應用線性迴歸模型與貝氏理論在統計預測與決策之研究

王棣撰

國立政治大學統計研究所碩士論文

指導教授：鄭堯柈先生

應用線性迴歸模型與貝氏理論在統計預測與決策之研究

研究生：王　棣

中　華　民　國　七　十　四　年　六　月

謝　　詞

感謝指導教授鄭堯柈在論文撰寫過程中的辛勤指導，使本文得以完成。盧師敦義、鄭師隆輝兩位教授對論文提供許多寶貴意見，更是獲益良多，另姚師興台、于師琮嘉兩位教授在撰寫過程中所給予精神上的鼓勵和支持，特在此一併致謝。

謹以此文獻給愛我的父母及文立，更感謝主，在整個過程中，讓我經歷主愛的長闊高深。願我一生的年日都在祂手中。

目　　錄

第一章　緒　論

第一節　前言

統計理論與機率論發展的結果，引起了統計學者從事不確定事實之研究，而統計預測乃是以現在及過去的資料來誘導未來的方向，以便提供決策者作決策之參考或及時修改以定目標與政策，每一個人，於日常生活中，無時不在對所面對的事情作決策，也對未來作預測，如何根據科學的方法、理論的基礎、過去的經驗、歷史的資料以及現有的資料來進行預測與決策，乃是統計學者所應負的重要使命。

第二節　研究動機

傳統的統計推論都是由樣本資料導出統計量的分配，以對未知母數作推論，以爲決策者的參考，亦即完全根據由抽樣樣本所提供的資訊來作推論未知母體的參數，此即吾人所熟知的抽樣理論。然而過去資料的趨勢、人類的經驗累積、以及決策者所作的判斷、理論上的考慮，莫不對於尋求理解現象及對現今情勢分析有所助益。因此在進行

統計推論時，事前所擁有的資訊是不容忽視的，此事前資訊和抽樣資料的結合，經由貝氏定理的判斷，即可產生未知母體參數的事後機率分配，以此分配進行統計推論，此種方法稱爲貝氏分析法。

第三節 研究目的與研究方法

本文主旨在應用貝氏方法結合常態迴歸模型來進行預測與決策問題，並試將此一方法進行實證分析，更進而應用所探討的結果可以提供研訂經濟政策參考。

第二章　迴歸模型的意義與貝氏理論

應用貝氏方法探討迴歸模型，基本上，即應用貝氏定理有關觀念來作迴歸分析。貝氏方法的優點在於，能將有關參數的事前情報，用適當的數學方法與由試驗所獲知的情報相互結合在一起，此種事前情報或取自一般理論上的考慮，或來自以前試驗的結果。貝氏理論將母體參數的事先資料列入考慮，其本身不僅很有用，且加强我們對古典統計學的各項限制的瞭解。所以利用貝氏理論結合迴歸模型所求得的迴歸參數之事後分布，進行參數之推定、檢定與預測，此即貝氏方法在迴歸分析上應用的重點。因此本章先就迴歸模型的意義與貝氏決策的基本理論討論之。

第一節　迴歸模型的意義

迴歸分析是一種統計工具，用來表示兩個或兩個以上計量變數的關係，係研究如何根據有關變數以預測某一變數的數值。為兩個或兩個以上變數分析方法的一種。一般言，變數間關係可分為〔註一〕：

(一)函數關係（ Function Relation ）或稱確定關係（

Deterministic Relation ），其關係式以數學公式表示 $Y=f(x)$，若 X 是獨立變數（自變數），Y 是相依變數（因變數），已知 X 值，透過函數關係式，僅有一個與之對應的 Y 值存在，或者 Y 取決於數個 X，即 $Y=f(X_1,X_2,\cdots X_k)$，$k=1,2,\cdots,n$，則稱 X 與 Y 具有函數關係，又可稱數學模式（ Mathematical Model ）。例如：若產品銷售額 (Y) 與銷售單位 (X) 之關係如下式，每單位產品銷售 2 元，則得 $Y=2X$，將觀察值繪於圖中，則所有點皆落入直線上，此即函數關係之特性。

㈡統計關係（ Statistical Relation ）或稱隨機關係，統計關係不同於函數關係，若對每一 X 值而言，透過關係式，與之對應的是許多 Y 值所形成的一機率分配，則稱 X 與 Y 具有隨機關係。

人類生活領域中，有許多事物的理論關係是隨機的，而迴歸模式是用來表示一統計關係兩個基本要素的最佳方法，此用來表示兩個變數間最簡單的隨機關係稱為簡單直線迴歸模型，模式為

$$Y_i=\beta_0+\beta_1 X_i+\epsilon_i \qquad i=1,2,\cdots,n \quad (2\text{-}1\text{-}1)$$

其中

Y_i：第 i 觀察值的相依變數（對應變數、被解釋變數或稱被預測變數）。

β_0, β_1：未知迴歸參數（Parameters）。

X_i：已知常數，爲第 i 觀察值的獨立變數（解釋性變數或稱預測性變數）。

ϵ_i：隨機誤差（或稱干擾項），$\epsilon_i \sim N(0, \sigma^2)$

模式中，Y_i 爲相依變數，Y_i 的變動可由兩部分組成，(1)可解釋的變動，由獨立變數 X_i 的變動來解釋。(2)不可解釋的變動，此一部分的考慮又可分爲二方面來解釋，①因觀察或測量時所產生之誤差。②一些阻礙變數，其對相依變數的變動影響較小，未被列入，或有些因子無法計量。爲了便於分析，一般學者對(2-1-1)之迴歸模式，均採下列假設〔註二〕：

(1)常態性：$\forall i, i = 1 - n \Rightarrow \epsilon_i \sim$ Normal Distribution

(2)零均數：$\forall i, i = 1 - n \Rightarrow E(\epsilon_i) = 0$

(3)變異數齊一性：$\forall i, i = 1 - n \Rightarrow Var(\epsilon_i) = \sigma^2$，而此變異數性爲未知。

(4)非自我相關性（Nonautocorrelation）：$\forall i, j, i \neq j \Rightarrow E(\epsilon_i, \epsilon_j) = 0$

(5)如 X 爲非隨機變數，在重覆抽樣中，其值固定，即

X的值爲可控制的值，但多數X爲隨機變數，則其分配與β_0、β_1、σ^2相互獨立且ϵ_i與X無關。

而迴歸分析基本上就是利用所建立之最佳迴歸模型，利用被預測變數與其他預測性變數的相互關係，估測其未來的可能趨勢，所以迴歸分析的主要目的，其一爲迴歸模型中未知母數的估計及檢定問題；其二即根據估計結果以進行預測及鑑定的問題。

當我們所建立之變數間關係模型爲(2-1-1)時，有關迴歸分析之基本推論：

㈠母數的估計

1. 點估計：當隨機變數的分配爲已知，模型中的未知母數可用最概法加以估計，可得β_0、β_1與σ^2之估計量分別爲

$$\hat{\beta}_1 = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2} \qquad (2\text{-}1\text{-}2)$$

$$\hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X} \qquad (2\text{-}1\text{-}3)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \qquad (2\text{-}1\text{-}4)$$

當隨機變數Y（或ϵ）的機率分配爲未知，則推定未知母數不能用最概法，而應用最小平方法（The Method of Squares ），解得之 β_0，β_1 如 (2-1-2) 及 (2-1-3) 式，而 σ^2 為

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-2} \quad \Sigma (Y_i - \hat{Y}_i)^2$$

$$= \frac{SSE}{n-2}$$

且 $E(S^2) = \sigma^2$，故滿足不偏性。

2. 區間估計與檢定

(1) β_0 的區間估計：若隨機誤差 ϵ 為常態變數，

則 $\beta_0 = \overline{Y} - \beta_1 \overline{X} = \Sigma \left(\frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\Sigma (X_i - \overline{X})^2} \right) Y_i$

，即為常態變數之線型組合，故亦為常態變數

，且 $E(\hat{\beta}_0) = \beta_0$，$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} \left[1 + \frac{n\overline{X}^2}{\Sigma(X_i - \overline{X})^2} \right]$

，因 $\frac{\hat{\beta}_0 - \beta_0}{[Var(\hat{\beta}_0)]^{\frac{1}{2}}}$ 是具有 n－2 個自由度的 t

分配，所以 β_0 之區間估計和檢定是應用 t 分配。

其在信賴係數為 $1-\alpha$ 之信賴區間為

$$\hat{\beta}_0 - t_{\frac{\alpha}{2}} \sqrt{\frac{\Sigma X_i^2 \ \Sigma(Y_i - \hat{Y}_i)^2}{n(n-2)\Sigma(X_i - \overline{X})^2}} < \beta_0 < \hat{\beta}_0$$

$$+ t_{\frac{\alpha}{2}} \sqrt{\frac{\Sigma X_i^2 \ \Sigma(Y_i - \hat{Y}_i)^2}{n(n-2)\Sigma(X_i - \overline{X})^2}} \qquad (2\text{-}1\text{-}5)$$

(2) β_0 的檢定

設 H_0 : $\beta_0 = 0$ H_1 : $\beta_0 \neq 0$，在顯著水準爲 α

，若 $t = \dfrac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}}$ ，$|t| > t_{\frac{\alpha}{2}}$ ，則拒絕 H_0，

否則接受 H_0 。

(3) β_1 的區間估計

因 $\hat{\beta}_1 = \dfrac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2} = \dfrac{\Sigma(X_i - \overline{X})Y_i}{\Sigma(X_i - \overline{X})^2}$

亦爲常態變數，且 $E(\hat{\beta}_1) = \beta_1$，$Var(\hat{\beta}_1) = \dfrac{\sigma^2}{2(X_i - \overline{X})^2}$

，取 σ^2 的不偏估計值 MSE 來代替，則設信賴係數 $1-\alpha$，β_1 之信賴區間

$$\hat{\beta}_1 - t_{\frac{\alpha}{2}} \frac{(MSE)^{\frac{1}{2}}}{\sqrt{\Sigma(X_i - \overline{X})^2}} < \beta_1 < \hat{\beta}_1 + t_{\frac{\alpha}{2}} \frac{(MSE)^{\frac{1}{2}}}{\sqrt{\Sigma(X_i - \overline{X})^2}} \qquad (2\text{-}1\text{-}6)$$

當 $\Sigma(X_i - \overline{X})^2$ 愈大，則 $Var(\hat{\beta}_1)$ 愈小，而 β_1 的估計結果愈爲可靠。

(4) β_1 的檢定

β_1 爲 X 變動一單位，Y 平均變動多少單位之變率測度。依此吾人說 Y 與 X 無直線關係，相當於設 $\beta_1 = 0$，因此在檢定假設時，定

$H_0 : \beta_1 = 0$　　　　$H_1 : \beta_1 \neq 0$

在顯著水準爲 α，其決策法則爲 $|t| > t_{\frac{\alpha}{2}}$，則接受 H_1，否則接受 H_0。

(二)預測

設 X，Y 兩變數之直線型迴歸模式爲

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

之假設下

$E(Y \mid x_k) = \overline{Y}_k = \beta_0 + \beta_1 x_k$

代表 $X = X_k$ 時 Y 的均值，即 $\mu_{Y|x_k}$。此關係依樣本 料估計得

$$\hat{Y}_k = \beta_0 + \beta_1 x_k \qquad (2\text{-}1\text{-}6)$$

故 $\hat{Y}_k$ 即爲 $X = x_k$ 時的預測值（Predicted Value），此預測值又分爲預測的期望值，即 $E(Y|x_k) =$

$\mu_{Y|x_k}$ 。(2) Y 的一數值，即 $Y_k = \beta_0 + \beta_1 x_k + \epsilon_k$

當 $X = x_k$ 時，Y 的眞值爲 Y_k（通常未知），期望值爲 $\overline{Y}_k$，預測值 $\hat{Y}_k$，其預測誤差 $\epsilon_k = Y_k - \hat{Y}_k$，

$$\sigma^2(Y_k - \hat{Y}_k) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_k - \overline{X})^2}{2(X_i - \overline{X})^2}\right]$$

故 Y_k 之 $(1-\alpha) \times 100\%$ 的信賴區間爲

$$\hat{Y}_k - t_{\frac{\alpha}{2}} S\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}} < Y_k < \hat{Y}_k + t_{\frac{\alpha}{2}} S\sqrt{1 + \frac{1}{n} + \frac{(x_k - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}} \qquad (2\text{-}1\text{-}7)$$

又因 $\sigma^2_{Y_{\hat{k}}} = \sigma^2\left[\frac{1}{n} + \frac{(X_k - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}\right]$，用 $\hat{Y}_k$ 估計 $\overline{Y}_k$ ($= \beta_0 + \beta_1 x_k$)，則 $\overline{Y}_k$ 的 $(1-\alpha) \times 100\%$ 之信賴區間爲

$$\hat{Y}_k - t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n} + \frac{(x_k - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}} < Y_k < \hat{Y}_k + t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n} + \frac{(x_k - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}} \qquad (2\text{-}1\text{-}8)$$

第二節　迴歸模型之建立

建立迴歸方程模型用於統計預測與決策，首先須擬訂預測計劃、進而尋求預測項目相關之自變數，如果所列自變數項目較多時，易發生變數間線性重合現象；同時亦難免有自我相關之困難。倘在建立模型時能循邏輯程序，固可減少以上兩問題所產生之結果，亦可減少計算所費之時間。

有關迴歸方程模型建立之五個步驟。

㈠選擇相關自變數

運用迴歸模型主要用途有三：⑴敍述，⑵控制，⑶預測。爲達此目的，須先明瞭應變數之特性，確定其範圍，儘可能設想到所有可能影響這個應變數有關係之自變數。

尋找自變數的方法，可由理論得知與應變數有關之自變數；或類於過去之統計觀察數列以尋找相關自變數。

㈡蒐集自變數資料

有關自變數項目選擇後，即進行蒐集與整理各該自變數。對於數量資料不易獲得可資利用者，或屬不很重要之自變數，或自變數間線性重合者，應予妥善之選擇，必要時予以合併或剔除。並進一步討論留下之自變數，是否

足以合理解釋應變數，若仍不足，則放棄此一模型。

㈢設計迴歸模型之選擇

蒐集自變數確定後，即設立模型，一般而言，迴歸模型可分爲線性與非線性，因非線性迴歸模型均可經由適當的轉換成線性模型，以便於統計之推定、檢定。

假設此例之迴歸模型爲

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots\cdots + \beta_p X_{pt} + \epsilon_t \qquad (2\text{-}3\text{-}1)$$

若以矩陣表示之，則

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix};\ X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ 1 & X_{12} & \cdots & X_{p2} \\ & \vdots & & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix};\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

則

$$Y = X\beta + \epsilon \qquad (2\text{-}3\text{-}2)$$

㈣估計迴歸係數

設 $\widehat{\beta_0}$，$\widehat{\beta_1}$，…$\widehat{\beta_p}$爲迴歸係數 $\beta_0, \beta_2 \cdots \beta_p$ 之估計