Chapter 12

迴歸分析與相關分析

12.1 迴歸分析之概念

迴歸分析（regression analysis）是用來研究兩個或兩個以上變數間的關係，而此方法的主要目的是要建立一個迴歸模式，然後根據一個或多個自變數（independent variable）之數值，再利用此迴歸模式來預測相依變數（dependent variable）之值。

一、迴歸分析之分類

迴歸分析之研究因自變數個數多寡及模式型態可將其區分爲：

- 迴歸分析
  - 一個自變數
    - 線性迴歸
    - 非線性迴歸
  - 多個自變數
    - 線性迴歸
    - 非線性迴歸

本章主要內容以簡單線性迴歸模式爲主。今先探討單一自變數之迴歸分析，再將其統計推論觀念推廣之多個自變數之迴歸分析問題。

Remark

一般所謂線性迴歸中之“線性（linear）”是指迴歸模式中所有迴歸參數皆為一次方的型態。

二、簡單直線迴歸模式的建立

㈠**直線迴歸模式之建立**：茲舉下列說明模式之建立，若觀察廣告費用與銷售量之資料如下，且將此資料繪於下圖（見圖12-1）。

X（廣告費）	5	5	6	6	6	7	8	8	8	9	9	10	10
Y（銷售量）	15	18	26	23	27	31	30	29	34	37	35	40	42

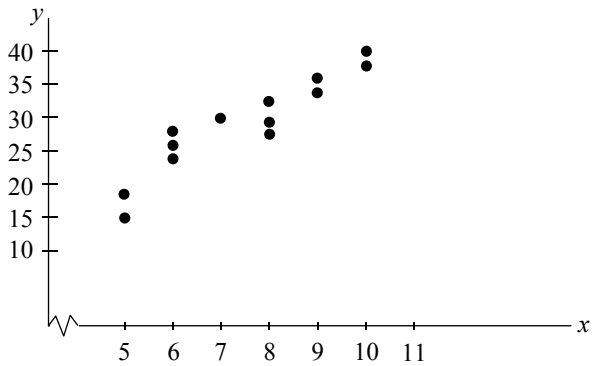

圖12-1　廣告費用與銷售量之散布圖

Remark

一般所謂散布圖（scatter diagram）是指將一組資料（x_i, y_i）點繪在二維空間上所構成的圖形。

在圖12-1中，可看出廣告費用與銷售量之關係並非直線關係，亦即兩變數並非函數關係，因此銷售量多寡並非完全受廣告費之影響。但可粗略看出此兩變數關係接近直線，因此可將其關係建立爲

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad ，i = 1,2,3,...,n$$

㈡**模式及其基本假設**：在直線迴歸模式 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ 中，有下列基本假設：

1. Y_i 是第 i 觀察值的相依變數（反應變數）。
2. X_i 爲非隨機變數（nonstochastically variable）是已知常數，爲自變數的第 i 觀察值。

3. ε_i 是隨機誤差項（random error term）或稱隨機干擾項（random disturbance），且 $E(\varepsilon_i)=0$，$\sigma^2(\varepsilon_i)=\sigma^2$ ，$i=1,2,3,...,n$
4. ε_i 和 ε_j 無相關（uncorrelated），所以對所有 i、j 而言，共變異數 $\sigma(\varepsilon_i,\varepsilon_j)=0$，$i\neq j$
5. ε_i 服從常態分配。

Remark

1. 在迴歸模式中，若迴歸係數及變數皆為線性模型，稱之為第一階模型（first-order model）。
2. 在上述模式中，β_0 與 β_1 則為迴歸係數（regression coefficient），而斜率 β_1 表示自變數 X 每變動一個單位，所引起反應變數 Y 的平均變動量，而截距項 β_0 表示在 Y 軸上的截點與原點之間的距離。

12.2 直線迴歸模式之參數估計

在直線迴歸模式中，迴歸係數 β_0 與 β_1 皆未知，因此必須利用樣本觀察資料 (x_i,y_i) 去估計參數 β_0 與 β_1。一般在統計理論中，尋找參數估計式之方法有許多種（見第8章），在此處只討論普通最小平方法（ordinary least squares method）及最大概似法（maximum likelihood method）兩種方法。茲分述如下：

一、普通最小平方法

㈠**最小平方估計式之推導**：簡單直線迴歸模式中，最主要的工作是針對模式

$$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$$

中未知參數 β_0 及 β_1 給予估計。換句話說我們要利用一組樣本觀察值 (X_i,Y_i) 所計算出來之樣本迴歸直線

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$$

來估計母體迴歸直線。今若要尋找 $\hat{\beta}_0$ 及 $\hat{\beta}_1$ 使得 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$ 爲“最

適合”此組樣本資料$(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$的直線，可利用實際值（$Y_i$）和預測值（$\hat{Y}_i$）之誤差$(e_i=y_i-\hat{y}_i)$之觀念來尋找，即要找$\hat{\beta}_0$及$\hat{\beta}_1$使誤差平方和

$$SSE=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2$$

爲最小，而要使*SSE*爲最小，則可利用微積分求極小值的概念，求出迴歸參數之估計式，且稱之爲普通最小平方估計式（OLSE）。

定理㈠：在迴歸模式$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$中，參數$\beta_0$及$\beta_1$之最小平方估計式爲

$$\hat{\beta}_1=\frac{n\sum_{i=1}^{n}X_iY_i-\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}=\frac{\sum_{i=1}^{n}X_iY_i-n\overline{X}\overline{Y}}{\sum_{i=1}^{n}X_i^2-n\overline{X}^2}$$

$$=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

$$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}$$

【證明】

因$SSE=\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2$，今對$\hat{\beta}_0$、$\hat{\beta}_1$作一階偏導數，即

$$\frac{\partial SSE}{\partial\hat{\beta}_0}=2\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)(-1)$$

$$\frac{\partial SSE}{\partial\hat{\beta}_1}=2\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)(-X_i)$$

再令$\frac{\partial SSE}{\partial\hat{\beta}_0}=0$及$\frac{\partial SSE}{\partial\hat{\beta}_1}=0$，即

$$\begin{cases}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)=0\\ \sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)X_i=0\end{cases}\Rightarrow\begin{cases}n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}X_i=\sum_{i=1}^{n}Y_i\quad\cdots\cdots(1)\\ \hat{\beta}_0\sum_{i=1}^{n}X_i+\hat{\beta}_1\sum_{i=1}^{n}X_i^2=\sum_{i=1}^{n}X_iY_i\quad\cdots\cdots(2)\end{cases}$$

此方程式稱爲標準方程式或稱正規方程式（normal equation），再利用crame rule來解此方程式，即

$$\hat{\beta}_1=\frac{\begin{vmatrix} n & \sum_{i=1}^{n}Y_i \\ \sum_{i=1}^{n}X_i & \sum_{i=1}^{n}X_iY_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n}X_i \\ \sum_{i=1}^{n}X_i & \sum_{i=1}^{n}X_i^2 \end{vmatrix}}=\frac{n\sum_{i=1}^{n}X_iY_i-\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}$$

$$=\frac{\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

又由標準方程式中第一式，兩邊同除 n 可知

$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}$$

Remark

再迴歸分析中常將各種差異平方和以下式符號簡寫，即

$$SS_Y=S_{YY}=\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2$$

$$SS_X=S_{XX}=\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum_{i=1}^{n}X_i^2-n\bar{X}^2$$

$$SS_{XY}=S_{XY}=\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}$$

故在定理㈠中之斜率可寫為 $\hat{\beta}_1=\dfrac{S_{XY}}{S_{XX}}$

例題 1

下表為5位電腦推銷員之年資（年）及最近三個月內之銷售量（台）：