Y}$



(c) 非獨立的誤差項

圖 12-3 典型的殘差散佈圖

誤差隨時間增加而增加，顯示誤差非獨立。可能因操作員疲勞等非隨機原因造成。

◎**複迴歸分析的陷阱**

共線性 當兩個獨立變數間存有高度的相關性時，將使得複迴歸的分析變成極為複雜且易發生誤導。

共線性（Multicollinearity)： 當預測變數X_1與X_2間有高度的相關性，表示彼此共同存在預測資料

12-2 評估迴歸式的品質

回歸分析：

1. 建立迴歸式
2. 評量迴歸式之整體適合度

　　探討整體迴歸式對於個別因變數解釋程度之高低，最常以判定係數 (Coefficient of Determination)來量度變數間的關係強度。

◎判定係數

任一觀測值均存在三個離差

$$(\mathrm{Y}-\overline{\mathrm{Y}})=[\hat{\mathrm{Y}}(\mathrm{X})-\overline{\mathrm{Y}}]+[\mathrm{Y}-\hat{\mathrm{Y}}(X)]$$

◎.上式左項表示任一觀測值與所有樣本Y的平均數之差異，稱為總離差 (Total Deviation)。如圖12-4， 第22個樣本觀測值所對應總離差151-87.8=63.2

◎.總離差由右邊兩個部分離差組成，說明如下

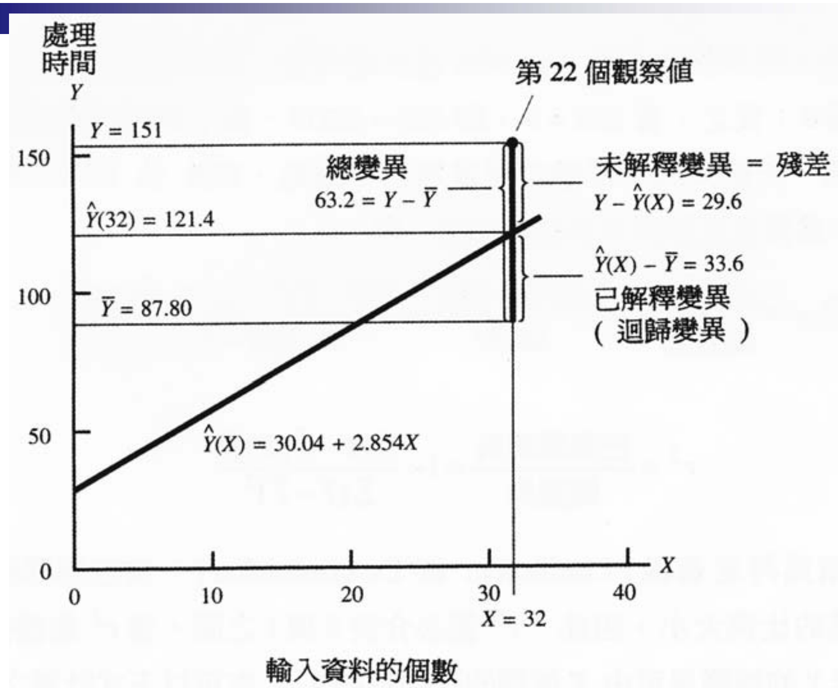


圖 12-4 迴歸變異的分析圖

- 第一個部分為**可解釋差異：**

該觀測值對應之迴歸式估計值與所有樣本平均數之差量，此差量為迴歸式可以解釋的部分，故稱之為**可解釋差異(Explained Deviation)**。

- 第二個部分為**未解釋的差異**

係由以量測個別的*Y*值與迴歸式估計值間的差異程度，此乃因為所有觀測點均落於迴歸線上的情況幾乎為不可能，而若有觀測點不落於迴歸線上，顯示有其他隨機因素造成非純由Y與X間差異所造成。故而稱此差異為**未解釋的差異(Unexplained Deviation)**。

$$(Y-\bar{Y})=[\hat{Y}(X)-\bar{Y}]+[Y-\hat{Y}(X)]$$

上式取平方，並將所有觀測點取總和

總變異 = 已解釋的變異 + 未解釋的變異

$$\sum_{i=1}^{n}[Y_i-\bar{Y}]^2=\sum_{i=1}^{n}[\hat{Y}(X_i)-\bar{Y}]^2+\sum_{i=1}^{n}[Y_i-\hat{Y}(X_i)]^2$$

$$SSTO = SSR + SSE$$

式中**總變異SSTO**表示不考慮迴歸關係下，個別Y與樣本平均值之差異程度，

$$SSTO=\sum_{i=1}^{n}[Y_i-\bar{Y}]^2=\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2$$

迴歸式的**可解釋差異，又稱回歸變異(Regression Sum of Squares)**

$$SSR=\sum_{i=1}^{n}[\hat{Y}(X_i)-\bar{Y}]^2=b\left[\sum_{i=1}^{n}(X_i-X))Y_i-Y)\right]$$

未解釋的差異係指離差值部份，用以測度樣本的誤差量，故又稱為誤差平方和SSE (Error Sum of Squares)

$$SSE=\sum_{i=1}^{n}[Y_i-\hat{Y}(X_i)]^2=SSTO-SSR$$

已解釋 = 總變異 − 未解釋的變異

$$SSR = SSTO - SSE$$

當SSR=0即SSE為最大(=SSTO)，表示回歸模式之解釋能力等於0 。

當SSE=0即SSR為最大(=SSTO)，表示總變異全部可以由**可解釋變異**來說明，回歸模式之解釋能力最大。

因此:SSR佔SSTO的比例，可作為衡量回歸關係強弱之指標(r^2)。

$$r^2=\frac{已解釋變異}{總變異}=\frac{SSTO-SSE}{SSTO}=\frac{\sum[Y-\bar{Y}]^2-\sum[Y-\hat{Y}(X)]^2}{\sum[Y-\bar{Y}]^2}$$

$$r^2=\frac{已解釋變異}{總變異}=1-\frac{\sum[Y-\hat{Y}(X)]^2}{\sum[Y-\bar{Y}]^2}$$

$$r^2=1-\frac{S^2_{Y\cdot X}}{S^2_Y}\left(\frac{n-2}{n-1}\right) \qquad r^2=1-\frac{(13.51)^2}{(33.99)^2}\left(\frac{25-2}{25-1}\right)=0.849$$

判定係數與相關係數間的關係

判定係數＝(相關係數)^2

12-3 迴歸分析的統計推論

◎迴歸分析的預測與信賴區間

◎平均數的信賴區間

母體平均數的信賴區間之估計式為

$$\mu=\bar{X}\pm t_{a/2}\frac{s}{\sqrt{n}}$$

在特定X值下Y的條件平均數的預測區間可表為下式：

$$\mu_{Y\cdot X}=\hat{Y}(X)\pm t_{a/2}估計的\sigma_{\hat{Y}(X)}$$

$\sigma_{\hat{Y}(X)}$ 為 $\hat{Y}(X)$ 的標準誤，係用以表示在特定X值下 Y 可能的變異量。

$\sigma_{\hat{Y}(X)}$ 的變異量包含以下的兩個部分：

$$\sigma^2_{\hat{Y}(X)}=Y平均數的變異量+X與\bar{X}所造成的變異量$$

第一部份之變異來源為樣本平均數，其變異取決於母體標準差與樣本大小。

第二部份之變異來自個別 X與$\bar{X}$ 間之差距(離差)。

小樣本時條件平均(期望)反應值的100(1-α)%信賴區間為

$$\mu_{Y\cdot X}=\hat{Y}(X)\pm t_{a/2}s_{Y\cdot X}\sqrt{\frac{1}{n}+\frac{(X-\bar{X})^2}{\sum X^2-(1/n)(\sum X)^2}}$$

1/n用以代表第一部份之變異。

$(X-\bar{X})$用以代表第二部份之變異。

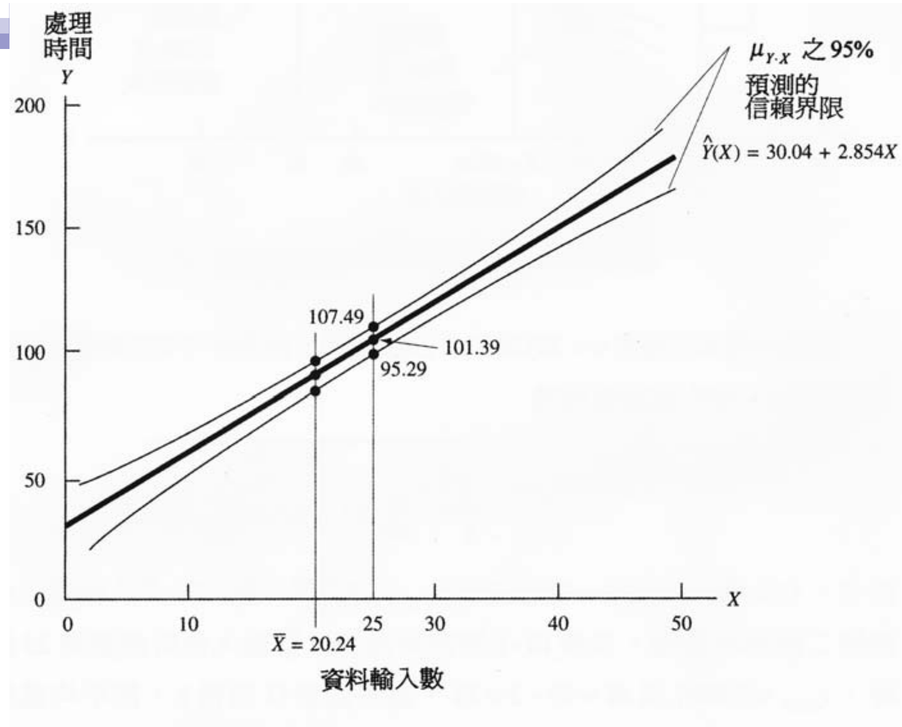


圖 12-6 反應值平均數預測值之信賴界限的圖示

◎ 特定X值下單一反應值Y的預測區間

Y之100(1-α)%預測區間為

$$Y_i(X) = \hat{Y}(X) \pm t_{\alpha/2} s_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{\sum X^2 - (1/n)(\sum X)^2} + 1}$$