第三章 研究方法

本章將介紹我們主要研究的方法：第一節介紹線性迴歸模型，第二節介紹本文重要的性質及其定義，第三節為介紹 James-Stein type 係數壓縮的概念，為 $JSWT^{+}$前身之介紹，第四節介紹 Sclove 估計量將 JS^{+}應用於線性迴歸模型，第五節介紹 James-Stein with Threshoding 估計量(JSWT)，最後在第六節闡述 $JSWT^{+}$於迴歸問題之應用，並建立 $JSWT^{+}$變數選取流程。

第一節 線性迴歸模型

考慮線性迴歸模型：$Y_i = \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i$，$(i = 1, 2, \cdots, n)$，假設$\varepsilon \sim N_n(0, \sigma^2 I_n)$；其矩陣形式如下：

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{xd} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = X\beta + \varepsilon$$

其中 Y 向量為觀測值，總共有 n 筆；X 矩陣為維度$(n \times d)$的迴歸矩陣或稱設計矩陣，本文考慮 X 矩陣為滿秩(full rank)，其秩為 d，即 X 矩陣每一行線性獨立；β 向量為未知係數，總共有 d 個。而 Y 即是反應變數；x_j $(j = 1, 2, \cdots, d)$ 為解釋變數。

由於 β 未知，所以必須估計 β，利用傳統最小平方法估計，就是找出 β 使得 $\|\varepsilon\|^2$ 有最小值，將 $\|\varepsilon\|^2$ 對 β 微分求極值，可得到對 β 的估計為 $\hat{\beta}^{OLS}$，其計算如下：

$$\|\varepsilon\|^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - 2\beta' X'Y + \beta' X'X\beta$$

$$\Rightarrow \quad \partial \|\varepsilon\|^2 / \partial \beta = -2X'Y + 2X'X\beta = 0$$

$$\Rightarrow \quad X'X\hat{\beta} = X'Y$$

$$\Rightarrow \quad \hat{\beta}^{OLS} = (X'X)^{-1} X'Y$$

即當 $\beta = \hat{\beta}^{OLS}$ 時，使得 $\|\varepsilon\|^2 = \|Y - X\beta\|^2$ 有最小值。

第二節 Minimax 與 Dominate

如果$\hat{\theta}$為θ的估計，則$\hat{\theta}$的損失(loss)定義為$\left\|\hat{\theta}-\theta\right\|^2$；而$\hat{\theta}$的風險(risk)定義為$\left\|\hat{\theta}-\theta\right\|^2$的期望值，即$E\left(\left\|\hat{\theta}-\theta\right\|^2\right)$，與我們常見的期望平方誤差(mean square error, MSE)相同。

如果$\hat{\theta}$為θ的估計，且當$\hat{\theta}$最大的風險不大於所有其他估計的最大風險時，定義$\hat{\theta}$具有 minimax 性質。換言之：

【定義】(Minimax)

當$\sup_{\theta} E\left(\left\|\hat{\theta}-\theta\right\|^2\right) \le \sup_{\theta} E\left(\left\|\tilde{\theta}-\theta\right\|^2\right)$，其中$\tilde{\theta}$為所有其他$\theta$的估計時，則$\hat{\theta}$為 minimax。

【定義】(Dominate)

當$E\left(\left\|\hat{\theta}-\theta\right\|^2\right) \le E\left(\left\|\tilde{\theta}-\theta\right\|^2\right)$時，則$\hat{\theta}$可控制(dominate)$\tilde{\theta}$。

當$E\left(\left\|\hat{\theta}-\theta\right\|^2\right) < E\left(\left\|\tilde{\theta}-\theta\right\|^2\right)$，則$\hat{\theta}$可強烈地控制(strictly dominate)$\tilde{\theta}$。

第三節 James-Stein type Shrinkage

壓縮(shrinkage)的概念源自於 James 與 Stein(1961)所提。假設觀察 d 維度 Z 向量且 $d>2$，即 $Z=(Z_1,Z_2,\cdots,Z_d)$，並假設 Z 向量服從 d 維度多元常態分布，其期望值為 θ，其共變異數矩陣為 $\sigma^2 I_d$，即 $Z \sim N_d(\theta,\sigma^2 I_d)$。顯然地，$\theta$ 的最大概似估計量為 Z。現在考慮計算 $\|Z\|^2$ 的期望值，即

$$E(\|Z\|^2)=\sum_{i=1}^{d}E\left(Z_i{}^2\right)=\sum_{i=1}^{d}\left(\sigma^2+\mu_i{}^2\right)=d\sigma^2+\|\theta\|^2>\|\theta\|^2$$

可發現 $\|Z\|^2$ 的期望值大於 $\|\theta\|^2$，所以此估計量 Z 以此觀點來看並不讓人滿意。因此以 $\|Z\|^2$ 的期望值來看，Z 估計量可能估的太大，隱含了將 Z 壓縮(shrinkage)的可能性。

現在考慮另一估計 $\tilde{\theta}$ 為 cZ，$0<c<1$。雖然 $\tilde{\theta}$ 是偏的(biased)，但是可利用微分求極值去找到合適的 c 使得 $\tilde{\theta}$ 具有最小的 MSE。即考慮：

$$f(c)=E\left(\left\|\tilde{\theta}-\theta\right\|^2\right)=c^2d\sigma^2+(1-c)^2\|\theta\|^2$$

$$\Rightarrow \quad f'(c)=2cd\sigma^2-2(1-c)\|\theta\|^2$$

$\Rightarrow$ 令 $f'(c)=0$ 可解得 $c=\|\theta\|^2/(d\sigma^2+\|\theta\|^2)$ 可使得 $f(c)$ 有最小值

所以理想的估計 $\tilde{\theta}$ 為：

$$\tilde{\theta}=cZ=\left(1-\frac{d\sigma^2}{d\sigma^2+\|\theta\|^2}\right)Z$$

由於 $\|\theta\|^2$ 未知，可利用上述推導所得 $\|Z\|^2$ 為 $d\sigma^2+\|\theta\|^2$ 的不偏估計量，將 $\tilde{\theta}$ 改寫為：

$$\tilde{\theta}=cZ=\left(1-\frac{d\sigma^2}{\|Z\|^2}\right)Z$$

James 與 Stein (1961)證得 $(1-b/\|Z\|^2)Z$ 此種形式的估計，其中當 $b=(d-2)\sigma^2$ 且 $(d>2)$時，具有最小的 MSE。即 James-Stein 估計量(JS)

$$\hat{\theta}^{JS}=\left(1-\frac{(d-2)\sigma^2}{\|Z\|^2}\right)Z$$

為了解決當括弧內為負值時出現的問題，Baranchik (1964)提出 James-Stein positive part 估計量，即 JS^+估計量，考慮當括弧為負值時將其壓縮為零，即

$$\hat{\theta}^{JS+}=\left(1-\frac{(d-2)\sigma^2}{\|Z\|^2}\right)_+ Z$$，其中 $(x)_+$ 這個符號即 $max(x,0)$。

所以 JS^+估計量其第 i 個元素即:

$$\hat{\theta}_i^{JS+}=\left(1-\frac{(d-2)\sigma^2}{\|Z\|^2}\right)_+ Z_i$$

則 JS^+估計量可控制(dominate)JS 估計量，即 MSE 更小。如圖 1 所示。

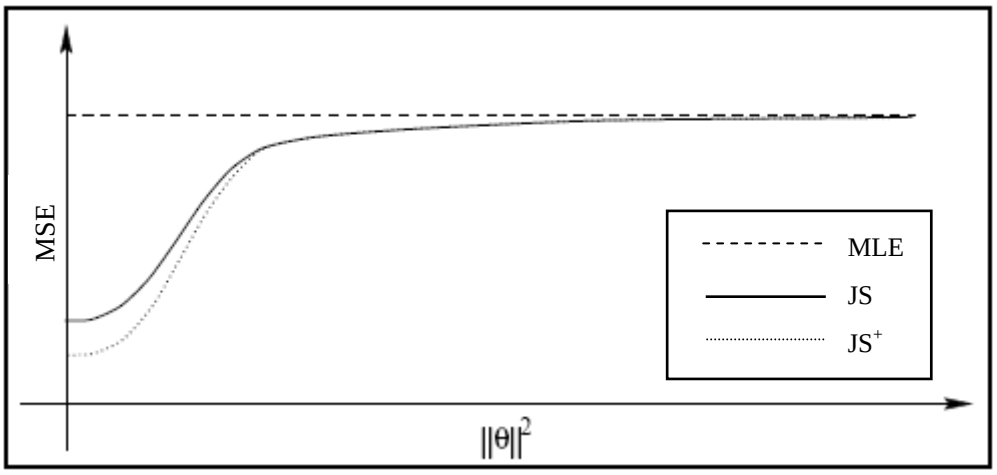

圖 1：JS^+估計量可控制 JS 估計量示意圖(Richards，1999)

第四節 Sclove^{+}估計量

Sclove(1968)承用 James 與 Stein (1961)所提的 JS 估計量及 Baranchik (1964)取正函數的概念，應用於線性迴歸模型建立 Sclove^{+}估計量。當σ^2未知時，Sclove 提出：

$\hat{\beta}_i^{\text{Sclove}^+} = \left(1 - \frac{\alpha \times RSS}{\left\|\hat{\beta}^{OLS}\right\|^2}\right)_+ \hat{\beta}_i^{OLS}$，其中 $\alpha \in (0, 2(d-2)/(n-d+2))$，$RSS$為殘差平方和

令 $F = \frac{\left\|\hat{\beta}^{OLS}\right\|^2}{RSS}$，則知道檢定統計量 $T = (n-d)F/d \sim F_{d,n-d}$。Sclove^{+}估計因此可改寫：

$$\hat{\beta}_i^{Sclove^+} = \left(1 - \frac{\alpha}{F}\right)_+ \hat{\beta}_i^{OLS}, \text{ 其中 } \alpha \in (0, 2(d-2)/(n-d+2))$$

但如果虛無假設為 $H_0 : \beta = 0$，α 為顯著水準，當 T 其檢定值小於顯著水準為 α 的臨界值時，即不拒絕虛無假設，則所有係數被壓縮為零，亦即拋棄了所有解釋變數；相反地，即拒絕虛無假設，則所有係數以固定比例被壓縮，但保留了所有解釋變數。由此看來，Sclove^{+}估計量利用 F 檢定只能在原始模型(origin model)和完全模型(full model)兩者去做挑選，並無法做其線性子集下的變數選取。

第五節 James-Stein with Thresholding 估計量(JSWT)

Zhou 與 Hwang (2005)為了改善 JS^{+}估計量僅有一個縮減模型(reduce model)的缺點，而建立一個具有 minimax 性質同時加上門檻值的估計量，即 JSWT^{+}估計量。為了建構 JSWT^{+}估計量，Zhou 與 Hwang 考慮以下標準模型：

假設觀察 $Z = (Z_1, ..., Z_d) \sim N_d(\theta, I_d)$，$\theta = (\theta_1, ..., \theta_d)$

- I_d ：共變異數單位矩陣
- Z_i ：資料的小波係數
- θ_i ：真實曲線
- d ：維度

不失一般性,考慮 $\delta(Z)=\left(\delta_1(Z),...,\delta_d(Z)\right),\delta_i\left(Z\right)=Z_i+g_i(Z)$ ，其中 $g_i(Z):R^d\to R$ 由於 Z 估計量本身即為 minimax 估計量，若建構一個可控制 Z 的估計量，即 $\delta(Z)$ ，則此估計亦具有 minimax 性質。

Zhou 與 Hwang 利用 Stein(1981)的【引理】去建構一個可控制(dominate)Z 的估計量，【引理】如下：

【引理】(Stein，1981)

Suppose that $g:R^d\to R^d$ is a measurable function with $g_i(\cdot)$ as the ith component. If for every i, $g_i(\cdot)$ is almost differentiable with respect to the ith component

and $$E\left(\left|\frac{\partial}{\partial Z_i}g_i(Z)\right|\right)<\infty \quad for\ i=1,...,d,$$

then

$$E_\theta\left\|Z+g(Z)-\theta\right\|^2=E_\theta\left\{d+2\nabla\cdot g(Z)+\left\|g(Z)\right\|^2\right\},\ where\ \nabla\cdot g(Z)=\sum_{i=1}^{d}\frac{\partial g(Z_i)}{\partial Z_i}$$

因此由上述【引理】可知，如果可以從微分不等式 $2\nabla\cdot g(Z)+\left\|g(Z)\right\|^2<0$ 解得 $g(Z)$ ，則 $Z+g(Z)$ 可強烈控制(strictly dominate)Z。而為了建構具有門檻值的估計量，將原本 JS$^+$估計量中的 $(d-2)$ 部份改寫成一個遞減函數 $h(\left|Z_i\right|)$ ，將 $\left\|Z\right\|^2$ 部份改寫成 D，即 JSWT 估計量：

$$\hat{\theta}_i^{JSWT}=\left(1-\frac{h(\left|Z_i\right|)}{D}\right)Z_i \quad \text{，其中 } D \text{ 未定且與第 i 元素獨立}$$

現在考慮 $h(\left|Z_i\right|)=a\left|Z_i\right|^{b-2}$，$D=\sum_{i=1}^{d}\left|Z_i\right|^b$，並將 JSWT 估計量改寫成 $Z_i+g_i(Z)$ 形式：

$$\begin{aligned}\hat{\theta}_i^{JSWT}&=\left(1-\frac{a\left|Z_i\right|^{b-2}}{D}\right)Z_i\\&=Z_i-\frac{Z_i\times a\left|Z_i\right|^{b-2}}{D}\\&=Z_i-\frac{sign(Z_i)\times\left|Z_i\right|\times a\left|Z_i\right|^{b-2}}{D}\\&=Z_i+\left(-\frac{a\times sign(Z_i)\times\left|Z_i\right|^{b-1}}{D}\right)\\&=Z_i+g_i(Z)\end{aligned}$$

利用 Stein(1981)的【引理】可證明當 $d\geq 3,\ 0<a\leq 2(b-1)d-2b$，$1<b\leq 2$

JSWT 估計量具有 minimax 性質，且除了 $b=2$ 以外，當 a 為上述範圍之上界時，JSWT 估計量可控制(dominate)Z。詳見【定理三】 (Zhou and Hwang，2005)。

而當此種形式估計量 $\tilde{\theta}_i=\left(1-h_i(Z)\right)Z_i$，其中 $h_i(Z)$ 為對稱函數時，Zhou 與 Hwang 指出 $E(\tilde{\theta}_i^{+}-\theta_i)^2<E(\tilde{\theta}_i-\theta_i)^2$，其中 $\tilde{\theta}_i^{+}=\left(1-h_i(Z)\right)_+ Z_i$，所以可得到 JSWT$^+$估計量：

$$\hat{\theta}_i^{JSWT^+}=\left(1-\frac{a\left|Z_i\right|^{b-2}}{D}\right)_+ Z_i \text{，其中 } D=\sum_{i=1}^{d}\left|Z_i\right|^b$$

同理可證明得當 $d\geq 3,\ 0<a\leq 2(b-1)d-2b$，$1<b\leq 2$，JSWT$^+$估計量可強烈地控制(strictly dominate) Z。同時可發現當 b=2 時，則 JSWT 估計量即是 JS 估計量，而 JSWT$^+$估計量即是 JS$^+$估計量。因此 Zhou 與 Hwang 建構出 JSWT$^+$估計量，同時具有 minimax 性質及門檻值允許我們可以在其線性子集下做變數選取。

然而 JSWT$^+$估計量中具有兩個參數 a 和 b 需要被估計，Zhou 與 Hwang (2003)指出，假設 $\theta_i \overset{i.i.d}{\sim} N(0,\tau^2)$，對於所有 τ^2，JSWT 估計量的貝氏風險(Bayes risk)比 Z 估計量的貝氏風險更小，且其若且唯若條件如下：

$$0\leq a\leq a_b=2\Big/E\left[\sum\nolimits_{i=1}^{d}\left|\xi_i\right|^{2b-2}\Big/(\sum\nolimits_{i=1}^{d}\left|\xi_i\right|^{b})^2\right],\ \text{其中 } \xi_i \overset{i.i.d}{\sim} N(0,1)$$

所以 JSWT 估計量可改寫成 b 的函數的形式，同理 JSWT$^+$ 估計量亦然，即：

$$\hat{\theta}_i^{JSWT^+}(b)=(1-a_b D_b^{-1}\left|Z_i\right|^{b-2})_+ Z_i \text{，其中 } D_b=\sum_{i=1}^{d}\left|Z_i\right|^b$$

為了推導 a_b 的廣義公式，Zhou 與 Hwang 利用取極限方式進一步對 a_b 估計，即：

$$\lim_{d\to\infty}\frac{a_b}{d}=C_b=4\left[\Gamma\left((b+1)/2\right)\right]^2\Big/\left[\sqrt{\pi}\Gamma\left((2b-1)/2\right)\right] \text{，其中 } 1/2<b<2$$

並建議 a_b 的廣義公式可利用 $0.97(d-2)C_b$ 來估計，其中 $1/2<b<2$。

則 JSWT$^+$估計量中只剩下 b 參數需要被估計，Zhou 與 Hwang 建議計算 JSWT$^+$ 估計量的 Stein unbiased risk estimator (SURE)，即：

$$SURE(b)=E\left(\left\|\hat{\theta}^{JSWT^+}-\theta\right\|^2\right)$$

$$=d+\sum_{i=1}^{d}(Z_i^{\,2}-2)I_i+a_b\left(\frac{a_b\left|Z_i\right|^{2b-2}}{D_b^{\,2}}-2(b-1)\frac{\left|Z_i\right|^{b-2}}{D_b}+2\frac{\beta\left|Z_i\right|^{2b-2}}{D_b^{\,2}}\right)I_i^{\,c}$$

其中 $I_i=\begin{cases}1,\ a_b\left|Z_i\right|^{b-2}>D_b\\ 0,\ o.w.\end{cases}$，$I_i^{\,c}=1-I_i$

而 b 的估計($\hat{b}$)，選擇使得 *SURE* 函數在 $1/2 < b < 2$ 有最小值者代入，則得到 JSWT$^+$ 估計量為：

$$\hat{\theta}_i^{JSWT^+}(\hat{b}) = (1 - a_{\hat{b}} D_{\hat{b}}^{-1} |Z_i|^{\hat{b}-2})_+ Z_i \text{ ，其中 } D_{\hat{b}} = \sum_{i=1}^{d} |Z_i|^{\hat{b}}$$

第六節 JSWT$^+$於迴歸問題之應用

Zhou 與 Hwang (2005)將 JSWT$^+$估計應用於小波分析對函數做估計，與一些可同時做估計與模型挑選的估計比較結果顯示 JSWT$^+$估計表現最佳(MSE 最小)，所以我們試著將 JSWT$^+$ 估計應用在線性迴歸模型，藉著 JSWT$^+$ 估計具有門檻值的特性將建立 JSWT$^+$變數選取流程。

現在我們將 JSWT 估計量引入線性迴歸模型，考慮線性迴歸模型：$Y = X\beta + \varepsilon$，假設 $\varepsilon \sim N_n(0, \sigma^2 I_n)$，則 $Y \sim N_n(X\beta, \sigma^2 I_n)$。利用最小平方法可以計算出 $\hat{\beta}^{OLS} = (X'X)^{-1}X'Y$ 且 $\hat{\beta}^{OLS} \sim N_d(\beta, \sigma^2 (X'X)^{-1})$。當設計矩陣為正交時，即 $(X'X)^{-1} = I_d$，則 $\hat{\beta}^{OLS} \sim N_d(\beta, \sigma^2 I_d)$。由於 $\hat{\beta}^{OLS}$ 為多維度常態分布，所以我們可直接將 JS 估計量中的 $\|Z\|^2$ 改為 $\|\hat{\beta}^{OLS}\|^2$，Z_i 改為 $\hat{\beta}_i^{OLS}$，當 σ^2=1 時，則 $\hat{\beta}_i^{JS}$ 與 $\hat{\beta}_i^{JS^+}$ 分別為：

$$\hat{\beta}_i^{JS} = \left(1 - \frac{(d-2)}{\|\hat{\beta}^{OLS}\|^2}\right)\hat{\beta}_i^{OLS} \qquad \hat{\beta}_i^{JS^+} = \left(1 - \frac{(d-2)}{\|\hat{\beta}^{OLS}\|^2}\right)_+ \hat{\beta}_i^{OLS}$$

同理我們將此概念套用於 JSWT$^+$ 估計量，可得到 $\hat{\beta}_i^{JSWT^+}$ ：

$$\hat{\beta}_i^{JSWT^+} = \left(1 - \frac{a|\hat{\beta}_i^{OLS}|^{b-2}}{D}\right)_+ \hat{\beta}_i^{OLS} \text{ ，其中 } D = \sum_{i=1}^{d} |\hat{\beta}_i^{OLS}|^b$$

變數選取流程

首先，我們將資料配適線性迴歸模型：$Y = X\beta + \varepsilon,\ \varepsilon \sim N_n(0, \sigma^2 I_n)$，其中解釋變數個數小於樣本個數且滿秩，考慮兩種狀況：

(i) 若設計矩陣 X 為正交集 $(X'X = I_d)$，我們直接計算其最小平方估 $\hat{\beta}^{OLS}$。

(ii) 若設計矩陣 X 非正交集 $(X'X \neq I_d)$，我們利用 Gram-Schmidt 正交化過程計算出正交集 Z。

在(ii)狀況中，我們考慮兩種轉換矩陣 L：

(1) 模型：$Y = X\beta + \varepsilon \Rightarrow Y = XLL'\beta + \varepsilon = (XL)(L'\beta) + \varepsilon = Z\gamma + \varepsilon$

其中 L 為轉換矩陣，使得 $Z = XL$ 為正交集，利用 QR 分解可以去解出轉換矩陣 L。

將原線性模型改成：$Y = Z\gamma + \varepsilon$，再計算其最小平方估計 $\hat{\gamma}^{OLS}$。

(2) 模型：$Y = X\beta + \varepsilon \Rightarrow LY = L(X\beta + \varepsilon) = (LX)\beta + L\varepsilon = Z\beta + L\varepsilon$

其中 L 為轉換矩陣，使得 $Z = LX$ 為正交集，利用廣義反矩陣可以去解出轉換矩陣 L。

將原線性模型改成：$LY = Z\beta + L\varepsilon$，再計算其最小平方估計 $\hat{\beta}^{OLS}$。

接下來，利用資料去計算 $\hat{b}$ 使得 $SURE(b)$ 在 $1/2 < b < 2$ 有最小值。由於將資料代入 $SURE(b)$ 函數後，$SURE(b)$ 函數顯得很複雜，無法直接求出其最小值。而 b 取值在 $1/2 < b < 2$，所以我們在 $1/2$ 與 2 之間等距生成 1000 點，將這 1000 點代入 $SURE(b)$ 函數直接找出使得 $SURE(b)$ 有最小值的 $\hat{b}$。

最後，在狀況(i)中，將 $\hat{b}$、$\hat{\beta}^{OLS}$ 代入 $\hat{\beta}_i^{JSWT^+}$，則當 $\hat{\beta}_i^{JSWT^+}$ 不為零即為挑選出之解釋變數。在狀況(ii) (1)中，將 $\hat{b}$ 代入 $\hat{\beta}_i^{JSWT^+}$，且將 $\hat{\beta}_i^{JSWT^+}$ 中 $\hat{\beta}^{OLS}$ 改為 $\hat{\gamma}^{OLS}$，則當 $L\hat{\beta}_i^{JSWT^+}$ 不為零即挑選出之解釋變數。在狀況(ii) (2)中，將 $\hat{b}$、$\hat{\beta}^{OLS}$ 代入 $\hat{\beta}_i^{JSWT^+}$，則當 $\hat{\beta}_i^{JSWT+}$ 不為零即挑選出之解釋變數。