第二章　基本理論之探討

2.1 模型架構與共線性之影響

首先將文中的線性迴歸模型定義如下：

$$Y = \beta_0 1 + X\beta + \varepsilon \tag{2.1.1}$$

$$\varepsilon \sim N_n(0, \sigma^2 I_n)$$

其中

Y : $n\times 1$ 的觀測值向量

X : $n\times k$ 的迴歸矩陣，且經過標準化，即 $\sum_i X_{ij} = 0, \sum_i X_{ij}^2 = 1$

β_0 :未知的常數項

1 : $n\times 1$ 的向量其元素均為 1

β : $k\times 1$ 的參數向量

ε : $n\times 1$ 的隨機誤差向量

令 $X = [x_1, x_2, \cdots, x_k]$ ，此處的 x_j 為 X 的第 j 行，也就是說 x_j 包含了第 j 個自變數的 n 個選定值。經由 X 的行向量間線性相依的關係可對「共線性」的概念下一個定義。如果存在一組不全為零的常數 $t_1, t_2, \cdots, t_k$ 使得

$$\sum_{j=1}^{k} t_j x_j = 0 \tag{2.1.2}$$

則稱向量 $x_1, x_2, \cdots, x_k$ 有線性相依的關係。若(2.1.2)式對 X 的某些行向量

剛好成立時，$X'X$ 的秩(rank) 便會小於k，也就是說$X'X$ 為不可逆矩陣。通常(2.1.2)式的等號右邊只會是一個接近於零的向量，此時稱$X'X$ 有近似線性相依關係(near-linear dependency)，而多重共線性(multicollinearity)也就發生了。但在實務上，完全線性相依的關係很少在資料中出現，反而是近似關係較常見。因此，共線性是一種程度輕重的問題，而非存在與否的問題。

考慮模型(2.1.1)中β的最小平方估計

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.1.3}$$

及其共變異矩陣(covariance matrix)

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{2.1.4}$$

當自變數間的線性相依關係很強時，$(X'X)^{-1}$的元素會急速膨脹，直接對最小平方估計及其變異數造成了重大的影響，這些影響稍後會加以介紹。

由於共線性帶來的影響相當嚴重，統計學家乃至於一般研究者從事研究工作時，都會設法儘可能減輕資料中的共線性。Montgomery, Peck, and Vining(2001)詳述了造成共線性的四個主要原因：

1. 資料搜集的方法出現瑕疵：當分析者在抽樣時，所取得自變數值大致滿足(2.1.2)式時，便會造成資料線性相依的關係。
2. 模型或母體本身的限制：例如在生產過程中，投入不同的生產要

素以得到不同的生產量。但是這些生產要素的耗用量總和，通常為固定的常數。這種限制便造成了各生產要素間的線性相依關係。

3. 模型選取的方式出現問題：分析者常為了增加解釋模型的變異，將二次項或交互項加入模型中，此舉將會導致明顯的共線性。
4. 一個過度定義的模型(overdefined model)：模型中自變數的個數超過樣本數時，稱為過度定義的模型。這種情形有時發生在醫藥和行為模式的研究中，此時收集了許多自變數的資訊卻只有少數的樣本被取得，因而導致共線性的發生。

而模型因發生共線性所導致最常見的影響不外乎下列兩點：

(1) 造成最小平方估計的變異數與共變異數變大

令$C=(X'X)^{-1}$，則矩陣C的對角線元素C_{jj}可表為

$$C_{jj}=\frac{1}{1-R_j^2}\text{，}j=1,2,\cdots,k \tag{2.1.5}$$

此處的R_j^2為變數X_j對其餘$k-1$個自變數的複相關係數的平方。當變數的共線性很強時，R_j^2會趨近於 1，C_{jj}則會急速的膨脹。 觀察個別的最小平方估計$\hat{\beta}_j$的變異

$$Var(\hat{\beta}_j)=C_{jj}\sigma^2=\sigma^2/(1-R_j^2) \tag{2.1.6}$$

可立刻得知，當共線性很強時，迴歸係數估計本身的變異數

會變大。一般而言，當對應的自變數之間存在共線性的關係時，迴歸係數估計彼此間的共變異數都會變的很大。

(2) 最小平方估計的絕對值亦會變大

今考慮$\hat{\beta}$與實際參數β的平方距離，其定義為

$$L_1^2=(\hat{\beta}-\beta)'(\hat{\beta}-\beta) \tag{2.1.7}$$

而(2.1.7)式的期望值為

$$\begin{aligned} E(L_1^2) &= E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) \\ &= \sum E(\hat{\beta}_j-\beta_j)^2 \\ &= \sum Var(\hat{\beta}_j) \\ &= \sigma^2 tr(X'X)^{-1} \end{aligned} \tag{2.1.8}$$

因為對稱矩陣的跡(trace)剛好等於該矩陣所有特徵值(eigen-value)的和，因此(2.1.8)式又可改寫為

$$E(L_1^2)=\sigma^2\sum_{j=1}^{k}\frac{1}{\lambda_j} \tag{2.1.9}$$

此處的λ_j>0 為矩陣$X'X$的特徵值，$j=1,2,\cdots,k$。

當發生共線性時，至少會有一個λ_j非常小，使得 (2.1.9)式的值的很大，也就是說$\hat{\beta}$與實際參數β的距離很遠。另外，(2.1.8)式也可改寫為

$$\begin{aligned} E(L_1^2) &= E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) \\ &= E(\hat{\beta}'\hat{\beta}-2\hat{\beta}'\beta+\beta'\beta) \end{aligned}$$

或

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 tr(X'X)^{-1} \tag{2.1.10}$$

由(2.1.10)式得知，一般而言 $\hat{\beta}$ 的長度都會比 β 長，這也證明了迴歸係數的最小平方估計在加了絕對值後會變的很大。

通常處理共線性問題的方法就是設法減少自變數，Mason, Gunst, and Webster(1975)提出了三個建議：

1. 以個數較少的自變數重新定義模型
2. 使用部份的原始自變數預作初步的研究
3. 以主成份迴歸(principal component regression)方法決定應刪除那些變數

但前兩個建議方法並未考慮到變數間的相關性，因而可能產生讓我們不滿意的結果，主成份迴歸的方法會在下一節介紹。

在了解共線性所帶來的影響後，多數人在從事研究時，大多希望能就現有的資料來判斷共線性的強弱；最常用來偵測共線性強弱的準則有下列三種:

1. 利用相關係數矩陣

這是判斷是否存在共線性最簡單的方法之一。令 $R=(R_{ij})$ 為解釋變數 $X_1, X_2, \cdots, X_k$ 的相關係數矩陣(correlation matrix)，R_{ij} 係變數 X_i 與 X_j 的相關係數。因 X 已經標準化，此時，$R = X'X$ 。當 X_i 與

X_j 呈近似線性相依關係時，$|R_{ij}|$ 會很靠近 1。但是，這個判斷的方式僅適用於偵測任兩個變數間是否存在近似線性相依的關係；當超過兩個以上的變數呈現近似線性相依關係時，並不保證任一組 R_{ij} 很大，為此法美中不足之處。

2. 利用變異膨脹因子(variance inflation factor,簡稱 VIF)

回到前面曾提到的矩陣 $C=(X'X)^{-1}$ 及(2.1.5)式，若 X_j 與其餘 $k-1$ 個變數呈現出近似線性相依時，(2.1.5)式中的 R_j^2 會很接近 1，使得 C_{jj} 變的很大。經由(2.1.6)式，我們了解 C_{jj} 是自變數間存在近似線性相依關係時，造成參數估計的變異劇增的主因，故稱 C_{jj} 為變異膨脹因子，並定義

$$VIF_j \equiv C_{jj} = (1-R_j^2)^{-1} \tag{2.1.11}$$

當一個或數個 VIF_j 的值很大時，顯示共線性已經發生。實務上，若任何 VIF_j 的值大於 5 或 10 時，資料中便存在很嚴重的共線性，此時參數的估計也不具參考價值。

3. 利用條件數(condition number)

分析者也可藉由矩陣 $X'X$ 的特徵值來判斷共線性在資料中的嚴重程度。如果服從近似線性相依關係的自變數不只兩個，則矩陣 $X'X$ 中至少會有兩個，甚至更多的特徵值會很小。

定義條件數(condition number) κ 如下：

$$\kappa \equiv \frac{\lambda_{\max}}{\lambda_{\min}} \tag{2.1.12}$$

其中 $\lambda_{\max}$ 係矩陣 $X'X$ 的最大特徵值，$\lambda_{\min}$ 係矩陣 $X'X$ 的最小特徵值。

條件數同時也是一個測量矩陣 $X'X$ 的特徵值分散程度的標準。一般而言，條件數在 100 以內時，共線性並不嚴重；若條件數在 100 與 1000 之間，表示有中度至強度間的共線性；當條件件數超過 1000 時，資料中的共線性非常嚴重。

2.2 主成份迴歸

主成份迴歸的主要概念，是把自變數的空間內所能解釋總變異較少的主成份刪除。與一般迴歸不同之處，在於所刪除的是一組自變數的線性組合，而非單一變數。

令 $V=[v_1,v_2,\cdots,v_k]$ 是一個 $k\times k$ 的正交矩陣，其中 v_j 是對應於矩陣 $X'X$ 的特徵值 λ_j 之特徵向量(eigenvector)。定義 $W\equiv X\ V$ 為 X 的主成份，透過特徵向量與主成份的性質可得到：(1) $V'V=VV'=I_k$ (2) $W'W=diag(\lambda_1,\lambda_2,\cdots,\lambda_k)\equiv\Lambda$。重新排序 λ_j 使得 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_k$ 且對應於特徵向量 $v_1,v_2,\cdots,v_k$。則模型(2.1.1)可改寫成

$$Y=\beta_0\,1+W\gamma+\varepsilon \tag{2.2.1}$$

此處的 $\gamma=V'\beta$。γ 的最小平方估計 $\hat{\gamma}=\Lambda^{-1}W'Y$，$\hat{\gamma}_j$ 的變異數為 σ^2/λ_j，

$j=1,2,\cdots,k$ 。而預測方程式則為

$$\hat{Y}=\overline{Y}+\sum_{j=1}^{k}u_j\hat{\gamma}_j \tag{2.2.2}$$

此處的 $u_j=z'v_j$ 為預定值向量 z 的第 j 個主成份。由於主成份迴歸會事先刪除那些對應於較小特徵值的主成份；換個方式來說，是保留那些對應於較大特徵值的主成份。假設被保留的主成份個數是 s ，則(2.2.2)式將變成

$$Y^*=\overline{Y}+\sum_{j=1}^{s}u_j\hat{\gamma}_j \tag{2.2.3}$$

根據(2.2.3)式，我們可導出 Y^* 的殘差平方和(residual sum of squares)為

$$SSE^*=\sum_{i=1}^{n}(Y_i-Y_i^*)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2-\sum_{j=1}^{s}\lambda_j\hat{\gamma}_j^2 \tag{2.2.4}$$

2.3 廣義壓縮最小平方估計

由於資料發生多重共線性時，最小平方估計的精確度很差。近年來發展出許多線性有偏估計的方法，可用來改善最小平方估計的缺點。王力群(1990)提出了模型(2.1.1)中參數向量 β 的廣義壓縮最小平方估計

$$\hat{\beta}_{GS}=VAV'\hat{\beta} \tag{2.3.1}$$

其中 $A=\mathrm{diag}(a_1,a_2,\cdots,a_k)$ ， $0\le a_j\le 1$ ， $j=1,2,\cdots,k$ ，而 $V=[v_1,v_2,\cdots,v_k]$

為正交矩陣(orthogonal matrix)使得$V'X'XV$ =diag($\lambda_1, \lambda_2, \cdots, \lambda_k$)≡ Λ 。事實上，V 的第 j 個行向量 v_j 為矩陣 $X'X$ 對應於第 j 個特徵值 λ_j 的特徵向量，且為了方便起見此處的 λ_j 滿足 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0$ 。

岳榮先(1991)導出了廣義壓縮最小平方估計的槓桿值公式及一些基本性質，並得到了特殊狀況下的脊估計、主成份估計、Stein 估計、疊代估計及根方估計等多種線性有偏估計的槓桿值公式，同時也證明了當共線性發生時，廣義壓縮最小平方估計的確優於最小平方估計 $\hat{\beta}$ 。

本文即是把廣義壓縮最小平方估計中將 $a_1, a_2, \cdots, a_k$ 設限在[0,1]範圍內的概念應用在 Mansfield et al.(1977)刪除自變數的理論中，希望在加了限制條件後，所得到刪除變數的結果能更嚴謹。