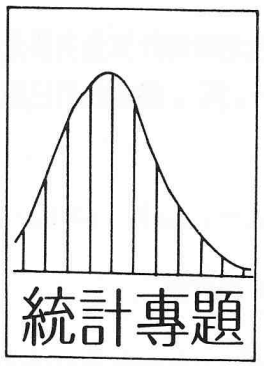
3. Kruskal , W. H. and Tanur , J . M , *International Encyclopedia of statistics* , The Free Press , Vol 2 , p 705 , 1978.
4. Hillier , F. and Lieberman , G. J . , *Introduction to Operations Research* , 4th Edition , Holden - Day , Inc. , 1986.
5. 孫自健，石仲拓，談談卜松過程，數學傳播第二卷第三期，民國67年。
6. Karlin , S. and Taylor , H. M. , *A First Course in Stochastic Proresses* , 2nd ed. Academic Press , New York , 1975.

—本文作者任教於國立中山大學應用數學研究所—



漫話線性迴歸

梁文騏

加拿大某個計算機中心曾經作過一項有趣的統計，即對上機算題者所解算的題目及上機時間進行統計，結果發現，全部上機時間的一半以上，是在計算線性迴歸。

這大概可以說明兩個問題，一是計算線性迴歸的次數頻繁，二是計算線性迴歸所需的上機時間長。

爲什麼會是這樣呢？

統計學大師H. Cramér 曾經說過，統計的主要目的是預報。其實，不獨統計如此，整個科學的價值，主要的也就是在於它的預見性，即根據給定的條件，指出將會出現何種結果。假使僅祇是局限於在結果已經出現了之後，解釋這種結果何以會出現，則解釋得再好，也不免是事後諸葛亮。當然，並不是說事後的檢討於事無補，往事不忘，後事之師。檢討往事

的價值，全在於產生後事的預見。

據以預報的那些條件或因素，我們以 t_1，…，t_K來表示，而預報的對象以 y 來表示。t_1，…，t_K 以及 y 可以是定量的，也可以是定性的。但是定性的東西也可以用定量的方式來表達，例如某種醫學檢驗的陽性反應和陰性反應可以用 1 和 0 來表示等等。所以，可將 t_1，…，t_K以及 y 都看作是數量，t_1，…，t_K 稱爲預報因子，y 稱爲預報量。

如果 t_1，…，t_K 包羅了一切能夠影響 y 的因素，那麼 t_1，…，t_K 確定了之後，y 的值就應該確定，換言之，y 即是 t_1，…，t_K 的函數 $f(t_1,\cdots,t_K)$，雖然此函數 f 的具體形式我們並不知道。考慮到實際上隨機干擾總是難以避免的，以 e 表示這個隨機干擾的數量，則我們可以寫出如下的關係式：

$$(1)\qquad y=f(t_1,\cdots,t_K)+e\ ,$$

其中 t_1，…，t_K 是已知的因子值，f 是未知的函數，e 是無法測知的隨機干擾量，而 y 是未知的預報量。

在多數情形下，我們有相當的理由，例如根據概率論的中心極限定理或高斯誤差律，假定 e 是一個遵守常態分布 $N(0,\sigma^2)$ 的隨機變量。其次，我們可以假定 f 具有泰勒展式，因爲幾乎在一切實際場合中，f 至少可以用一個具有泰勒展式的函數來無限逼近，寫出 f 的泰勒展式，

$$(2)\quad f(t_1,\cdots,t_K)$$

$$=\sum_{j=0}^{N}\frac{1}{j!}\Big[\sum_{i=1}^{K}(t_i-t_{i0})\frac{\partial}{\partial t_i}\Big]^j f(t_{10},\cdots,t_{K0})+R_N,$$

其中 R_N 爲泰勒展式餘項。由此可見，上式右側一般項是未知函數 f 的一個若干階偏導數在 $(t_{10},\cdots,t_{K0})$ 點的未知值 β 和 $t_1,\cdots,t_K$ 的一個已知多項式 $p(t_1,\cdots,t_K)$ 之乘積。若作變數替換，令此已知多項式 $p(t_1,\cdots,t_K)=x$，則(2)式右側一般項即形如 βx。由於(2)式右側含有若干個這種形狀的項，所以(2)式可寫成下列形狀

$$(3)\qquad f(t_1,\cdots,t_K)=\sum_{i=1}^{k}\beta_i x_i+R_N\ 。$$

將(3)式代入(1)式，並且由於餘項 R_N 可以任意小，忽略 R_N 不計，即得

$$(4)\qquad y=\sum_{i=1}^{k}\beta_i x_i+e\ ,$$

其中各 β_i 爲未知係數，各 x_i 爲已知的（經變換後的）預報因子，e 爲遵守常態分布 $N(0,\sigma^2)$ 的隨機干擾項，σ^2 爲未知參數。

(4)式即稱爲線性迴歸模型。

我們看到，線性迴歸模型(4)是在非常一般的假定下導出的，這可以說明人們爲什麼會頻繁地計算線性迴歸。

其次，線性迴歸的計算是否需要相當長的上機時間呢？

我們知道，線性迴歸的計算量主要是在於估計未知係數 $\beta_1,\cdots,\beta_k$。假定我們已積累了 n 組觀測數據

$$(y_i;x_{1i},\cdots,x_{ki}),\quad i=1,\cdots,n\ ,$$

於是我們即可列出 n 個方程

$$(5)\quad \begin{cases} y_1=\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_kx_{1k}+e_1\\ y_2=\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_kx_{2k}+e_2\\ \qquad\vdots\\ y_n=\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_kx_{nk}+e_n \end{cases}$$

採用向量和矩陣的表示法，令（黑體字母表示向量和矩陣）

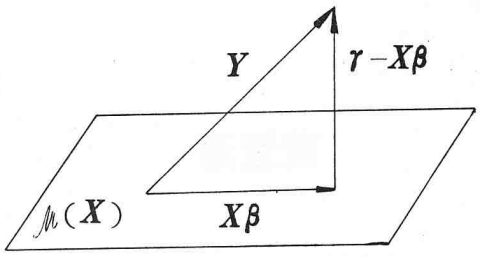
$$\boldsymbol{Y}=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix},\ \boldsymbol{X}=\begin{pmatrix}x_{11}\cdots x_{1k}\\ \cdots\\ x_{n1}\cdots x_{nk}\end{pmatrix},$$

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_1\\ \vdots\\ \beta_k\end{pmatrix},\ \boldsymbol{e}=\begin{pmatrix}e_1\\ \vdots\\ e_n\end{pmatrix},$$

則方程組(5)即寫爲

(6) $$Y=X\beta+e$$

β的最大似然估計和最小二乘估計$\hat{\beta}$是使"殘差向量"$Y-X\beta$的模達到最小。如以$\mathcal{M}(X)$表示X的列向量所張成的空間，注意到$X\beta$是$\mathcal{M}(X)$中的一個向量，則我們有如下的示意圖：



欲使$Y-X\beta$的模最短，顯然$X\beta$應爲Y在$\mathcal{M}(X)$上的投影，也就是說$Y-X\beta$應垂直$\mathcal{M}(X)$，此等價於$Y-X\beta$垂直X的各列，因之我們得到

(7) $$X'(Y-X\beta)=0,\quad (X'\text{爲}X\text{的轉置})$$

此稱爲正規方程，由此立即得出解$\hat{\beta}$如下

$$\hat{\beta}=(X'X)^{-1}X'Y\ 。$$

依照(8)式來計算β的估計量$\hat{\beta}$，主要計算量是在於求逆方陣$(X'X)^{-1}$。$X'X$是一個k行k列的方陣，不論用何種方法求它的逆，算術運算的次數的數量級總是k^3。當預報因子的個數k不大時，k^3次算術運算對於電子計算機來說是輕而易舉的。但是，在許多實際問題中，可能影響預報量y的因子非常之多。例如假若要預報明日本地的最低氣溫y，則各地的從過去到現在，從地面到高空的各種氣象因素都對y有或強或弱的影響，如將它們一一列爲預報因子，則k將是無限大。任何電子計算機的存儲和速度都無法應付，即使是能夠計算，由於電子計算機所能夠進行的僅祇是在一定有效位數之內的近似計算，計算次數過多時，四捨五入的誤差的積累將會使得計算結果面目全非，毫無實際用途。

在實用上，大多將k限制在10以內，不過，在數目衆多的可供考慮的因子中，怎樣才能遴選出少數堪充實用的因子來呢？

例如，在氣象預報中，如果考慮10個觀測地點，每地考慮5－10個氣象因素，則即共有50－100個可供考慮的因子，如何從中選用10個因子呢？原則上說，一種選取法的效果優劣應由試算來檢定，現在一切可能的選取法共有$\binom{50}{10}\approx 10^{10}$乃至$\binom{100}{10}\approx 10^{13}$種，如果對每一種選取法都試算一次，僅以求逆方陣一項計算來說，每求一個10階方陣的逆姑以10^3次算術運算來計，即共需$10^{13}\sim 10^{16}$次算術運算。假使電子計算機每秒約可進行10^8次算術運算，則共需時約27小時以至三年，即使27小時猶可，三年則免談了。

由此觀之，對於線性迴歸中非常重要却又足以令人頭大的因子選擇問題，光靠計算機硬體的發展是不行的，就算計算速度再提高三個數量級，達到每秒$10''$次運算，那麼200個因子選10個共有$\binom{200}{10}\approx 10^{16.3}$種選取法，比$\binom{100}{10}$提高了三個數量級以上，仍是無法計算，所以必須發展統計和數學的軟體，來避免計算一切可能的選取法同時降低每種選取法的計算次數，這方面的研究已有大量成果並且方興未艾。

不過，從目前以至可預見的將來的硬體和軟體發展來說，計算中心的線性迴歸顧客高居首座的狀況，恐怕還要繼續下去。

—本文作者現為中央研究院統計所客座專家—