中國統計學報
第41卷第4期
九十二年十二月
401-420頁

超大資料集的迴歸分析

任眉眉[1]　林億雄[1]　林共進[2]

[1]國立成功大學統計學系

[2]美國賓州州立大學管理科學及資訊系統系

摘　要

在常態誤差假設下，最小平方估計量亦是最大概似估計量，其具有良好的統計性質。爲了減少計算時間與記憶體空間，對於超大資料集迴歸係數的估計，我們提出二個變通的估計，一爲平均估計法，另一爲平均數最小平方估計法。我們藉由矩陣分割法討論資料分組下，最小平方估計法計算之可行性，並針對分組資料分別估計各組的迴歸係數，在各組殘差變異一致下，我們提出各組間迴歸係數重疊性的檢定方法。我們亦針對超大資料集簡單線性迴歸問題，提出決定使用平均估計法或平均數最小平方估計法的適當時機，並以一實例作爲個案探討。

關鍵詞：資料探勘，最小平方法，矩陣分割法，平均估計法，平均數最小平方估計法。

美國數學會分類索引：主要62J05；次要62F03。

1. 前言

在資訊科技不斷進步的過程中，利用科技系統自動化收集資料將爲十分普遍。而如何在超大資料集中汲取有用的資訊，則爲資料探勘研究領域最大的課題。例如：公司行號想從上下班打卡資料，了解員工上下班的統計關係；信用卡發卡公司想研究持卡者的刷卡消費金額與工作收入的關係；量販百貨公司想了解顧客當日消費金額與逗留商場時間的關係；海測中心想了解距離海平面之深度與水溫高低的關係；太空總署想了解人造衛星距離地球表面的高度與其所承受的壓力大小關係等，均是具題的研究課題。對於一超大資料集而言，在固定欄位（變數項）下資料筆數可能達到上百萬筆，且其檔案大小可能高達好幾 TB (1TB=1024GB) ；參考Hand *et al.* (2000) 或林億雄和林共進(2001)一文。在統計方法中，線性迴歸分析 (linear regression analysis) 最常被用以討論反應變數與解釋變數間的統計關聯，一般常用的統計軟體均有這項功能。當利用統計軟體作超大資料集線性迴歸分析時，電腦在儲存及運算資料時，會因爲記憶體 (memory) 不足，導致無法直接進行線性迴歸分析。

在常態(normal)誤差假設下，傳統最小平方估計量(least square estimator)，也是最大概似估計量 (maximum likelihood estimator)，其具有良好的統計性質，故本文主要針對超大資料集，探討線性迴歸係數最小平方估計量之簡化研究。首先我們藉由矩陣分割法 (matrix partition) 討論在資料分組下，傳統最小平方估計法計算之可行性。同時我們提出二種估計法針對已分組之資料，分別估計各組的迴歸係數。在各組殘差變異一致下，我們提出各組間迴歸係數重疊性的檢定方法。若整體資料滿足迴歸係數重疊性的假設，則我們將各組的迴歸係數估計量作平均，當作整體迴歸係數的估計，此法稱爲平均估計法 (averaging method)。此外，若各分組資料滿足迴歸係數重疊性 (coincidence，參考 Seber 1977) 的假設，由於各組樣本平均點必通過整體共同的最小平方迴歸直線，我們針對各組資料的樣本平均數，處理線性迴歸分析並作成參數估計，此法稱爲平均數最小平方估計法 (least square estimation on the average)。當計算工具發生記憶體不足導致無法進行最小平方估計時，可以考慮使用矩陣分割法來計算估計量，或採用本文所提出的兩種變通估

計法。相對於傳統最小平方估計法，我們所提出的兩種估計法需要較少的記憶體及較短的計算時間 (computing time)，但仍具有相當的精確度。本研究方法亦適合於處理一般超大資料集的模式分析。文中我們特別針對簡單線性迴歸問題，提出決定使用平均估計法或平均數最小平方估計法的適當時機。

本文主要章節分別如下：第二節介紹傳統線性迴歸係數的點估計，並利用矩陣分割演算法解決超大資料集記憶體不足的問題。在第三節中我們推導各分組迴歸係數重疊性的檢定理論，從而提出平均估計法與平均數最小平方估計法二種變通估計法，並研究其統計性質。在第四節中，針對簡單線性迴歸問題提出決定使用平均估計法或平均數最小平方估計法的適當時機，並利用電腦模擬說明上述二估計法使用的適當時機。第五節針對某公家機關員工上下班打卡資料進行統計分析以作爲個案探討，第六節爲結論，相關證明則置於附錄中。

2. 傳統線性迴歸係數的點估計

一般線性迴歸模式可表爲

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \quad i = 1, \ldots, N$$

其中 y_i 代表反應變數，$x_{i,1}, \ldots, x_{i,p-1}$ 代表自變數，$\beta_0, \ldots, \beta_{p-1}$ 爲迴歸係數，ϵ_i 爲獨立且具有 $E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2$ 之共同分配。其矩陣表示式爲

$$\mathrm{Y} = \mathrm{X}\beta + \epsilon \tag{1}$$

其中

$$\mathrm{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathrm{X} = \begin{bmatrix} 1 & x_{1,1} & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \ldots & x_{N,p-1} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

迴歸分析問題主要在處理參數向量 β 的統計推論，首先令 $Q(\beta)=(Y-X\beta)'(Y-X\beta)$ 代表誤差項之平方和，再尋找 β 之可能值使得 Q 爲最小。利用矩陣微分法可得

$$\tilde{\beta}=(X'X)^{-1}X'Y \tag{2}$$

使 Q 最小，此時 $\tilde{\beta}=[\tilde{\beta}_0,\tilde{\beta}_1,\cdots,\tilde{\beta}_{p-1}]'$ 即稱爲迴歸係數最小平方估計量，其具有不偏性 (unbiased)，亦即 $E(\tilde{\beta})=\beta$，而且 $Cov(\tilde{\beta})=(X'X)^{-1}\sigma^2$。在常態誤差假設下，最小平方估計量 $\tilde{\beta}$ 亦爲最大概似估計量。同時，根據 Gauss-Markov 定理最小平方估計也爲最佳線性不偏估計 (best linear unbiased estimator 簡稱 BLUE)，相關文獻請參閱 Christensen (1984), Graybill (1982) 或 Neter (1996)。

在超大資料集中，$X'X$ 之運算會佔用很大的記憶體及計算時間，此時可以利用矩陣分割運算來節省記憶體空間及運算時間。針對一超大資料集資料總數爲 N 時，考慮分爲 k 組，每組資料數爲 n，滿足 $N=k\times n$。爲方便說明矩陣分割法，我們重新定義向量 Y 及矩陣 X 如下：令

$$Y_i=\begin{bmatrix} y_{1,i} \\ y_{2,i} \\ \vdots \\ y_{n,i} \end{bmatrix}_{n\times 1}, \quad X_i=\begin{bmatrix} 1 & x_{1,i,1} & \cdots & x_{1,i,p-1} \\ 1 & x_{2,i,1} & \cdots & x_{2,i,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,i,1} & \cdots & x_{n,i,p-1} \end{bmatrix}_{n\times p} \tag{3}$$

其中 $y_{j,i}=y_{j+(i-1)\times n}$ 爲第 i 組中第 j 筆反應變數的值，$x_{j,i,s}=x_{j+(i-1)\times n,s}$ 爲第 i 組第 j 筆中第 s 個自變數的值 $i=1,\ldots,k$，$j=1,\ldots,n$ 且 $s=1,\ldots,p-1$。則 (1) 中的 Y 與 X 可表爲

$$Y=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{bmatrix}_{N\times p}, \quad X=\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}_{N\times p}$$

此時 $X'X = \sum_{i=1}^{k} X_i' X_i$ ，且 $X'Y = \sum_{i=1}^{k} X_i' Y_i$ 。透過矩陣分割運算可以節省 $X'X$ 及 $X'Y$ 的計算時間並解決記憶體不足的問題。由於 N 相對於 p 很大時，矩陣 X 爲全秩且 $(X'X)^{-1}$ 必定存在，因此 (2) 式中 $\tilde{\beta}$ 可表爲

$$\tilde{\beta} = (\sum_{i=1}^{k} X_i' X_i)^{-1} (\sum_{i=1}^{k} X_i' Y_i)。 \tag{4}$$

註一：現行的統計軟體未有 (4) 式之計算公式，故建議軟體公司應適度修正程式以切合超大資料集之迴歸分析問題。

3. 變通估計法

在第二節中爲了解決電腦計算時記憶體不足之問題，我們將超大資料集分組並利用矩陣分割法，計算超大資料集線性迴歸係數的最小平方估計值。雖然在超大資料集中，傳統最小平方估計法透過矩陣分割法是可以計算的，但關於迴歸分析之殘差值檢驗工作依然會受限於電腦記憶體不足之問題，故本節主要針對已分組之資料集探討一些變通估計法之可行性。假設資料筆數 n 不超過一般計算工具所能處理的資料筆數，在矩陣分割的原則下原本的超大資料集總筆數爲 N ，將可分爲 k 組，每組 n 筆。

3.1 平均估計法

針對這 k 組資料，若各分組資料均符合模式 (1)，亦即有相同迴歸係數向量 β 及相同殘差分配。對於第 i 組資料，迴歸係數 β 的最小平方估計向量，記爲

$$\hat{\beta}^i = (X_i' X_i)^{-1} X_i' Y_i, \quad i = 1, \cdots, k。$$

在常態誤差假設下，若各組的迴歸模式一致，則我們很自然地將各組迴歸係數估計值作平均，當作整體資料的迴歸係數估計值，此法稱爲平均估計法 (averaging method) ，其估計向量記爲

$$\hat{\beta}_{\mathrm{A}} = \frac{1}{k}\sum_{i=1}^{k}\hat{\beta}^{i} = \frac{1}{k}\sum_{i=1}^{k}(\mathrm{X}_i'\,\mathrm{X}_i)^{-1}\,\mathrm{X}_i'\,\mathrm{Y}_i \circ$$

根據計算，我們有 $\mathrm{E}(\hat{\beta}_{\mathrm{A}}) = \beta$ 且 $\mathrm{Cov}(\hat{\beta}_{\mathrm{A}}) = \sigma^2[\sum_{i=1}^{k}(\mathrm{X}_i'\mathrm{X}_i)^{-1}]/k^2$ 。在超大資料集下，平均估計法與傳統最小平方估計法的差異，在於前者先將資料分組進行最小平方估計，隨後將這些估計向量平均，共計作了 k 次的最小平方法；而後者針對整體資料僅進行 1 次最小平方估計，但需要利用矩陣分割法計算和矩陣，再求該和矩陣的反矩陣，因此需要較長的計算時間與較多的記憶體空間。有關各分組殘差變異一致檢定及迴歸係數重疊性 (coincidence) 檢定將於 3.3 節作一探討。

3.2 平均數最小平方估計法

上節中，爲了避免因電腦記憶體不足無法運算的問題，我們考慮在分組下利用平均估計法作爲變通。由於在分組後，各組的樣本平均值必定過該組的迴歸直線，因此我們以各組的樣本平均代表各該組，並針對這 k 組新資料進行迴歸分析，以做爲另一變通方法，其執行細節如下。首先將資料分 k 組，每組之樣本平均值記作$(\overline{\mathrm{Y}}_1, \overline{\mathrm{X}}_1), \ldots, (\overline{\mathrm{Y}}_k, \overline{\mathrm{X}}_k)$ ，其中

$$\overline{\mathrm{Y}}_i = \frac{1}{n}\mathbf{1}'\,\mathrm{Y}_i, \quad \overline{\mathrm{X}}_{\mathrm{i}} = \frac{1}{\mathrm{n}}\mathbf{1}'\,\mathrm{X}_{\mathrm{i}},$$

$\mathbf{1}$ 爲所有元素均爲 1 的行向量， Y_i 與 X_i 同 (3)。令 $\mathrm{X}^* = (\overline{\mathrm{X}}_1, \cdots, \overline{\mathrm{X}}_k)'$, $\mathrm{Y}^* = (\overline{\mathrm{Y}}_1, \cdots, \overline{\mathrm{Y}}_k)'$，再根據此 k 筆資料，我們進行迴歸係數的最小平方估計，並以此估計結果當作整體資料的迴歸係數估計，此法稱爲平均數最小平方估計法 (least square estimation on the average)，其估計量記爲

$$\hat{\beta}_{\mathrm{LA}} = (\mathrm{X}^{*'}\,\mathrm{X}^*)^{-1}\,(\mathrm{X}^{*'}\,\mathrm{Y}^*),$$

直接推導，我們有 $E(\hat{\beta}_{\mathrm{LA}}) = \beta$ 且 $Cov(\hat{\beta}_{\mathrm{LA}}) = \sigma^2\,[\sum_{i=1}^{k}(\mathrm{X}_i'\mathrm{J}\mathrm{X}_i)^{-1}]/n$ ，其中 J 爲所有元素均爲 1 的 $n \times n$ 矩陣。注意：平均數最小平方估計法與前二種估計方法最大差異在於本法明顯降低資料的筆數，資料數從原始 N 筆降至爲 k 筆。

3.3 分組殘差變異一致性與迴歸係數重疊性之檢定

本節我們將探討常態誤差下，使用平均估計法的合理性，亦即各分組資料有相同迴歸係數向量 β 及相同殘差分佈。故而各分組資料除了均需滿足迴歸分析基本假設外，還需要在各組間殘差變異一致下，討論各分組迴歸係數是否具有重疊性。針對這 k 組資料，我們分別假設其線性迴歸模式的矩陣表示式爲

$$\mathrm{Y}_i = \mathrm{X}_i \beta^i + \epsilon^i, \quad i = 1, \cdots, k$$

其中 X_i 如同 (3)，β^i 爲第 i 組迴歸係數向量，ϵ^i_j 爲獨立且具有 $E(\epsilon^i_j) = 0$ 及 $Var(\epsilon^i_j) = \sigma^2_i$ 之常態分配。首先，我們考慮各組間殘差變異一致性的問題，若各組殘差變異有明顯差異，我們建議使用加權 (weighted) 平均估計法來取代平均估計法估計整體資料的迴歸係數。因此，檢定各組殘差變異是否有明顯差異是平均估計法之首要前題，其檢定假設爲

$$\mathrm{H}_0 : \sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_k = \sigma^2 ,$$

H_1 : 至少有一等式不成立。

我們可用拔雷特式檢定 (Bartllett's test)，進行檢查各組間殘差變異是否有明顯差異，相關文獻請參閱 Anderson (1958), Chao and Glaser (1978), Draper and Smith (1981) 及 Graybill (1976)。

接續，我們討論各組迴歸係數重疊性的問題，其檢定假設爲

$$\mathrm{H}_0 : \beta^1 = \beta^2 = \ldots = \beta^k = \beta ,$$

H_1 : 至少有一等式不成立。

在各組間殘差變異一致下，我們推導出其檢定統計量具有 $\mathrm{F}((k-1)p, (n-p)k)$ 分配；相關證明請參閱附錄。若僅考慮二組簡單線性迴歸係數估計比較問題，此時 $p = 2$, $k = 2$ ，則檢定統計量具有 $\mathrm{F}(2, (N-4))$ 分配。

4. 使用變通估計法之適當時機

4.1 適當時機的判斷

在第 3 節中我們提出平均估計法與平均數最小平方估計法，二個簡便的變通估計方法。本節，我們以簡單線性迴歸爲例，說明使用上述變通估計法的適當時機。根據 (2)，當 $p=2$ 時，我們知道傳統最小平方估計量 $\tilde{\beta}=[\tilde{\beta}_0,\tilde{\beta}_1]'$ 之共變異數矩陣爲

$$\mathrm{Cov}(\tilde{\beta})=\begin{bmatrix} \frac{1}{N}+\frac{\bar{x}_{..}^2}{\sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{..})^2} & \frac{-\bar{x}_{..}}{\sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{..})^2} \\ \frac{-\bar{x}_{..}}{\sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{..})^2} & \frac{1}{\sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{..})^2} \end{bmatrix}\sigma^2$$

其中 $\bar{x}_{..}=\sum_{j=1}^{n}\sum_{i=1}^{k}x_{j,i}/N$ 。在給定 N 及 σ^2 下，$\tilde{\beta}_0$ 和 $\tilde{\beta}_1$ 的變異數會受自變數 $x_{j,i}$ 的間距影響，亦即受 $\sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{..})^2$ 大小影響。若其值越大，則 $\tilde{\beta}_0$ 和 $\tilde{\beta}_1$ 的變異數越小。由於

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{..})^2 &= \sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{.,i}+\bar{x}_{.,i}-\bar{x}_{..})^2 \\ &= \sum_{j=1}^{n}\sum_{i=1}^{k}(x_{j,i}-\bar{x}_{.,i})+n\sum_{i=1}^{k}(\bar{x}_{.,i}-\bar{x}_{..})^2,\end{aligned}$$

我們知道整體自變數的間距和 SSE 爲組內間距和 SSW 與組間間距和 SSB 之加總，上式可記爲

$$\mathrm{SSE}=\mathrm{SSW}+\mathrm{SSB}。$$

回顧 3.1 節迴歸係數之平均估計法，因其在分組後先在各組內進行迴歸係數估計，其迴歸係數的估計變異主要源自於組內間距；在 3.2 節中，迴歸係數之平均數最小平方估計法，其主要在分組後以每組的樣本平均數作最小平方估計，其迴歸係數的估計變異主要源自於組間間距。因爲主要間距貢獻較大

者，其對應的變通估計方法在迴歸係數估計上會有較小的變異數。因此，探討自變數間距主要貢獻來自何處，用以決定何種變通估計方法較適合。綜合分析如下：

(i) 當 SSE $\approx$ SSW 時，其指整體資料自變數間距和大小與分組資料的組內自變數間距和相近，此時則建議使用平均估計法，例如資料集爲例行性收集或分組資料爲全部資料的縮影；

(ii) 當 SSE $\approx$ SSB 時，即自變數間距和與組間間距和相近，則建議使用平均數最小平方估計法，例如資料集爲依自變數大小收集。因此，我們有下列建議：

(i) 當 $\frac{\text{SSB}}{\text{SSW}} \ll 1$, 建議使用平均估計法，

(ii) 當 $\frac{\text{SSB}}{\text{SSW}} \gg 1$, 建議用平均數最小平方法。

4.2 模擬分析

本模擬分析主要探討在簡單線性迴歸模式下使用平均估計法與平均數最小平方估計法的適當時機。模擬方法如下，考慮簡單線性迴歸模式

$$y_{j,i} = \beta_0 + \beta_1 x_{j,i} + \epsilon_{j,i}$$

其中自變數 $x_{j,\ i}$ 模擬自範圍 $(0, 10)$ 間的均勻分配 (uniform distribution)；誤差項 $\epsilon_{j,i}$ 爲獨立且具有 $E(\epsilon_{j,i}) = 0$, $Var(\epsilon_{j,i}) = \sigma^2$ 的常態分配，N 、k 及 n 分別爲資料總筆數、分組數及每組的資料筆數。本模擬分析爲比較起見，將 N 取爲 10^5 筆以方便將二變通估計法與傳統最小平方法作一比較。注意：事實上，在 N 很大時，受限於計算工具之記憶體大小，直接計算最小平方估計量基本上是不可行，需利用矩陣分割法間接對迴歸係數作最小平方估計。

表一說明當分組資料爲整體資料之縮影時，使用平均估計法的估計標準差會較平均數最小平方法所得的估計標準差小，括弧內爲樣本標準差。表二則說明當資料集爲依自變數大小收集使用分段取樣時，使用平均數最小平方法的估計標準差會較使用平均估計法所得的估計標準差來的小。同時，由表

一及表二均可發現使用平均數最小平方估計誤差項的標準差均明顯比使用平均估計法所得的估計標準差大；這是因爲利用平均數最小平方法所得的均方差 (Mean Square Error) 爲估計 σ^2/n ，而使用平均估計法所得的均方差爲直接估計 σ^2 所造成。

表一 全體資料爲例行性收集 ($N = 10^5$, k =10, $n = 10^4$)

眞值		傳統最小平方法	平均估計法	平均數最小平方法
β_0	1	0.997 (0.009)	0.997 (0.009)	0.986 (0.709)
β_1	2	2.000 (0.001)	2.000 (0.001)	2.002 (0.118)
σ^2	1	1.000 (0.005)	1.000 (0.005)	1.000 (0.522)
β_0	10	9.997 (0.009)	9.997 (0.009)	9.986 (0.709)
β_1	20	20.000 (0.001)	20.000 (0.001)	20.042 (0.118)
σ^2	1	1.000 (0.005)	1.000 (0.005)	1.000 (0.522)

表二 全體資料爲依自變數大小收集 ($N = 10^5$, k =10, $n = 10^4$)

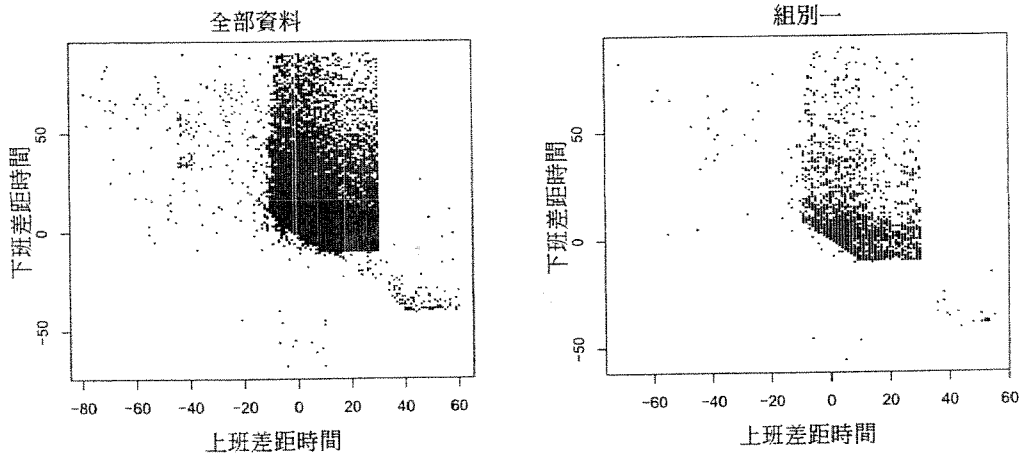
	眞值	傳統最小平方法	平均估計法	平均數最小平方法
β_0	1	0.999 (0.006)	1.000 (0.059)	0.999 (0.006)
β_1	2	1.999 (0.001)	1.999 (0.010)	1.999 (0.001)
σ^2	1	1.001 (0.005)	1.001 (0.005)	0.947 (0.455)
β_0	10	10.000 (0.006)	10.000 (0.060)	10.000 (0.006)
β_1	20	19.999 (0.001)	20.000 (0.010)	19.999 (0.001)
σ^2	1	1.000 (0.005)	1.000 (0.005)	1.081 (0.620)

5. 個案分析

本個案收錄某公立機關員工上下班打卡時間資料集，用於分析的欄位計有：工作日期、員工編號、上班打卡時間、下班打卡時間等。本資料集共記錄 42 個工作天，時間爲 2002 年 4 月 16 日至 2002 年 6 月 16 日止，共有 25,025 筆。該機關標準上、下班時間分別爲上午 8 點與下午 5 點，並且有實

施彈性上班，我們有興趣了解該機關彈性上班族羣與一般正常上下班族羣上班差距時間與下班距時間的統計關聯。我們定義上班差距時間 (x) 爲上班標準時間減去上班打卡時間，例如某員工上午 7 點 50 分到，比標準時間早到 10 分鐘，計爲 +10；若 8 點 10 分到，晚到 10 分鐘，計爲 −10。另外定義下班差距時間 (y) 爲下班打卡時間減去下班標準時間，例如某員工 17 點 10 分下班打卡，比標準時間晚退 10 分鐘，計爲 +10；若 16 點 50 分退，提早離開 10 分鐘，計爲 −10。

首先，我們考慮含有彈性上班族羣的資料，針對該機關員工上、下班時間與標準上、下班時間差距前後 90 分鐘內的資料進行分析，共有 22,800 筆資料滿足上述要求。我們依時間順序分爲十組，圖一中，左圖爲全體資料的上班差距時間與下班差距時間之散佈圖，右圖第一組的散佈圖，其餘九組之散佈圖亦類似。



圖一　上班差距時間與下班差距差距時間之散佈圖

由於資料集各分組的散佈圖極爲類似，在此我們僅報告第一組資料的迴歸分析結果。經由配適簡單線性迴歸模式後，變異數分析 (ANOVA) 表爲

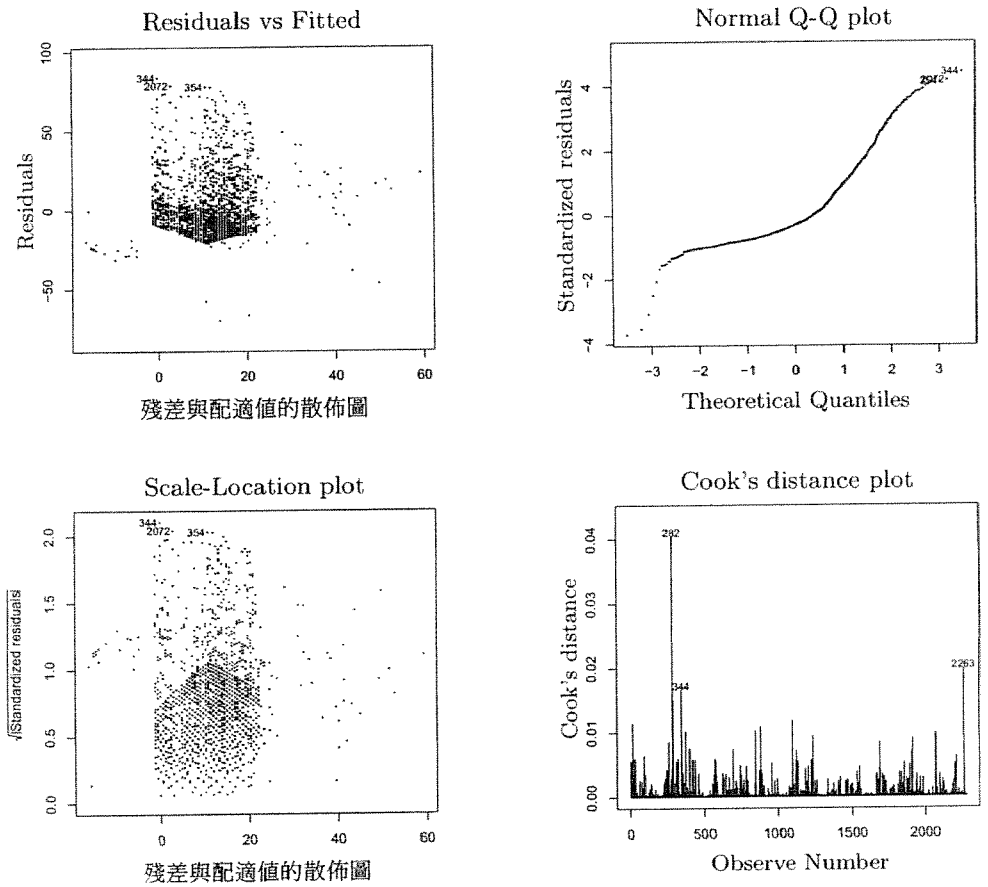
Source of variation	SS	df	MS	F-value	p-value
Regression	115355	1	115355	326.54	0.000
Error	804741	2278	353		

	估計值	標準誤	t-value	p-value
截距	16.724	0.473	35.360	0.000
x	−0.599	0.033	−18.070	0.000

其迴歸直線爲

$$y = 16.724 - 0.599x \circ$$

有關殘差檢查之結果列於圖二中，其中左上圖顯示殘差與配適值呈凌亂散佈並無特殊趨勢；右上圖呈現殘差的常態 Q-Q 圖可檢驗常態分布；左下圖爲配適值與標準化後殘差的散佈圖；右下圖利用 Cook's 距離檢查出該分組資料並無異常點。



圖二 彈性上班資料分組後的第一組之殘差檢查圖

在完成各分組資料迴歸分析及殘差檢查後，根據 3.3 節，我們先處理分組間殘差變異一致性與迴歸係數重疊性之檢定，其資料分析結果如下：

(i) 分組間變異數一致性檢定（Bartllett's test）：p-value $= 0.000$，

(ii) 迴歸係數重疊性檢定：p-value $= 0.000$，

(iii) SSB / SSW $= 0.002 \ll 1$。

這說明了本組資料分組間殘差變異程度並不太一致，進而會影響迴歸係數重疊性的檢定結果。若忽略此二前置處理程序，在簡單線性迴歸的模式下，根據 (iii) 及 4.1 節的結論，我們建議採用平均估計法，其迴歸直線爲

$$y = 16.031 - 0.527x,$$

表示在標準時間以前上班的員工，平均至少會晚 16 分鐘下班；同時，員工上班時間每提早 10 分鐘到者平均會晚 10 分鐘下班，而遲到的員工會自動延長下班時間平均遲到 10 分鐘者會晚 21 分下班。爲了利於比較，我們利用第 2 節的矩陣分割法，間接計算整體資料的迴歸係數最小平方估計，並將平均數最小平方估計法之結果列於表三中。

表三　三種估計法之比較(N=22,800, k=10, n=2,280)

	傳統最小平方估計法	平均估計法	平均數最小平方法
$\hat{\beta}_0$	16.050	16.031	17.010
$\hat{\beta}_1$	−0.528	−0.527	−0.648
$\hat{\sigma}^2$	354.774	354.774	1,443.920

其次，我們考慮非彈性上班族羣的資料，亦即上班時間差距在正負 10 分鐘內的資料，滿足此條件共有 14,850 筆資料。我們亦依序將此資料分十組，其散佈圖與圖一類似，故略去。對於本資料分組後的迴歸模式建構及殘差檢查類似彈性上班族羣，並無顯著違反迴歸模式的假設條件，其殘差檢查圖類似圖二，故略去。經執行分組殘差變異一致性與迴歸係數重疊性之檢定，其數據分析結果如下：

(i) 分組間變異數一致性檢定 (Bartllett's test)：p-value =0.000，

(ii) 迴歸係數重疊性檢定：p-value = 0.123，

(iii) SSB / SSW = 0.248 $<$ 1。

忽略資料分組間殘差變異程度不一致的問題，在顯著水準爲 0.05 下，本組資料滿足迴歸係數重疊性檢定。因此，採用平均估計法，其迴歸直線爲

$$y = 14.157 - 0.929x,$$

表示在標準時間以前上班的員工，平均至少會晚 14.157 分鐘下班；同時，員工上班時間每提早 10 分鐘到者平均晚5分鐘下班，而遲到 10 分鐘的員工會自動延長 25 分鐘下班，其比較結果列於表四中。

表四 三種估計法之比較(N=14,850, k=10, n=1,485)

	傳統最小平方估計法	平均估計法	平均數最小平方法
$\hat{\beta}_0$	14.135	14.157	16.578
$\hat{\beta}_1$	−0.924	−0.929	−2.631
$\hat{\sigma}^2$	3768.893	3767.100	4266.400

最後，我們考慮正常上下班族羣的資料，亦即上班時間差距在正負 10 分鐘內且下班時間差距在正負 5 分鐘內的資料，滿足此條件共有 4,800 筆資料。我們亦依序將此資料分十組，其散佈圖與圖一類似，故略去。對於本資料分組後的迴歸模式建構及殘差檢查類似彈性上班族羣，並無顯著違反迴歸模式的假設條件，其殘差檢查圖類似圖二，故略去。經執行分組殘差變異一致性與迴歸係數重疊性之檢定，其數據分析結果如下：

(i) 分組間變異數一致性檢定 (Bartllett's test)：p-value=0.084，

(ii) 迴歸係數重疊性檢定：p-value = 0.055，

(iii) SSB / SSW = 0.02 $\ll$ 1。

在顯著水準爲 0.05 下，本組資料滿足分組間變異數一致性檢定與迴歸係數重疊性檢定。因此，採用平均估計法，其迴歸直線爲

$$y = 2.028 - 0.272x,$$

表示在標準時間上下班的員工，平均晚 2.028 分鐘下班；同時，此族羣雖然提早 10 分鐘上班卻會準時下班，堪稱標準上班族，其比較結果於表五中。由以上分析發現，該機關施行彈性上班制度是有利於提昇員工工作時間。

表五　三種估計法之比較(N=4,800, k=10, n=480)

	傳統最小平方估計法	平均估計法	平均數最小平方法
$\hat{\beta}_0$	2.033	2.028	2.941
$\hat{\beta}_1$	−0.273	−0.272	−0.542
$\hat{\sigma}^2$	6.289	6.275	14.400

註二：本個案分析之資料收集使用的作業系統爲 Unix，其資料庫爲 Informix 資料庫系統；在統計分析方面使用的作業系統爲 Window98，利用免費的統計軟體 R 進行分析。在資料讀取上透過資料庫連接 (Open Database Connection 簡稱 ODBC) 輔助，讀取後端資料庫的資料，充分利用資料庫專負責資料儲存，統計軟體負責資料分析的優點。

6. 結論

超大資料集的統計分析問題，主要在於一般計算工具普遍存在電腦執行計算或儲存上記憶體不足的問題。針對線性迴歸分析，本文提出利用矩陣分割運算，使得傳統最小平方法在超大資料集中依然可以計算，但需要適度的修正現行統計軟體的計算公式。雖然矩陣分割運算法可解決傳統最小平方估計量之計算，但對於迴歸分析中的殘差資料檢驗工作依舊受限於超大資料集之分析問題。本文根據分組資料，首先進行各組間殘差變異數一致性的檢定，從而進行各組間迴歸係數重疊性之檢定。在上述二種前置檢定通過下，根

據分組資料中自變數爲全體資料之縮影或依自變數大小分組收集，我們分別建議平均估計法或平均數最小平方估計法作爲簡便的變通方法，其特點在於所使用較的記憶體較小，且計算時間較短，但仍然具有相當的精確度。

致謝詞：作者感謝評審們的寶貴意見使文章更爲流暢易讀。

附錄

在常態誤差假設下，若各組間殘差變異一致，有關各分組迴歸係數重疊性的問題，其對等於檢定

$$\mathrm{H}_0 : \beta^1 = \beta^2 = \ldots = \beta^k = \beta ,$$

$$\mathrm{H}_1 : \text{至少有一等式不成立。}$$

我們可推導出其概似比 (likelihood ratio) 爲

$$\Lambda \propto \left(\frac{\mathrm{SSE}}{\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{SSE}_i} \right)^{\frac{-N}{2}}$$

其中 SSE 表全體資料的誤差平方和， SSE_i 爲第 i 組的誤差平方和。根據矩陣運算公式，我們知道$\mathrm{SSE} = \mathrm{Y}'(\mathrm{I}_N - \mathrm{H})\mathrm{Y}$, $\mathrm{SSE}_i = \mathrm{Y}_i'(\mathrm{I}_n - \mathrm{H}_i)\mathrm{Y}_i$，其中 $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$, $\mathrm{H}_i = \mathrm{X}_i(\mathrm{X}_i'\mathrm{X}_i)^{-1}\mathrm{X}_i'$， I_N 及 I_n 爲對應的單位矩陣。由於

$$\sum_{i=1}^{k} \mathrm{SSE}_i = \sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{Y}_i' \left(\mathrm{I}_n - \mathrm{H}_\mathrm{i}\right) \mathrm{Y}_\mathrm{i} = \mathrm{Y}' \left(\mathrm{I}_N - \mathrm{H}^*\right) \mathrm{Y}$$

其中

$$\mathrm{H}^* = \begin{bmatrix} \mathrm{H}_1 & 0 & 0 & 0 \\ 0 & \mathrm{H}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathrm{H}_k \end{bmatrix}_{\mathrm{N}\times\mathrm{N}},$$

我們有

$$\mathrm{SSE} = \sum_{i=1}^{k} \mathrm{SSE}_i + \mathrm{Y}'(\mathrm{H}^* - \mathrm{H})\mathrm{Y} = \mathrm{A} + \mathrm{B},$$

其中 $\mathrm{A} = \sum_{i=1}^{k} \mathrm{SSE}_i$，$\mathrm{B} = \mathrm{Y}'(\mathrm{H}^* - \mathrm{H})\mathrm{Y}$。在 H_0 之下，我們知道 SSE 及 A 具有卡方分配 (Chi-square)，自由度分別爲 $N-p$ 及 $(n-p)k$。若能證明 A 與 B 獨立，則 B 將具有卡方分配，自由度爲 $(k-1)p$，且概似比檢定統計量與 B/A 成比例，故問題的關鍵在於證明 A 與 B 是否獨立，下列兩個性質可幫助證明 A 與 B 獨立。

性質 1: $\mathrm{H}^*\mathrm{H}^* = \mathrm{H}^*$。

證明: $\mathrm{H_i H_i} = \mathrm{H_i}$，再根據 H^* 的定義，直接計算即可得證。

性質 2: $\mathrm{H}^*\mathrm{H} = \mathrm{H}$。

證明: 根據 $\mathrm{H_i}$ 的定義，我們有

$$\begin{bmatrix} 0 & \cdots & \mathrm{H}_i & \cdots & 0 \end{bmatrix} \mathrm{X} = \begin{bmatrix} 0 & \cdots & \mathrm{H}_i & \cdots & 0 \end{bmatrix} \begin{bmatrix} \mathrm{X}_1 \\ \vdots \\ \mathrm{X}_i \\ \vdots \\ \mathrm{X}_k \end{bmatrix} = \mathrm{X}_i \circ$$

直接計算可以得

$$H^*H = \begin{bmatrix} H_1 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & H_i & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & H_k \end{bmatrix} X(X'X)^{-1}X' = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_k \end{bmatrix} (X'X)^{-1}X'$$

$$= X(X'X)^{-1}X' = H,$$

故得證。

性質 3: A 與 B 爲統計獨立。

證明: 由於 $A = \sum_{i=1}^{k} SSE_i = Y'(I_N - H^*)Y$ 及 $B = Y'(H^* - H)Y$ ，僅需證明

$$(I_N - H^*)(H^* - H) = 0。$$

根據性質 1 與性質 2，可得

$$(I_N - H^*)(H^* - H) = H^* - H - H^*H^* + H^*H = 0,$$

故得證，相關文獻請參考趙民德 (1999)。

根據性質 3，迴歸係數重疊性檢定之檢定統計量爲

$$F = \frac{B}{(k-1)p} \Big/ \frac{A}{(n-p)k} = \frac{Y'(H^* - H)Y}{Y'(I_N - H^*)Y} \cdot \frac{(n-p)k}{(k-1)p},$$

其具有 $F((k-1)p, (n-p)k)$ 分配。

參考文獻

林億雄、林共進 (2001)。龐大資料集的統計推論方法。國立成功大學統計學報，**24** (Data Mining專刊)，85-99。

趙民德 (1999)。淺談不相關的二次式。中國統計學報，**37**，309-318。

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

Chao, M.T. and Glaser, R. E. (1978). The exact distribution of Bartlett's test statistic for homogeneity of variance with unequal sample sizes. *J. Amer. Statist. Asso.* **73**, 422-426.

Christensen, Ronald (1984). *Plane Answers to Complex Questions: The Theory of Linear Models.* Springer-Verlag, New York.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis.* 2nd Ed. Wiley, New York.

Graybill, F. A. (1976). *Theory and Application of the Linear Model.* Belmont, CA, Wadsworth.

——— (1982). *Matrices with in Statistics.* 2nd Ed. Belmont, CA, Wadsworth.

Hand, D. J., Blunt, G., Kelly, M. G. and Adams, N. M. (2000). Data Mining for fun and profit. *Statistical Sciences* **15**, 111-131.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Statistical Models.* Wiley, New York.

Seber, G. A. F. (1977). *Linear Regression Analysis.* Wiley, New York.

[民國91年9月23日收稿，民國92年12月6日接受。]

Regression Analysis for Large Dataset

Mei-Mei Zen[1] Yi-Hsiung Lin[1] Dennis, K. J. Lin[2]

[1]Department of Statistics

National Cheng-Kung University

[2]Department of Management Science and Information System

Penn State University

ABSTRACT

Under the normal assumption, the least square estimator is also the maximum likelihood estimator, which has certain well-known properties. The statistical analysis on a large dataset, however, is usually limited by the computing memory and time. We consider the estimation problem of regression analysis on large datasets. Two alternative estimators are proposed here to avoid the problem of lack of computing memory and time; (1) averaging estimator and (2) least square estimator on the average. First, to save the computing meory and time, we use the techniques of matrix partition to make the traditional least square estimator feasible. Secondly, for the partitioned subgroups of the dataset, we deal with the homogeneity of variation and the coincidence of the regression model for all subgroups. Comparison among these estimation methods are made. A real life example is used for illustration.

Key words and phrases: Averaging estimator, data mining, least square estimator, least square estimator on the average, matrix partition.

AMS 2000 subject classifications: Primary 62J05; secondary 62F03.