第二章 迴歸分析

迴歸分析的目的通常是要推估隨機應變數 y，迴歸分析是探討某個隨機應變數(dependent variable) y 與影響變數(explanatory variables)或自變數(independent variables) x 之間的函數關係（迴歸方程式）。迴歸分析的基礎，是自變數和應變數之間具備某種統計關係。迴歸分析的工作是要找出可以有效推估隨機應變數 y 的影響變數（可能為多變數），以及推估的不確定性。

2.1 相關係數和統計檢定

協變異性指的是兩個隨機變數之間的統計相關關係，通常用二階動差、相關係數或聯合機率密度函數表示。相關係數是兩個變數之間關係強度的一種數學表示法，所有測度方法均將相關係數轉為無因次，介於 -1 與 $+1$ 之間的數值。相關係數為零，表示兩種變數不相關；若相關係數等於 -1 或 $+1$，表示兩種變數完全關連。一般利用相關係數評量的「關係強度」有兩種形式，一為評量應變數與影響變數之間是否為「單調變化」(monotonic)或是「非單調變化」(non-monotonic)的相關關係，常用方法是 Kendall 相關係數；另一種是評量兩者之間的關係是線性(linearity)或是非線性(non-linearity)的相關關係，主要方法是 Pearson 相關係數，又稱為線性相關係數。

一、Kendall 相關係數和統計測試(非參數化分析法)

Kendall 相關係數又稱為 Kendall's τ。Kendall 相關係數是一種利用排序方法計算的相關係數，為非參數化的統計值。與其他非參數化方法類似，Kendall 相關係數受極端值的影響度較低，適於分析高偏度係數與高峰度係數(coefficient of kurtosis)的資料，並且可以量測非線性變數的相關情形。

Kendall 相關係數的計算方法如下：

1. 將 n 組資料對(x,y)，按 x 值的大小排列，使 x_1 為最小、x_n 為最大，排列結果為 $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$。其中，$x$ 為自變數或影響變數，y 是應變數。
2. 檢視所有的 $n(n\text{-}1)/2$ 個 y_i 與 y_j 的比較，計算當序號 $i > j$ 時 $y_i > y_j$ 的個數

P，以及 $y_i < y_j$ 的個數 M，令 $S = P\text{-}M$。

3. Kendall 相關係數 τ：$\tau_a = S/\left[n(n-1)/2\right]$ 或 $\tau_b = S/\sqrt{(D-T_x)(D-T_y)}$。

τ_a 是無論自變數 x 或應變數 y 均無數值相等情形的計算公式，τ_b 是自變數 x 或應變數 y 有數值相等情形的 Kendall 相關係數計算公式；如果有數值相等的情形，τ_b 才是計算 Kendall 相關係數的正確公式；即 τ_a 是 τ_b 的特例，在完全沒有任何數值相等情形下，$\tau_a = \tau_b$。τ_b 中，$D = n(n-1)/2$，$T_x = \sum_{i=1}^{n} t_{xi}(t_{xi}-1)/2$、$T_y = \sum_{i=1}^{n} t_{yi}(t_{yi}-1)/2$，$t_{xi}$ 和 t_{yi} 分別為 x 和 y 中，相同數值出現相同數字的個數。例如 13 筆 x 資料樣本由小到大排為：5, 5, 6, 7, 8, 8, 8, 10, 10, 11, 12, 12, 14。第一組$(i=1)$相同數字是 5，出現 2 次，$t_{x1} = 2$；第二組$(i=2)$相同數字是 8，出現 3 次，$t_{x2} = 3$；第三組$(i=3)$相同數字是 10，出現 2 次，$t_{x3} = 2$；第四組$(i=4)$相同數字是 12，出現 2 次，$t_{x4} = 2$。因此，$T_x = \left[2\times 1/2\right]_{i=1} + \left[3\times 2/2\right]_{i=2} + \left[2\times 1/2\right]_{i=3} + \left[2\times 1/2\right]_{i=4} = 6$。

Kendall 變數相關統計測試是測試兩個變數彼此為相關或是不相關，虛無假設是不相關，即 $\tau = 0$；替代假設是兩者相關，$\tau \neq 0$。測試步驟如下：

1. Kendall 變數相關統計測試的測試統計值為 $S = P - M$。

當觀測資料數 n>10，測試統計值接近常態分佈。計算標準常態分佈的測試統計值：

$$z = \begin{cases} (S-1)/\sqrt{var(S)} & \text{if} \quad S > 0 \\ 0 & \text{if} \quad S = 0 \\ (S+1)/\sqrt{var(S)} & \text{if} \quad S < 0 \end{cases}$$

其中，$\boldsymbol{var}(S)$ 的數值，同樣受到數值有無重複出現的影響，分為無重複數值的 $\boldsymbol{var}(S_a)$，和有數值相等情形的計算公式 $\boldsymbol{var}(S_b)$：

$$\boldsymbol{var}(S_a) = n(n-1)(2n+5)/18$$

$$\begin{aligned} \boldsymbol{var}(S_b) = {} & \left[n(n-1)(2n+5) - \sum_{i=1}^{n} t_{xi}(t_{xi}-1)(2t_{xi}+5) - \sum_{i=1}^{n} t_{yi}(t_{yi}-1)(2t_{yi}+5)\right]/18 \\ & + \left[\sum_{i=1}^{n} t_{xi}(t_{xi}-1)(t_{xi}-2)\right]\cdot\left[\sum_{i=1}^{n} t_{yi}(t_{yi}-1)(t_{yi}-2)\right]/9n(n-1)(n-2) \\ & + \left[\sum_{i=1}^{n} t_{xi}(t_{xi}-1)\right]\cdot\left[\sum_{i=1}^{n} t_{yi}(t_{yi}-1)\right]/2n(n-1) \end{aligned}$$

2. 當標準化測試統計值$|z| > z_{1-\alpha/2}$時，排斥虛無假設，接受替代假設。

當自變數或影響變數 x 為時間，則 Kendall 相關係數測試即相當於時間趨勢測試 (test for trend)，其中最有用的測試是「季節 Kendall 測試」。將一年的時間分為 n 個季節，例如 n=2, 3, 4 或是 6 個季節，計算每個季節的測試統計值 s_i，$i = 1 \sim n$。「季節 Kendall 測試」的測試統計值為 $S' = \sum_{i=1}^{n} s_i$、$var(S') = \sum_{i=1}^{n} var(s_i)$。「季節 Kendall 測試」可檢驗是否有年際的長期趨勢存在，而不會受到一年之內季節起伏變化量的混淆。但此測試的缺點是對於部份季節具有逐年增加趨勢、部份季節具有逐年遞減趨勢的情形不敏感。

二、線性相關係數和統計測試

「皮爾森相關係數」r 是一個專門量測「線性相關」的係數，因此又稱為「線性相關係數」。若影響變數與應變數的所有配對均落在一直線上，則 $r = +1$ (當 x 增加時 y 亦增加)或 $r = -1$(當 x 增加時 y 會減少)；在此情形下，Kendall 的排序法相關係數 τ 值同樣為 $|\tau| = 1$。假若 x 與 y 的關係是單調、非線性的，同時所有點落在一個非線性的曲線上，Kendall 的排序法相關係數 τ 值仍將保持 $|\tau| = 1$，但皮爾森相關係數便會變為 $|r| < 1$。

和 Kendall 相關係數相同，線性相關係數亦可用來測試兩組資料是否相互獨立。統計假設測試的虛無假設是：隨機變數 y 與 x 為線性獨立(linearly independent)，並且隨機變數 y 是常態分佈。測試統計值為 $T = r\sqrt{n-2}/\sqrt{1-r^2}$，若 $|t| > t_{1-\alpha/2,n-2}$ 便排斥虛無假設、接受替代假設。由於 $r = b_1\sqrt{S_{xx}/S_{yy}}$，因此相同的檢定方法可以用來檢定 b_1 係數（斜率）是否顯著，檢定測試的內容為：

虛無假設 H_0：$\beta_1 = 0$

替代假設 H_1：$\beta_1 \neq 0$

測試統計值 $T = b_1\sqrt{S_{xx}}/s$，T 的自由度 $\nu = n-2$；

若 $|t| > t_{1-\alpha/2,n-2}$ 便排斥虛無假設、接受替代假設。

2.2 線性迴歸法

單一影響變數的線性迴歸方程式可寫為：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, 3, \ldots, n$$

其中，y_i 為應變數；x_i 為影響變數；β_0、β_1 為迴歸係數，β_0 是截距，β_1 為斜率；ε_i 為應變數的實際觀測 y_i 與利用 x_i 及迴歸方程式估計的 $\hat{y}_i$ 之間的差距，$\varepsilon_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$。

迴歸方程式的另一種表示方法，是條件機率表示法。即當 $x = x_0$ 時，y 的平均值與變異數可以條件機率分別表示為：

$$\boldsymbol{E}(\mathrm{Y}|x_0) = \beta_0 + \beta_1 x_0 \qquad \boldsymbol{var}(\mathrm{Y}|x_0) = \sigma^2$$

一、估計迴歸係數

迴歸方法假設 ε_i 的分佈不隨影響變數 x 而變，基於此定常性假設(stationarity)乃可以將不同影響變數 x_i 對應的誤差值 ε_i 放在一起，找尋使誤差平方和為最小的 β_0 和 β_1 值。

目標函數：$\boldsymbol{min}\left\{L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right\}$

求最佳解的方法是聯立解以下方程組：

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)\cdot(-1) = 0$$

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)\cdot(-x_i) = 0$$

以上二聯立方程式可以改寫為矩陣式如下：

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

除了上述以「最小誤差平方和」原則，建立迴歸係數的線性聯立方程組，再聯立解取得迴歸係數的方式外；另一種方式，是採用「不偏估」與「最小誤差平方和」兩原則的組合，亦可得到完全相同的線性迴歸係數。不偏估原則將 X、Y 皆視為隨機變數，要求滿足：

$$E[Y] = \beta_0 + \beta_1 E[X] \quad \text{即} \quad b_0 = \bar{y} - b_1 \bar{x}$$

表2-1迴歸方程式中用到的統計值、計算公式與符號

統 計 值 名 稱	計 算 公 式
x 平均值	$\bar{x}=\sum_{i=1}^{n}x_i\big/n$
y 平均值	$\bar{y}=\sum_{i=1}^{n}y_i\big/n$
x 差值平方和	$S_{xx}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$
y 差值平方和	$S_{yy}=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$
x-y 差值乘積和	$S_{xy}=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)=\left[\sum_{i=1}^{n}x_iy_i\right]-n\overline{xy}$
斜率 β_1 的樣本估計值	$b_1=S_{xy}\big/S_{xx}$
截距 β_0 的樣本估計值	$b_0=\bar{y}-b_1\bar{x}$
y_i 估計值	$\hat{y}_i=b_0+b_1x_i$
估計殘餘值	$e_i=y_i-\hat{y}_i$
樣本估計誤差變異數	$s^2=\left(S_{yy}-b_1S_{xy}\right)\big/\left(n-2\right)=\left[\sum_{i=1}^{n}e_i^{\,2}\right]\big/\left(n-2\right)$
誤差平方和	$S_{ee}=\sum_{i=1}^{n}e_i^{\,2}$
迴歸估計標準偏差	$s=\sqrt{s^2}$
線性相關係數	$r=S_{xy}\big/\sqrt{S_{xx}S_{yy}}=b_1\sqrt{S_{xx}\big/S_{yy}}$
迴歸所解釋的變異量係數	$R^2=\left[S_{yy}-s^2(n-2)\right]\big/S_{yy}=1-\left(S_{ee}\big/S_{yy}\right)=r^2$

上述結果與表 2.1 中 b_0 的方程式相同。其次，再將不偏估條件代入迴歸方程式得到 $\hat{y}_i-\bar{y}=b_1\left(x_i-\bar{x}\right)$，化簡為 b_1 是唯一參數的最小化問題：

$$\boldsymbol{min}\left\{L=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left[y_i-\bar{y}-\beta_1\left(x_i-\bar{x}\right)\right]^2\right\}$$

$$\frac{\partial L}{\partial\beta_1}=\sum_{i=1}^{n}2\left[\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)-\beta_1\left(x_i-\bar{x}\right)^2\right]\cdot(-1)=0$$

由上式得到 $b_1=S_{xy}\big/S_{xx}$ 的結果，與單獨使用誤差平方和最小化原則求 β_0、β_1 兩參數的結果相同。

二、開發線性迴歸模式的步驟

1. 繪出資料的(*x*,*y*)資料散佈圖(data scatter plot)，觀察兩項重要事實，第一項是資料是否呈線性變化？第二項是資料的變異數是否隨影響變數值而改

變 (同分佈態 homoscedastic vs. 異分佈態 heteroscedastic)？若資料呈現非線性變化、同分佈態，則可對自變數 x 作函數轉換。常用的轉換方法，是嘗試不同次冪的轉換，$z = x^p$，例如 $p = -2, -1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{2}, 1, 2$ 等，觀察轉換後的(z,y)資料散佈圖，選擇最接近線性與同分佈態的結果。（以上次冪轉換中 $p = 0$ 的轉換，僅代表對數轉換，$z = \ln(x)$）。若資料呈現異分佈態的情形，則必須對應變數 y 作轉換，轉換應變數有特別的考量，詳情見下節。

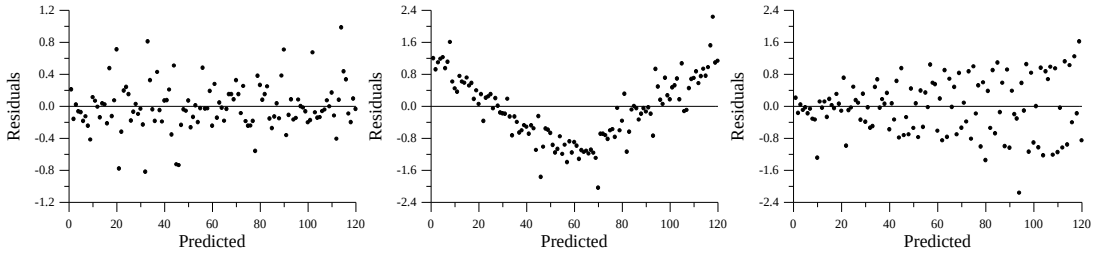
2. 估計迴歸模式的迴歸係數值，b_0 和 b_1。

3. 使用 t 測試判斷迴歸模式中的斜率 b_1 是否顯著。對於單一影響變數的迴歸方程式而言，斜率 b_1 是否顯著的測試和以上的線性相關係數顯著測試相同。若檢定結果顯示斜率 b_1 不顯著，則使用 $\hat{y} = b_0 = \bar{y}$ 的方程式估計 y。

4. 計算應變數殘餘值，$\varepsilon_i = y_i - b_0 - b_1 x_i$，繪 (x_i, ε_i) 殘餘值散佈圖。殘餘值散佈圖應該顯示：**平均值為零、沒有規則曲率**，以及**誤差變異數不隨影響變數數值而改變**等三項特性。圖 2-1(a)、2-1(b)與 2-1(c)顯示三種不同的殘餘值散佈圖，其中圖 2-1(a)是正常分佈的殘餘值，顯示資料沒有非線性分佈的情形，同時變異數不隨影響變數數值而改變；圖 2-1(b)顯示(x,y)資料具有非線性關係，必須先進行資料轉換；圖 2-1(c)顯示殘餘值變異數有隨影響變數數值變化的異分佈態情形。若有非線性的情形，則必須回到第(1)步驟，進行自變數轉換。若無法取得有效的迴歸模式，可考慮使用單一變數多項式迴歸模式，或是多變數迴歸模式。若有「異分佈態」的情形，也可以使用「權重最小方差法」(Weighted least-square)建立模式。

5. 繪出殘餘值的機率分佈梯狀圖、累積機率圖或是方盒圖，以了解殘餘值是否約呈常態分佈。一般估計迴歸模式的信心區間時，是假設殘餘值呈常態分佈，因此若殘餘值的機率分佈與常態分佈相去甚遠，則在建立信心區間時必須將殘餘值的分佈考慮在內。

6. 將殘餘值按資料性質、x 值範圍分割區間、y 值範圍分割區間，將資料分為數類或數個區間，分別繪 side-by-side 方盒圖或梯狀圖，以了解其對於 x 或 y 或是隨資料類別變化的趨勢。



(a)正常分佈的殘餘值　　(b)非線性分佈的殘餘值　　(c)異分佈態的殘餘值

圖2-1 三種不同的殘餘值散佈圖

三、信心區間與預報區間

預報區間（prediction interval）：根據資料樣本，推估**可觀測的隨機變數**，未來**個別樣本**實現值，某指定出現機率的數值範圍。

信心區間（confidence interval）：在指定機率下，**推估不可觀測的某統計參數的出現範圍，例如變數的真實平均值、迴歸係數等。**

當(*x*,*y*)資料確實為線性關係，殘餘值呈常態分佈並且變異數不隨影響變數數值改變時，則迴歸係數 β_0、β_1 的信心區間可以估計，對於斜率 β_1 而言，$(1-\alpha)$的信心區間為：

$$\left(b_1 - \frac{s \cdot t_{1-\alpha/2,n-2}}{\sqrt{S_{xx}}},\ b_1 + \frac{s \cdot t_{1-\alpha/2,n-2}}{\sqrt{S_{xx}}} \right)$$

截距 β_0 的 $(1-\alpha)$ 信心區間為：

$$\left(b_0 - s \cdot t_{1-\alpha/2,n-2}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} \quad , \quad b_0 + s \cdot t_{1-\alpha/2,n-2}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} \right)$$

此外，若已知影響變數值為 x_0，則**應變數 *y* 的條件平均值**為 $\hat{y} = E(y|x_0) = b_0 + b_1 x_0$，**條件平均值**的(1−α)信心區間為：

$$\left(\hat{y} - s \cdot t_{1-\alpha/2,n-2}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}},\ \hat{y} + s \cdot t_{1-\alpha/2,n-2}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} \right)$$

若已知影響變數值為 x_0，對於任意 *y* 值的預報區間(prediction interval)而非條件平均值，其不確定性包括參數的不確定性，以及迴歸模式所不能解

釋的殘餘值變異性(variability)：

$$\left(\hat{y} - s \cdot t_{1-\alpha/2,n-2} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\ \hat{y} + s \cdot t_{1-\alpha/2,n-2} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

圖 2-2 為以西元 1960 年至 1996 年淡水雨量站年雨量記錄(自變數 x)對基隆雨量站年雨量記錄(應變數 y)線性迴歸結果，以及應變數 y 值的 95%預報區間與 y 值條件平均值的 95%信心區間的示意圖。

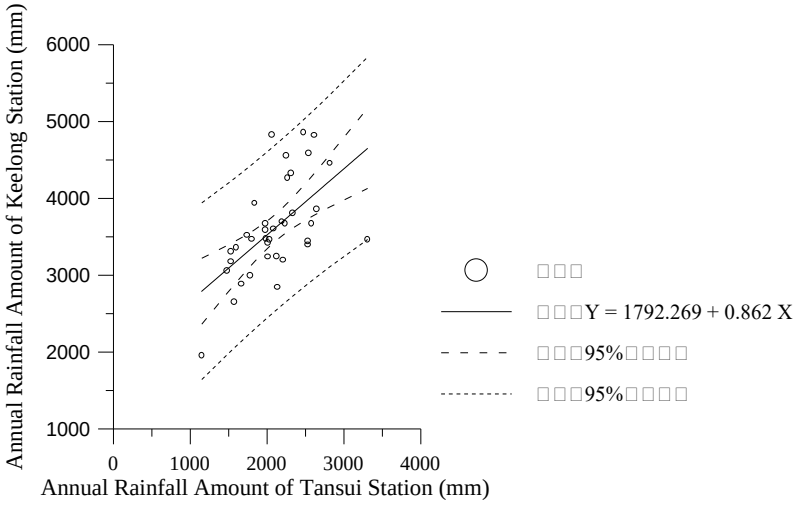


圖2-2 西元 1960 年至 1996 年淡水雨量站年雨量記錄對基隆雨量站年雨量記錄之線性迴歸結果及基隆站的迴歸線 95%信心區間與資料點 95%預報區間示意圖

2.3 應變數 y 的函數轉換

當應變數 y 經過某種函數轉換，迴歸方程式是 F(y) 對於影響變數 x 或是 G(x) 的線性迴歸，在作反函數轉換時，$y = \mathrm{F}^{-1}(\cdot)$，必須給予特殊考慮，尤其是當應變數 y 是時間序列資料，必須將應變數 y 對於時間積分時，更是必需特別注意。舉例說明，令 L 代表水中的懸浮物總重量，Q 為流量，兩者均經過對數轉換後，得到一線性關係：

$$\ln L = \beta_0 + \beta_1 \ln Q + \varepsilon$$

其中 L 的單位是 tons/day，流量 Q 的單位是 cfs。假設兩者的資料分佈狀況、迴歸線與 50%、95%預報區間(prediction interval)的結果如圖 2-3，將 ln$L$ 與 ln$Q$ 反轉換為 L 與 Q 得到圖 2-4。

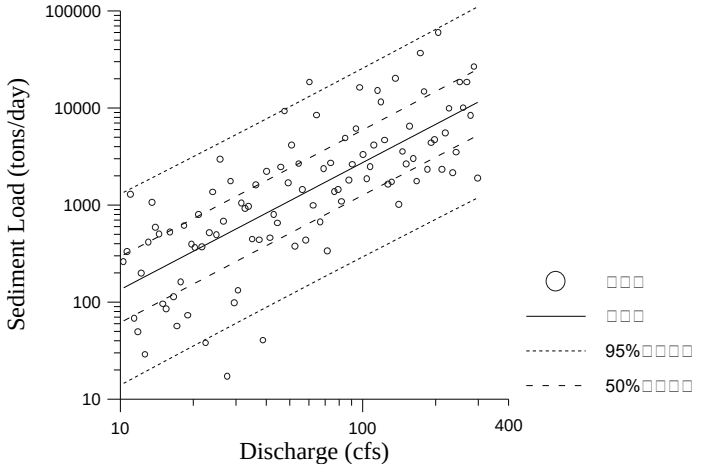


圖2-3 以對數座標系統求得流量對水中懸浮物總重量之線性迴歸結果及懸浮物總重量之 50%與 95%預報區間例

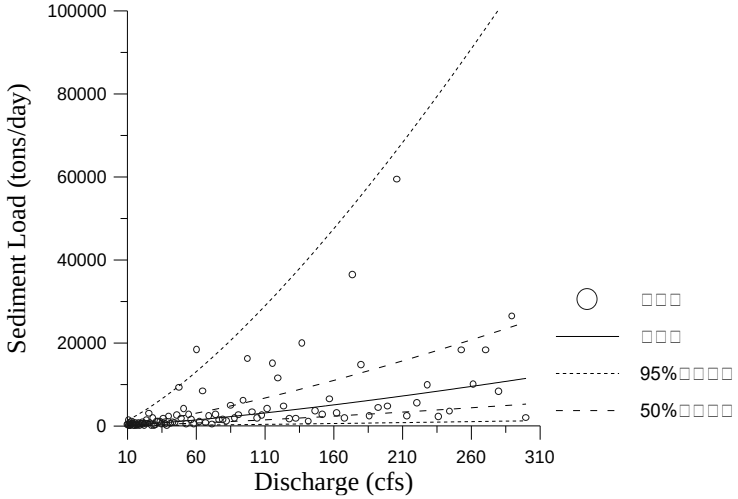


圖2-4 經對數反轉換後之流量對水中懸浮物總重量之迴歸線及懸浮物總重量之 50%與 95%預報區間例

當隨機變數呈常態分佈時，因為分佈對稱的關係，變數的中值與平均值相等：

$$(\ln L)_{50} = \mu_{\ln L}$$

所以在圖 2-3 中的迴歸線同時代表平均值與中值。圖 2-4 中的點與線，是圖 2-3 中點、線的指數轉換。對於任何給定的流量，L 的條件分佈是對數常態分佈，而非常態分佈。由於中值是數值按大小排序的中央排序數值，將隨機變數經過指數函數的單調函數轉換並不改變中值的順序，因此對數常態分佈變數的中值數值等於常態分佈中值的指數值：

$$L_{50} = exp\left[(\ln L)_{50}\right]$$

而對數常態分佈的變數平均值會大於變數取對數後平均值的指數值：

$$\mu_L > exp\left[\mu_{\ln L}\right] = exp\left[(\ln L)_{50}\right] = L_{50}$$

因此對數常態分佈的平均值會大於其中值。因為此一轉換特性，若採用一年的小時流量記錄帶入迴歸式，估計累積輸砂量，得到：

$$(\Sigma L)' = \sum_{d=1}^{365}\sum_{h=1}^{24} exp\left(\beta_0 + \beta_1 \ln Q_{h,d}\right) = \sum_{d=1}^{365}\sum_{h=1}^{24} exp\left[(\ln L)_{50}\right]_{h,d}$$

則因為$(\Sigma L)'$是個別條件中值的累積值，會低於 L 的累積值；意即$(\Sigma L)'$是一個偏估的估計值(biased estimate)，傾向低估 L 的累積值：

$$(\Sigma L)' < (\Sigma L)$$

以上反轉換的偏估問題的一種解決方法，是在樣本數量較大(n>30)並且對數座標中殘餘值變異數較小(σ^2<0.5)的情形下，假設迴歸係數值β_0、β_1為已知(true parameter or known without error)，以及在對數座標中殘餘值呈常態分佈，則使用以下的估計方式，可得一個「**不偏估的平均值估計**」：

$$\hat{L}_F = exp\left(b_0 + b_1 \ln Q + \tfrac{1}{2}s^2\right)$$

但在不符合以上多種假設的情形下，上式可能會出現過度補償偏估的情形，造成高估的偏估情形。另一種解決方式是使用「補償平均值估計」(smearing estimate)，此一方法僅需假設所有的殘餘值為相互獨立且來自同一分佈的假設(i.i.d. or homoscedastic)，殘餘值可以是常態分佈以外的其他分佈。在對數轉換的情形下，「**補償平均值估計**」的表示式為：

$$\hat{L}_D = \boldsymbol{exp}\left(b_0 + b_1 \boldsymbol{\ln} Q\right) \cdot \tfrac{1}{n}\sum_{i=1}^{n} \boldsymbol{exp}(\varepsilon_i)$$

當殘餘值分佈為常態分佈時，「補償平均值估計」與「不偏估的平均值估計」，$\hat{L}_F$，的表現相當；在殘餘值分佈不是常態分佈時，「補償平均值估計」的表現優於「不偏估的平均值估計」，因此在一般情形下，建議使用「補償平均值估計」。除了以上對數轉換的「補償平均值估計」形式以外，對於任何單調函數轉換(例如開根號、倒數、指數或對數等)，「補償平均值

估計」的表示法為：

$$\hat{Y}_D = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{F}^{-1}\left(b_0 + b_1 X + \varepsilon_i\right)$$

其中，Y 為轉換前的原始應變數；X 為自變數；$y = F(Y) = b_0 + b_1 X + \varepsilon$ 中的 F(·) 為轉換函數，$F^{-1}(\cdot)$ 為其反函數。修正值不會因為 X 變數的樣本資料分布與母群分布差異而不同；當樣本資料數量 n 愈大，則 $\hat{Y}_D$ 愈趨近於真值 Y_D。原因是每個轉換前的誤差，都是來自相同的常態分布，且和影響變數值無關；因此，影響變數樣本分布不會造成修正值的差異，誤差分佈型態差異才會。

2.4 多變數迴歸

任意水文應變數可能受到不只一種影響變數的影響，在迴歸式中增加影響變數通常會降低估計誤差的變異數。多變數的線性迴歸方程式表示法為：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

以上方程式的係數，$b_0 = \bar{y} - \sum_{i=1}^{k} b_i \bar{x}_i$ ；b_i ，$i = 1 \sim k$ 為以下矩陣式的解：

$$\begin{bmatrix} \sum\left(x_{1_i} - \bar{x}_1\right)\left(x_{1_i} - \bar{x}_1\right) & \sum\left(x_{2_i} - \bar{x}_2\right)\left(x_{1_i} - \bar{x}_1\right) & \cdots & \sum\left(x_{k_i} - \bar{x}_k\right)\left(x_{1_i} - \bar{x}_1\right) \\ \sum\left(x_{1_i} - \bar{x}_1\right)\left(x_{2_i} - \bar{x}_2\right) & \sum\left(x_{2_i} - \bar{x}_2\right)\left(x_{2_i} - \bar{x}_2\right) & \cdots & \sum\left(x_{k_i} - \bar{x}_k\right)\left(x_{2_i} - \bar{x}_2\right) \\ \vdots & \vdots & \ddots & \vdots \\ \sum\left(x_{1_i} - \bar{x}_1\right)\left(x_{k_i} - \bar{x}_k\right) & \sum\left(x_{2_i} - \bar{x}_2\right)\left(x_{k_i} - \bar{x}_k\right) & \cdots & \sum\left(x_{k_i} - \bar{x}_k\right)\left(x_{k_i} - \bar{x}_k\right) \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum\left(x_{1_i} - \bar{x}_1\right)\left(y_i - \bar{y}\right) \\ \sum\left(x_{2_i} - \bar{x}_2\right)\left(y_i - \bar{y}\right) \\ \vdots \\ \sum\left(x_{k_i} - \bar{x}_k\right)\left(y_i - \bar{y}\right) \end{bmatrix}$$

例如某種水中溶質的濃度與流量(Q)及時間(T)有關，迴歸關係形式如下：

$$C = \beta_0 + \beta_1 \boldsymbol{ln} Q + \beta_2 \left(\boldsymbol{ln} Q\right)^2 + \beta_3 T + \varepsilon$$

上式比假設濃度僅隨時間單一變數變化的迴歸關係更具有解釋能力(more efficient)，因為後者將濃度因流量變化的部份由時間變數來解釋，會大幅減少對濃度的解釋能力。

不同次冪的同一變數多項式可以用作為不同的影響變數，以解釋非線性的變化趨勢。在此一情形下，建議使用減去平均值的形式，以避免不同次冪變數之間的多變數線性相關性(multi-colinearity)，造成迴歸係數估計與迴歸統計值解釋的問題：

$$y = \beta_0 + \beta_1\left(x - \bar{x}\right) + \beta_2\left(x - \bar{x}\right)^2 + \beta_3\left(x - \bar{x}\right)^3 + \varepsilon$$

使用正交的多項式(orthogonal polynomial)可以完全解決以上不同次冪變數之間的相關性問題，但減去平均值形式的多項式通常已經足夠。

許多水文變數是時間上的週期函數，例如降雨的年週期效應、溫度的日差效應(diurnal cycle)、潮汐等，週期變數可以使用正弦函數與餘弦函數作為迴歸的自變數：

$$y = \beta_0 + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \varepsilon$$

若應變數無法由以上簡單週期函數描述，可以增加使用週期的整數倍數的正弦與餘弦函數：

$$y = \beta_0 + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \beta_3 \sin(4\pi t) + \beta_4 \cos(4\pi t) + \varepsilon$$

例如美國東部 Chesapeake 海灣由上游流域帶來的多種水中養分負荷，均可用以下綜合時間趨勢、流量與週期等三種影響變數的迴歸方程式有效描述：

$$\ln C = \beta_0 + \beta_1 \ln Q + \beta_2 \ln^2 Q + \beta_3 t + \beta_4 t^2 + \beta_5 \sin(2\pi t) + \beta_6 \cos(2\pi t) + \varepsilon$$

同樣的六變數迴歸方程式被用來迴歸不同的水中養分濃度，而不管係數的值是否顯著非零。

§ 假變數-整合分類迴歸的工具

另一類的影響變數是只有 0 與 1 兩種數值的假變數(pseudo variables)，使用假變數的迴歸稱為共變數分析法(Analysis of Covariance)。共變數分析法適用於分析類別化變數(categorical variables)，例如某變數的白天與晚上的數值、汙水處理廠的有無等。在某一種類別時，令假變數的數值為 0，在另一種類別時，令假變數的數值為 1。使用假變數共變數分析法的兩個應用例子如下：

(1) $y = \beta_0 + \beta_1 x + \beta_2 Z + \varepsilon$

(2) $y = \beta_0 + \beta_1 x + \beta_2 Z + \beta_3 xZ + \varepsilon$

$$y = \beta_0 + \beta_1 x + \beta_3 (x - x_0) Z + \varepsilon$$

其中 Z 為假變數，x 為連續的影響變數。以上模式(1)中，當假變數數值由 0 變為 1 時，增加了一個截距 β_2，有如由下圖(圖 2-5)中的虛線變為細實線；模式(2)當假變數數值由 0 變為 1 時，除了增加了一個截距 β_2 以外，還增加一個斜率，使斜率由原來的 β_1 變為 $\beta_1+\beta_3$，有如由圖 2-5 中的虛線變為粗實線。使用共變數分析法，其實是在一個迴歸方程式中，包含兩個迴歸方程式，經由對包含假變數 Z 的迴歸係數作顯著統計測試，可決定在不同類別下，變數的迴歸方程式是否相同。

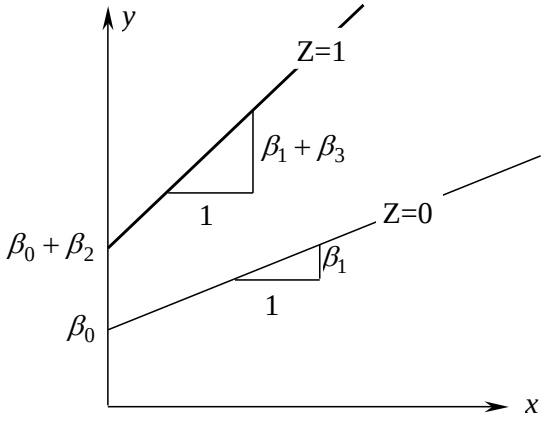


圖2-5 包含假變數迴歸模型示意圖。細線和粗線分別為假變數 Z=0 與 Z=1 時 $y=\beta_0+\beta_1x+\beta_2Z+\beta_3xZ+\varepsilon$ 模式的迴歸線位置

2.5 多變數迴歸方程式的評估

使用多變數迴歸方程式，永遠必須將殘餘值對於應變數繪圖，以了解是否有曲率存在，同時殘餘值是否符合「**同分佈態**」假設。除此以外，多變數迴歸還必須做「**多變數線性相關測試**」與「**個別觀測資料測試**」。

一、多變數線性相關測試

多變數迴歸方程式中，至少有一個影響變數與其他某一變數或某些變數的組合，具有高度線性相關性，便是具有「多變數線性相關」問題(multi-colinearity)。發生多變數線性相關問題時，迴歸係數會發生不穩定的現象，即略為變動少數觀測資料，會導致非常不同的迴歸係數。一個分析診斷多變數線性相關問題的方法是「變異數膨脹係數」法(variance inflation factor, VIF)。第 j 個影響變數的 VIF_j 的計算步驟，是：

1. 將第 j 個影響變數對於所有其他變數做迴歸，求出迴歸係數：

$$x_j = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_{j-1} x_{j-1} + \alpha_{j+1} x_{j+1} + \ldots + \alpha_k x_k + \varepsilon$$

2. 求出第 j 個影響變數對其他影響變數迴歸的「解釋變異數係數」(coefficient of determination fraction of the variance explained by regression)，r_j^2：（其中 $S_{\varepsilon\varepsilon}$ 和 S_{jj} 分別為上式中誤差 ε 的變異數和 x_j 變數的變異數）

$$r_j^2 = 1 - \frac{S_{\varepsilon\varepsilon}}{S_{jj}}$$

3. 計算第 j 個影響變數的「變異數膨脹係數」，VIF_j

$$VIF_j = \frac{1}{1 - r_j^2}$$

最理想的狀況是第 j 個影響變數與其他影響變數均不具線性相關性，或彼此為相互正交的函數，若是如此，則 VIF_j=1；反之，若第 j 個影響變數與其他影響變數高度線性相關，此時 $VIF_j \to \infty$。當 VIF_j>10，表示此一變數與其他變數的「多變數線性相關性」存在，解決此一「多變數線性相關性」問題的方式有以下數種選擇：

1. 改採用各影響變數減去其平均值的數值，作為影響變數。
2. 若「多變數線性相關性」的原因是第 j 個影響變數可由其他變數線性組合而成，將第 j 個影響變數自迴歸方程式中除去。
3. 蒐集更多與原樣本群不同的資料，以突顯第 j 個影響變數資料的獨立性與解釋能力。
4. 採用脊線迴歸(ridge regression)或主成份分析(principle component analysis)。

二、個別觀測資料測試

迴歸個別資料的測試、診斷方法有數種，診斷、測試的目的是判斷個別資料對於整個迴歸方程式係數的影響。此處介紹的方法，是應用於單一影響變數迴歸的診斷方法。

1. 槓桿統計值(leverage statistics)

槓桿統計值是描述個別觀測資料與所有觀測資料「重心」的距離，統計值是對於影響變數 x 所作的計算，非針對應變數 y 的統計值。槓桿統計值的定義為：

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \text{(單變數)} \quad \text{或} \quad h_i = \frac{1}{n} + \left[\mathbf{x}_i - \bar{\mathbf{x}}\right]^{\mathrm{T}} \mathbf{\Lambda}^{-1} \left[\mathbf{x}_i - \bar{\mathbf{x}}\right] \text{(多變數)}$$

以上多變數方程式中的 $\mathbf{\Lambda}$ 是以下矩陣，不是共變矩陣：

$$\begin{bmatrix} \sum(x_{1_i} - \bar{x}_1)(x_{1_i} - \bar{x}_1) & \sum(x_{2_i} - \bar{x}_2)(x_{1_i} - \bar{x}_1) & \cdots & \sum(x_{k_i} - \bar{x}_k)(x_{1_i} - \bar{x}_1) \\ \sum(x_{1_i} - \bar{x}_1)(x_{2_i} - \bar{x}_2) & \sum(x_{2_i} - \bar{x}_2)(x_{2_i} - \bar{x}_2) & \cdots & \sum(x_{k_i} - \bar{x}_k)(x_{2_i} - \bar{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \sum(x_{1_i} - \bar{x}_1)(x_{k_i} - \bar{x}_k) & \sum(x_{2_i} - \bar{x}_2)(x_{k_i} - \bar{x}_k) & \cdots & \sum(x_{k_i} - \bar{x}_k)(x_{k_i} - \bar{x}_k) \end{bmatrix}$$

所有觀測資料槓桿統計值的總和 Σh_i 應該等於 p。一個（或一組）觀測資料 x_i（或 $x_{1i} \sim x_{ki}$）的槓桿統計值 h_i 若大於 $2p/n$，便是具有高槓桿的觀測資料。其中 n 是觀測資料個數，p 是迴歸方程式中迴歸係數的個數，對於單一影響變數的迴歸方程式，p=2；若是 k 個影響變數的多變數迴歸，則 $p = k + 1$。具有高槓桿並不一定構成迴歸的問題，因為僅具有長的力臂並不一定發生大的力矩，還需要結合應變數的迴歸誤差或殘餘值，方可決定某特定資料對於迴歸統計的影響。

2. 標準化殘餘值(Studentized residual)

Student 化殘餘值 e_{si} 是將殘餘值 e_i 除以其標準偏差的統計值：

$$e_{si} = \frac{e_i}{s\sqrt{1 - h_i}}$$

若殘餘值 e_i 呈常態分佈，則 $|e_{si}|$>2 的發生機率應僅約為資料數量的5%，$|e_{si}|$>3 的發生機率應僅約為資料數量的 0.3%。

將一個資料自樣本群組中去除，使用 n–1 筆資料迴歸得到的迴歸係數若與原迴歸係數(用 n 組資料迴歸之結果)相差甚大，則該組資料對於迴歸係數有高度影響力。高影響力需具備兩項條件，高槓桿與大殘餘絕對值。一個診斷觀測資料影響力的統計值是 Cook's D 統計值(Cook's

measure of influence or Cook's distance)，其計算方法為：

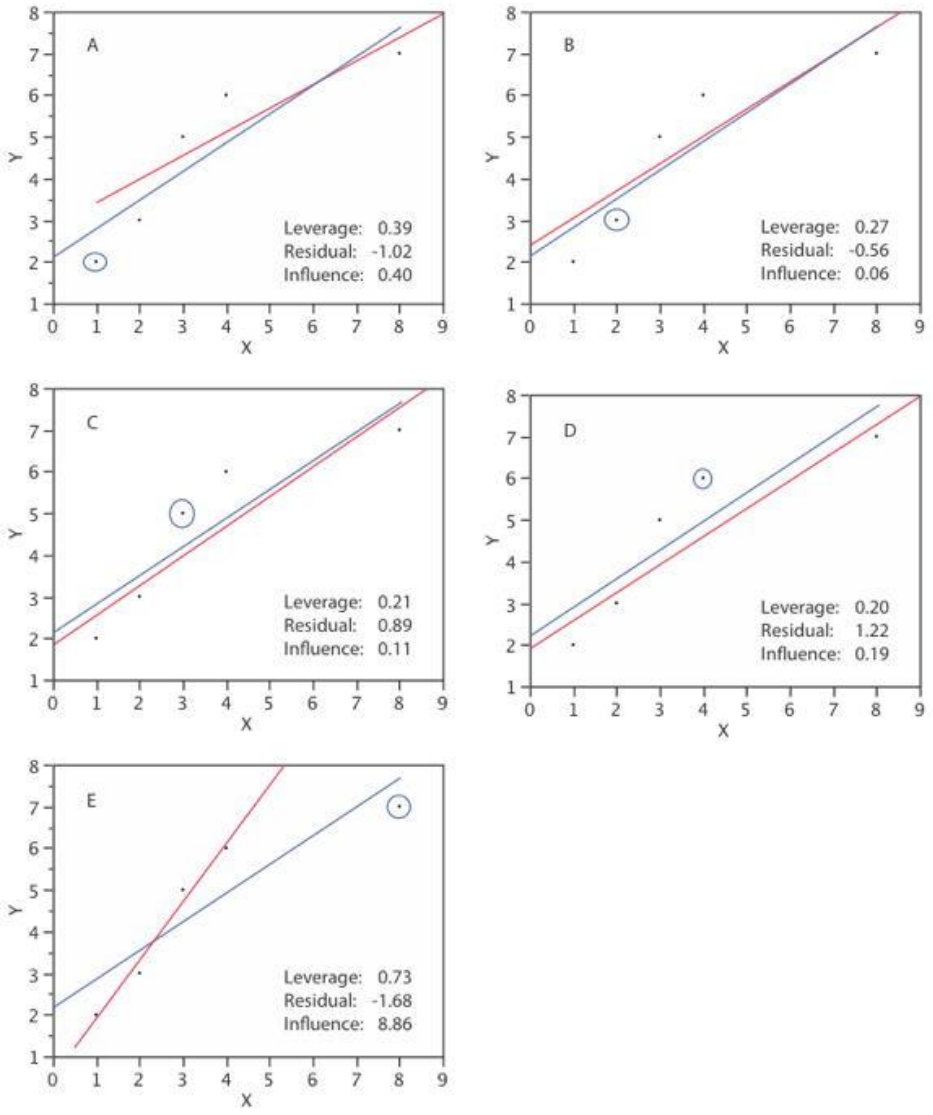
$$D_i = \frac{e_i^2 h_i}{ps^2(1-h_i)^2} = e_{si}^2 \frac{h_i}{p(1-h_i)}$$

當 α=0.5 時 $D_i > F_{p,n-p}$，則第 i 個觀測資料便視為具有高影響力。此處需注意，採用的 α 值為 0.5 而非 0.05。例如，有超過 30 組資料的單一影響變數迴歸問題中，D_i 的門檻值約為 0.7，若有數個影響變數，則門檻值約為 0.8 至 0.9；與 $F_{2,30,0.95}$=3.32 甚為不同。一個比較簡單迅速的判斷標準是若 $D_i > 1$ 則該筆資料具有高影響力。

若發現某筆資料具有高影響力，則應該檢查該筆資料是否有誤，若資料有誤，則修正或去除該筆資料；若證實資料無誤，則考慮採用更複雜的迴歸模式，包括納入其他影響變數、採用非線性項等，或是採用「權重最小方差法」求取迴歸係數。

例 2-1：利用下表中 A~E 的 5 筆數據建立單變數迴歸模式，並計算其槓桿統計值 h_i、e_{si} 和 Cook's D 統計值。資料數 $n=5$，判斷槓桿、Student 化殘餘絕對值和 Cook's D 值偏高的簡單標準，分別是 $2p/n$=0.8、2 和 1。下圖中，利用 5 個點迴歸得到的是藍色線，去除其中某點，利用其他 4 點得到的迴歸線為紅色。數據和圖顯示，其中 E 點的影響力極高，其次是 A 點，也具有較高的影響力，但未達顯著標準。（資料來源：http://onlinestatbook.com/2/regression/influential.html 但有修正其錯誤）

Point i	x_i	y_i	e_i	h_i	e_{si}	D_i
A	1	2	-0.80	0.43	-1.02	0.40
B	2	3	-0.49	0.29	-0.56	0.06
C	3	5	0.82	0.21	0.89	0.11
D	4	6	1.12	0.21	1.22	0.19
E	8	7	-0.64	0.86	-1.68	8.86



三、*PRESS* 統計值

預報殘餘值 $e_{(i)}$ 是比標準化殘餘值更有用的一種個別觀測資料診斷測試方法，其定義為 $e_{(i)} = y_i - \hat{y}_{(i)}$；其中 $\hat{y}_{(i)}$ 是使用除去資料 (x_i, y_i)，使用其他 n-1 筆觀測資料求出一組迴歸係數，將 x_i 代入此一迴歸方程式，得到的應變數估計值便是 $\hat{y}_{(i)}$。但實際上計算預報殘餘值 $e_{(i)}$ 不需要每次去除一個觀測資料進行 n 次迴歸，使用以下公式計算即可：

$$e_{(i)} = \frac{e_i}{1 - h_i}$$

利用預報殘餘值可以計算一個評估迴歸方程式品質的統計值 *PRESS*：

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2$$

PRESS 可用來比較不同的迴歸模式。若欲由數種迴歸模式中選擇一個最佳模式，可以使用 *PRESS* 統計值法，選擇 *PRESS* 統計值最小的，作為迴歸方程式。

2.6 模式選擇

模式選擇的第一步，是決定使用什麼影響變數，以及使用多少個影響變數。決定使用或不使用某個影響變數，必須具有理論上能夠成立的理由，**影響變數個數不應該大於資料數的十分之一，*n*/10。迴歸模式中增加任何解釋變數，總會使模式的解釋能力增加，或估計應變數的誤差變異數減少，決定是否增加某個影響變數的評量原則有二，一是其迴歸係數 β 是否通過顯著不為 0 的假設測試；另一是增加一個影響變數，估計應變數的誤差變異數 S_{ee} 減少的比率是否顯著。**

若使用多項式或者是周期性函數模式，則必須遵守一個原則，即**若模式包含高次項的多項式，應該包括所有此一次冪以下的所有低次冪多項式進行迴歸。若包含正弦函數，則必然要包含餘弦函數，反之亦然。**

若模式使用 *k* 個變數，理論上包括無解釋變數模式，共有 2^k 個可能的模式組合。在此所有模式當中求最佳解釋變數組合迴歸模式的方法如下：

1. 對於所有可能的模式作擬合，並且求出最佳的 *k* 個解釋變數的迴歸模式，最佳的 *k-1* 個解釋變數的迴歸模式，．．．，直到最佳的 1 個解釋變數的迴歸模式，共有 *k* 組最佳模式。所謂最佳，指的是在同樣變數個數的不同模式中，誤差變異數 S_{ee} 最小，或是解釋變異數比例 R^2 最大。在此共有 *k* 組最佳模式中，選擇最佳預測模式方法，是比較此 *k* 組模式的 *PRESS* 值，*PRESS* 值最小者即為最佳模式。
2. 對於兩種不同複雜程度的多變數迴歸模式，並且較為複雜的模式包含較為簡單的模式(巢套模式，nested model)，要決定複雜模式複雜模式增加的變數項，是否提供實質有效的額外解釋功能，可以用 *F* 測試進行檢驗。

令 *s* 代表簡單模式，*c* 代表複雜模式：

$$y=\hat{y}_s+\varepsilon_s=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\varepsilon_s$$

$$y=\hat{y}_c+\varepsilon_c=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k+\beta_{k+1}x_{k+1}+\cdots+\beta_mx_m+\varepsilon_c$$

若以上兩個模式的樣本資料個數為 n，因為簡單模式有 $k+1$ 個參數，因此有 $n-(k+1)$ 個自由度(df_s)，其殘差變異數總和為 $SSE_s=\sum\varepsilon_s^2$ ；複雜模式有共有 $m+1$ 個參數，因此有 $n-(m+1)$ 個自由度(df_c)，其殘差變異數總和為 $SSE_c=\sum\varepsilon_c^2$ ；並且 $\mathrm{df}_s-\mathrm{df}_c=m-k$ 。F 測試的虛無假設與替代假設分別為：

H_0 ： $\beta_{k+1}=\beta_{k+2}=...=\beta_m=0$

H_1 ：以上 $m-k$ 個係數中，至少有一個非零

測試的統計值 F 為： $F=\dfrac{(\mathrm{SSE}_s-\mathrm{SSE}_c)/(\mathrm{df}_s-\mathrm{df}_c)}{(\mathrm{SSE}_c/\mathrm{df}_c)}$

若是測試統計值超過表列的 $F_\alpha(m-k,\mathrm{df}_c)$ 數值，則排斥虛無假設，接受替代假設。若複雜迴歸模式只比簡單迴歸模式多一個解釋變數，即 $m=k+1$，並且巢套，則 F 測試退化為檢測 $\hat{y}_c=\hat{y}_s+\beta_{k+1}x_{k+1}$ 模型中第 $k+1$ 個影響變數 X_{k+1} 減少 Y 估計誤差變異數的 Student t 測試。利用 Student t 測試判斷 X_{k+1} 解釋變異數貢獻是否顯著的計算方法如下。

在隨機變數 Y 與 X_{k+1} 彼此線性獨立(linearly independent)的虛無假設，和隨機變數 Y 是常態分佈的假設下，統計假設測試是測試 b_{k+1} 的顯著性，測試統計值 $t=\hat{b}_{k+1}\Big/\sqrt{\boldsymbol{var}(\hat{b}_{k+1})}$ ， $\boldsymbol{var}(\hat{b}_{k+1})=\left[SSE_c/(n-k-1)\right]\cdot\boldsymbol{\Lambda}^{-1}(k+1,k+1)$ ，其中 $\boldsymbol{\Lambda}=\begin{bmatrix}\sum(x_{1_i}-\bar{x}_1)(x_{1_i}-\bar{x}_1) & \sum(x_{2_i}-\bar{x}_2)(x_{1_i}-\bar{x}_1) & \cdots & \sum(x_{k+1_i}-\bar{x}_{k+1})(x_{1_i}-\bar{x}_1)\\ \sum(x_{1_i}-\bar{x}_1)(x_{2_i}-\bar{x}_2) & \sum(x_{2_i}-\bar{x}_2)(x_{2_i}-\bar{x}_2) & \cdots & \sum(x_{k+1_i}-\bar{x}_{k+1})(x_{2_i}-\bar{x}_2)\\ \vdots & \vdots & \ddots & \vdots\\ \sum(x_{1_i}-\bar{x}_1)(x_{k+1_i}-\bar{x}_{k+1}) & \sum(x_{2_i}-\bar{x}_2)(x_{k+1_i}-\bar{x}_{k+1}) & \cdots & \sum(x_{k+1_i}-\bar{x}_{k+1})(x_{k+1_i}-\bar{x}_{k+1})\end{bmatrix}$，

$\boldsymbol{\Lambda}^{-1}(k+1,k+1)$ 為該矩陣反矩陣的第 $(k+1,k+1)$ 個元素。若 $|t|>t_{1-\alpha/2,n-k-1}$ 便排斥虛無假設、接受替代假設，意即 X_{k+1} 減少 Y 估計誤差變異數的貢獻顯著。

1. 增加的影響變數至少和其他影響變數中的一個線性相關：

$n-k$ 個自由度。統計假設測試的虛無假設是：

建立迴歸模式的相關問題與處理分析方法的總結：

2. **若問題是：檢驗自變數是否需要作轉換？**

分析方法：首先計算 R^2、*SSE* 或 *PRESS* 等統計值，選擇最佳迴歸模式；其次繪殘差圖，檢驗最佳模式是否符合(1)同分佈態(homoscedasticity)，(2)殘差為常態分佈，(3)殘差不隨自變數變化而改變趨勢等假設。

3. **若問題是：檢驗**應變**數是否需要作轉換？**

分析方法：繪圖比較何者資料(1)呈同分佈態(homoscedasticity)，(2)殘差為常態分佈，(3)殘差不隨自變數變化而改變趨勢。不使用 R^2、SSE 或 *PRESS* 等統計值，因為這些統計值在轉換與不轉換模式之間，會有不同的單位，不適合作為模式比較的標準。

4. **若問題是：解釋變異數**個數**相同的多個不同迴歸模式(每個單一影響變數均已通過有效解釋能力測試)，該如何選擇最佳模式？**

分析方法：使用 R^2、*SSE* 或 *PRESS* 等統計值選擇最佳迴歸模式。其次再繪殘差圖，進行檢驗。

5. **若問題是：在數個巢狀模式中何者為最佳模式？**

分析方法：可以採用 *F* 檢定或是 *PRESS* 統計值。若目的是模式選擇(判斷哪些自變數不具有解釋能力)或各個影響變數係數值非零，採用 *F* 檢定；若目的是了解模式估計的準確度，則採用 *PRESS* 統計值。

6. **若問題是：在數個解釋變數數量不同、並且未必是巢狀的迴歸模式中，何者為最佳模式？**

分析方法：採用 PRESS 統計值為最小的模式，即為估計準確度最佳的模式。

2.7 主成分分析（Principle Component Analysis）

主成份分析是將彼此線性相關的原始多變數，以線性權重組合原始變數的方式「轉換」為數量相同、相互獨立的新變數，稱為主成分。

假設原始變數為 m 個空間隨機變數 x_i，$i=1,\cdots,m$；在 n 個不同時間，此 m 個空間隨機變數共有 $m\times n$ 筆**配對樣本**，$x_{i,k}$，$k=1,\cdots,n$。經過主成分分析，轉換為數量相同的主成分變數 $y_{j,k}$，$j=1,\cdots,m$，$k=1,\cdots,n$。

最常見的應用，是分析空間中不同量測點隨機變數的時間序列資料，例如 m 個臨近水井或雨量站的 $m\times n$ 筆資料的變異情形。

二、分析目的：

1. **概述變數之間的關係：**分析空間多變數的共變數矩陣或相關係數矩陣，得到多變數的成分或模(component or mode)；由變異量最大的前幾個主成分向量（特徵向量）所個別呈現的空間態勢(pattern)，協助瞭解原始多變數最主要的共同變化模式。
2. **分析多變數受到環境變數的影響：**透過變異量最大的前幾個主成分向量與其他環境變數的關聯性分析，瞭解空間多變數受到哪些環境變數影響，以及影響的方式。

三、分析步驟：

1. **計算空間隨機變數的共變矩陣 $\mathbf{\Lambda}_{xx}$ 或相關係數矩陣 $\mathbf{R}_{xx}$**。
2. **計算 $\mathbf{\Lambda}_{xx}$ 或 $\mathbf{R}_{xx}$ 矩陣的特徵值與特徵向量：**解特徵方程 $\left|\lambda\mathbf{I}-\mathbf{\Lambda}_{xx}\right|=0$ 或是 $\left|\lambda\mathbf{I}-\mathbf{R}_{xx}\right|=0$，求出特徵值 λ_i，$i=1,\cdots,m$。
3. **求出對應於每個特徵值的特徵向量：**$\vec{e}_i=[e_{i1},e_{i2},\cdots,e_{im}]^T$，$i=1,\cdots,m$。
4. **計算主成分變數時間序列數值：**計算 $y_{j,k}=\vec{e}_j\cdot\vec{x}_{t=k}$，for $j=1,\cdots,m$，$k=1,\cdots,n$，共 $m\times n$ 筆主成分變數資料數值。

四、數據特性：

1. 若採用共變數矩陣 $\mathbf{\Lambda}_{xx}$ 進行主成分分析，則此 m 個空間隨機變數的變異數總和，即為 $\mathbf{\Lambda}_{xx}$ 矩陣的對角線元素數值的和 $\sum_{i=1}^{m}\sigma_i^2$，採用此做法則將各隨機變數變異量大小不同的現象納入考量，分析得到的主成分是解釋 $\sum_{i=1}^{m}\sigma_i^2$ 的分量。若採用相關係數矩陣 $\mathbf{R}_{xx}$ 進行主成分分析，則

因為已經先將變數正規化，變異數總和為 m，分析得到的主成分百分比是以 m 為分母的數值。

2. 每個特徵值代表一個主成分分量解釋多變數總變異量（$\mathbf{\Lambda}_{xx}$）的數量，或正規化總變異量（$\mathbf{R}_{xx}$）的數量，$\sum_{i=1}^{m}\lambda_i=\sum_{i=1}^{m}\sigma_i^2$ 或是 $=m$。若按特徵值大小順序排列 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m\geq 0$，則第一主成分代表變異量最多的主成分，其他依序遞減。

3. 主成分樣本 $y_{j,k}=\vec{e}_j\cdot\vec{x}_{t=k}$ 具有正交特性：

$$E\left[(Y_i-\bar{y}_i)(Y_j-\bar{y}_j)\right]=\begin{cases}\lambda_i & \text{if} \quad i=j\\ 0 & \text{if} \quad i\neq j\end{cases}$$

4. 以 y_j 作為迴歸方程式的影響變數，如以下的簡單模型與複雜模型；由主成分變數的正交特性，可知增減變數項不會影響其他變數的迴歸係數，$\beta_{is}=\beta_{ic}$，$\forall i=1,\cdots,k$；$\beta_{0s}-\beta_{0c}=\beta_{kc+1}\bar{y}_{k+1}+\cdots+\beta_{lc}\bar{y}_l$。

$$z=\beta_{0s}+\beta_{1s}y_1+\beta_{2s}y_2+\cdots+\beta_{ks}y_k+\varepsilon$$

$$z=\beta_{0c}+\beta_{1c}y_1+\beta_{2c}y_2+\cdots+\beta_{kc}y_k+\beta_{kc+1}y_{k+1}+\cdots+\beta_{lc}y_l+\varepsilon$$

五、舉例說明：

例 2-2 若 $\mathbf{R}_{xx}=\begin{bmatrix}1 & 0.6\\ 0.6 & 1\end{bmatrix}$，求其特徵值與特徵向量。

解：特徵數值（eigenvalues）必須滿足 $\begin{vmatrix}1-\lambda & 0.6\\ 0.6 & 1-\lambda\end{vmatrix}=0$，則

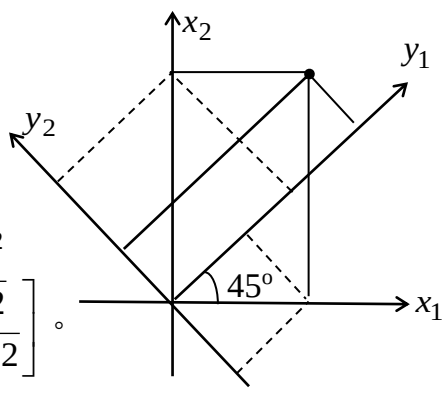
$$\lambda^2-2\lambda+1-0.36=0 \;\Leftrightarrow\; \lambda=\frac{2\pm\sqrt{4-2.56}}{2} \;\Leftrightarrow\; \lambda_1=1.6\text{、}\lambda_2=0.4\text{；}$$

當 $\lambda_1=1.6$，特徵向量使得 $\mathbf{R}_{xx}\vec{e}_1=\lambda_1\vec{e}_1$

因此 $\vec{e}_1=\begin{bmatrix}1\\1\end{bmatrix}$，正規化：$\vec{e}_1=\begin{bmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{bmatrix}$；

當 $\lambda_2=0.4$，特徵向量使得 $\mathbf{R}_{xx}\vec{e}_2=\lambda_2\vec{e}_2$

因此 $\vec{e}_2=\begin{bmatrix}1\\-1\end{bmatrix}$，正規化：$\vec{e}_2=\begin{bmatrix}1/\sqrt{2}\\-1/\sqrt{2}\end{bmatrix}$。



正交特性：$\vec{e}_1 \cdot \vec{e}_2 = \frac{1}{2} - \frac{1}{2} = 0$ 。

轉換變數：$y_{1k} = \frac{x_{1k}}{\sqrt{2}} + \frac{x_{2k}}{\sqrt{2}}$ ，$y_{2k} = \frac{x_{1k}}{\sqrt{2}} - \frac{x_{2k}}{\sqrt{2}}$ ，for $k = 1, \cdots, n$ 。

若將 x_1 和 x_2 視為卡氏座標的兩個座標軸，則 y_1 和 y_2 構成逆時鐘旋轉 45° 的座標系統。相同資料點在兩個座標系統間之座標轉換：

$$y_1 = \frac{x_1 - \bar{x}_1}{S_{x_1}} \cos 45^\circ + \frac{x_2 - \bar{x}_2}{S_{x_2}} \sin 45^\circ$$

$$y_2 = -\frac{x_1 - \bar{x}_1}{S_{x_1}} \sin 45^\circ + \frac{x_2 - \bar{x}_2}{S_{x_2}} \cos 45^\circ$$

主成分變數：$\frac{1}{n}\sum_{k=1}^{n} y_{1k}^2 = \lambda_1 = 1.6$ 和 $\frac{1}{n}\sum_{k=1}^{n} y_{2k}^2 = \lambda_2 = 0.4$ 。

例 2-3 若 $\mathbf{\Lambda}_{xx} = \begin{bmatrix} 16 & 7.2 \\ 7.2 & 9 \end{bmatrix}$，求其特徵值與特徵向量。

解：特徵數值（eigenvalues）必須滿足 $\begin{vmatrix} 16-\lambda & 7.2 \\ 7.2 & 9-\lambda \end{vmatrix} = 0$ ，

$$\Rightarrow \lambda^2 - 25\lambda + 144 - 51.84 = 0 \quad \Leftrightarrow \quad \lambda = \frac{25 \pm \sqrt{625 - 4 \times 92.16}}{2}$$

$\Leftrightarrow \lambda = 20.51 \ \ or \ \ 4.49$ 。

特徵向量：$\mathbf{\Lambda}_{xx}\vec{e}_1 = \lambda_1 \vec{e}_1 \ \Rightarrow \ \begin{bmatrix} 16 & 7.2 \\ 7.2 & 9 \end{bmatrix} \cdot \begin{bmatrix} e_{11} \\ e_{12} \end{bmatrix} = \begin{bmatrix} 20.51e_{11} \\ 20.51e_{12} \end{bmatrix}$，求解得到

正規化向量 $\vec{e}_1 = \begin{bmatrix} 0.847 \\ 0.531 \end{bmatrix}$ ；

$\mathbf{\Lambda}_{xx}\vec{e}_2 = \lambda_2 \vec{e}_2 \ \Rightarrow \ \begin{bmatrix} 16 & 7.2 \\ 7.2 & 9 \end{bmatrix} \cdot \begin{bmatrix} e_{21} \\ e_{22} \end{bmatrix} = \begin{bmatrix} 4.49e_{21} \\ 4.49e_{22} \end{bmatrix}$，求解得到

正規化向量 $\vec{e}_2 = \begin{bmatrix} -0.531 \\ 0.847 \end{bmatrix}$ ；

正交特性：$\vec{e}_1 \cdot \vec{e}_2 = 0.847 \times (-0.531) + 0.847 \times 0.531 = 0$ 。

轉換變數：$y_{1k} = 0.847x_{1k} + 0.531x_{2k}$ ，$y_{2k} = -0.531x_{1k} + 0.847x_{2k}$ ，

for $k = 1, \cdots, n$。

若將 x_1 和 x_2 視為卡氏座標的兩個座標軸，則 y_1 和 y_2 構成逆時鐘旋轉 $\theta = \cos^{-1} 0.847 = 32.1^\circ$ 的座標系統。相同資料點在兩個座標系統間之座標轉換：

$$y_1 = \left(x_1 - \bar{x}_1\right)\boldsymbol{cos}\,32.1^\circ + \left(x_2 - \bar{x}_2\right)\boldsymbol{sin}\,32.1^\circ$$

$$y_2 = -\left(x_1 - \bar{x}_1\right)\boldsymbol{sin}\,32.1^\circ + \left(x_2 - \bar{x}_2\right)\boldsymbol{cos}\,32.1^\circ$$

主成分變數：$\frac{1}{n}\sum_{k=1}^{n} y_{1k}^2 = \lambda_1 = 20.51$ 和 $\frac{1}{n}\sum_{k=1}^{n} y_{2k}^2 = \lambda_2 = 4.49$。

本例題和例題 2-1 的差別，在於 x_1 和 x_2 的標準偏差不同，分別是 $\sigma_{x1} = 4$ 和 $\sigma_{x2} = 3$；而例題 2-1 用的是相關係數矩陣，正規化的結果使得 $\sigma_{x1} = \sigma_{x2} = 1$。

六、IMSL 程式庫說明：

1. **主成分分析**（Principle Component Analysis）：主要是分析時間-空間的資料，利用求固有值（eigen-value）及固有向量（eigen-vector）之方法，過濾出佔最大變異數的型態，此即為最主要之型態。例如：太平洋的海面溫度經主成分分析後，第一主成分即為聖嬰現象模式（El Nino mode），這是最為大家熟知的型態。其他也可以應用於高度場、某地區平均氣溫或雨量的分析，可以得到各種場的主要型態。在 IMSL 統計副程式庫裡，有 EVCSF 副程式，可以簡單的計算出固有值及固有向量，求出主要型態。因為氣象資料一般都很龐大，矩陣反覆運算容易有大截斷誤差，可能會得到不確實的結果，IMSL 統計副程式庫亦有 EPISF 函數，提供判斷 EVCSF 副程式輸出結果可靠性之指標，使得運用上更能安心。

2. **正準相關分析**（Canonical Correlation Analysis，CCA）：主要是在找尋兩個不同資料向量場間最好的相關，是考慮兩組向量場間大範圍互相的變化。和主成分分析只分析單一向量場之變化略有不同，不過仍和主成分分析一樣，必須計算固有值及固有向量。IMSL 統計副程式

庫裡，有 CANCR 可以計算出兩組向量間最佳之線性組合（正準相關向量），分析出兩向量間，最配合的型態。在氣象運用上較常見的，有求出海面溫度和 500 百帕高度場間的相互關係，可知道聖嬰現象時大氣環流可能出現的型態。

3. **特異值分解**（Singular Value Decomposition，SVD）：在應用上和 CCA 相似，主要也是求出兩向量場間最佳組合型態。有報告指出 SVD 的結果優於 CCA。IMSL 統計副程式庫裡，亦有 LSVRR 直接算出最佳組合的兩向量場。