國立台北大學統計學系
碩士論文

指導教授: 黃怡婷 博士

建構多變量二元聯合機率分配並應用於長期追蹤資料模型

Building the Multivariate Joint Distribution for Binary Data and its Application in Longitudinal Marginal Model

研究生：葉麗芬

中華民國 一 〇 三 年 七 月

謝 誌

謹以此檔案, 感謝台北大學統計學系在我求學期間的照顧, 黃怡婷老師及汪群超老師對我的指導, 當然還要謝謝支持我逐夢的家人朋友。希望這些從所撰論文截取出的 cwTEX 編輯技巧, 能對撰寫論文中的學弟妹們有所幫助, 也祝福你在接下來的寫作過程順利。

葉麗芬謹誌于

中華民國 一〇三 年 七 月

國立台北大學一〇二學年度第二學期碩士學位論文提要

論文題目: 建構多變量二元聯合機率分配並應用於長期追蹤資料模型

論文頁數: 85

所 組 別: 統計學系碩士班 系 (所) ____ 組 (學號: 710133121)

研 究 生: 葉麗芬 指導教授: 黃怡婷

論文提要內容:

現今許多科學研究常藉由觀察相同群體多個時間點的狀況來了解所關心事件對此群體所產生的長期平均影響, 這類型研究需要使用長期追蹤資料分析方法來瞭解影響平均反應趨勢的變數, 以供後續決策或研究參考。

在一階馬可夫鏈 (First-Order Markov Chains) 的假設下, 本論文利用 Biswas 和 Hwang (2002) 提出之二元二項分配 (Bivariate Binomial Distribution) 建構出多變量聯合二元機率分配, 並討論該分配的特性, 推導參數的最大概似估計式, 及其大樣本性質。藉由此多變量分配, 本論文提出利用最大概似估計法來估計長期追蹤資料之廣義線性模型中參數, 最後運用統計模擬來探討新的多變量二元聯合機率分配參數與長期追蹤資料之廣義線性模型參數的最大概似估計式表現, 再與現行研究者常採用的廣義估計方程式的參數估計方法進行比較。

關鍵詞: 二元長期追蹤資料、多變量二元分配、長期追蹤資料的廣義線性模型、最大概似估計法

ABSTRACT

Building the Multivariate Joint Distribution for Binary Data and its Application in Longitudinal Marginal Model

by

YE, LI-FEN

July 2014

ADVISOR: Dr. HWANG, YI-TING

DEPARTMENT: DEPARTMENT OF STATISTICS

MAJOR: STATISTICS

DEGREE: MASTER OF SCIENCE

Many recent studies often observe the response variables repeatedly to understand the influence of certain conditions longitudinally. The general linear model and generalized linear model for longitudinal data are used to make inference of this kind of data. Since the response variable is observed repeatedly, the model settings and estimations would need the multivariate distribution. Many continuous multivariate distributions have been proposed in the literatures. However, owing to the complexity of describing the association among the multivariate discrete random variables, it is lack of the well-known distribution. To estimate the parameter in the generalized linear model for longitudinal data, the generalized estimating equation (GEE) proposed by Liang and Zeger (1986) is a commonly used estimating method.

KEY WORDS: Binary Longitudinal Data, Multivariate Binomial Distribution, Marginal Model, Maximum Likelihood Estimation.

目 錄

圖目錄

表目錄

第 1 章
緒 論

現今許多科學研究藉由分析相同群體在多個時間點的狀況, 來衡量所關心之事件對此群體所產生的長期趨勢影響。例如: 流行病學家想找出某項疾病之致病基因, 醫生或藥廠想知道某藥物或療法對病人的治療效果, 證券分析師想研究某產業下之公司近年的經營狀況 ⋯⋯ 等。此類型資料因同時具有時間序列 (Time Series) 與橫斷面 (Cross Section) 訊息, 所提供的訊息較單一層面資料更為充份完整, 故不管是在生物醫藥科技、商業行為、或是公共政策研究領域, 收集此類型資料進行分析之研究與日俱增。然在流行病學、遺傳醫學等健康相關研究領域稱此為長期追蹤資料 (Longitudinal Data); 而在財務金融、計量經濟等商業相關研究領域, 則稱之為縱橫斷面資料 (Panel Data)。

分析長期追蹤資料時, 若反應變數為連續型或二元資料時, 常會繪製反應變數的平均反應曲線 (Mean Response Curve), 以便初步了解平均反應隨時間而變化的軌跡。圖 1.1 係以兩個長期追蹤資料為例繪製的平均反應曲線, 圖 1.1(a) 為研究者想了解 4 種不同飼養策略, 是否影響實驗室老鼠在實驗期間體重增加之趨勢, 故分別在實驗的一開始 (第 0 週) 及後續的第 1, 2, 4, 8, 12 週測量老鼠的體重, 屬連續型長期追蹤資料; 圖 1.1(b) 則為二元 (Binary) 長期追蹤資料, 係臨床研究者想了解某種積極治療法是否可改善患有呼吸上疾病病人之呼吸狀態, 此研究採隨機指派方式決定病人接受的治療為積極治療

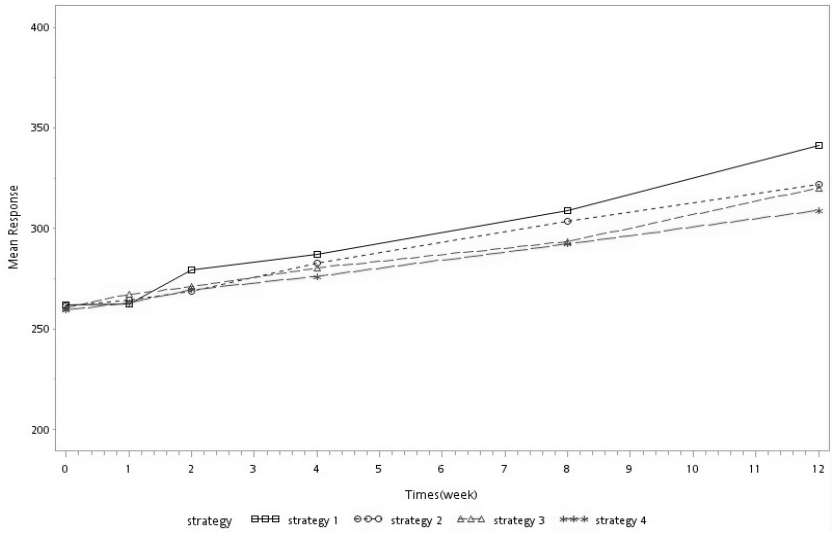
法或是安慰劑 (Placebo) 治療, 病人每週回診 1 次, 共回診 4 次, 紀錄呼吸狀態有改善者為 1, 未改善者為 0, 以計算、分析改善比率。

此外, 藉由反應變數之共變異數矩陣 (Covariance Matrix) 來了解同一研究對象各次試驗結果間的關係。令 σ_{tu} 代表時間點 t 與 u 試驗結果的共變異數, $t, u = 1, \ldots, k$, 並將第 i 個研究對象之共變異數矩陣表示如下:

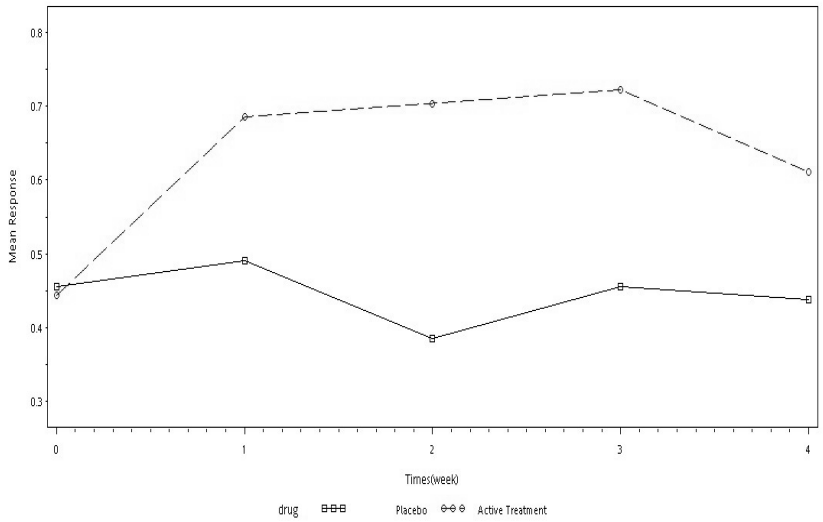
$$\text{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{kk} \end{pmatrix},$$

在分析資料時, 共變異數矩陣可以假設與研究個體 i 有關, 但一般實務上研究個數在同一時間點通常不會有重覆測量的資料, 故為簡化以下討論, 本論文假設每一研究對象均有相同共變異數結構。由於生物性, 長期追蹤資料之關係隨著間隔時間的增加而遞減, 且試驗結果間會有正相關, 故在資料分析時多會針對各試驗時間點資料間之共變異數模型 (Covariance Pattern) 進行假設。

本論文將在第 2 章說明多變量二元聯合機率分配的建構概念及此分配的特性; 第 3 章介紹二元長期追蹤資料如何以邊際模型配適平均反應模型; 第 4 章為數值模擬分析結果, 包括多變量二元分配在不同樣本和參數假設下之參數估計結果比較, 及在不同資料型態和工作矩陣假設下, 二種平均反應模型之參數估計結果比較; 最後, 在第 5 章總結本研究的結果, 並對本論文所建立的分析方法及後續研究提出建議。



(a) 連續型資料實例: 不同飼養策略對實驗室老鼠於實驗期間體重增加趨勢之影響



(b) 類別型資料實例: 不同治療方法是否改變患有呼吸上疾病病人呼吸狀態研究之改善比率曲線

圖 1.1: 長期追蹤資料平均反應曲線

NTPU

第 2 章
多變量二元聯合機率分配

本論文以指標 i 代表某研究對象, 指標 t 代表某觀察或試驗時間點; 並假設重複試驗次數為 k 次, 各時間點參與試驗之人數皆為 n。令二元隨機變數 Y_{it}, $i = 1, \ldots, n,\ t = 1, \ldots, k$, 代表第 i 個對象在時間點 t 的試驗結果, 並假設 Y_{it} 之邊際分配為參數 p_t 的白努利分配, $t = 1, \ldots, k$, 且不同研究對象間的試驗結果為獨立, 但同一研究對象各次試驗結果間可能存在關係。此外, 令 y_{it} 為其所對應的樣本值, $y_{it} = 0, 1$, 並令

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ik} \end{pmatrix},$$

為一個 $k \times 1$ 的向量, 表示第 i 個試驗對象 k 次試驗結果。於論文中若未重新進行符號定義, 在不造成閱讀困擾的前提下, 將以 Y_t 表示第 i 個對象在時間點 t 的試驗結果, 以簡化符號。本章將先說明成對資料之聯合機率函數的建構概念, 之後再將此方法一般化, 提出新的多變量二元資料的聯合機率函數。

2.1 多變量聯合機率質量函數

本論文假設 $\{Y_t, 1 \le t \le k\}$ 為一階馬可夫鏈 (First-Order Markov Chains), 亦即時間 $t+1$ 的狀態僅與時間 t 有關; 利用 Biswas 和 Hwang (2002) 提出的條件機率公式定義時間 t 與 $t+1$ 間診斷試驗結果的條件機率, 提出隨機變數 $Y_1, \ldots, Y_k$ 之 JPMF 如下:

$$\begin{aligned} f(y_1, \cdots, y_k) &= f(y_1) f(y_2|y_1) f(y_3|y_1, y_2) \cdots f(y_k|y_1, \ldots y_{k-1}) \\ &= f(y_1) f(y_2|y_1) f(y_3|y_2) \cdots f(y_k|y_{k-1}) \\ &= p_1^{y_1}(1-p_1)^{1-y_1} (\mathrm{P}^{12}_{y_1,1})^{y_2} (1-\mathrm{P}^{12}_{y_1,1})^{1-y_2} (\mathrm{P}^{23}_{y_2,1})^{y_3} (1-\mathrm{P}^{23}_{y_2,1})^{1-y_3} \\ &\quad \times \cdots \times (\mathrm{P}^{k-1,k}_{y_{k-1},1})^{y_k} (1-\mathrm{P}^{k-1,k}_{y_{k-1},1})^{1-y_k} \\ &= p_1^{y_1}(1-p_1)^{1-y_1} \prod_{t=1}^{k-1} (\mathrm{P}^{t,t+1}_{y_t,1})^{y_{t+1}} (1-\mathrm{P}^{t,t+1}_{y_t,1})^{1-y_{t+1}} \end{aligned} \tag{2.1}$$

在此函數假設下，任意兩時間點 $Y_t = 1$ 與 $Y_u = 1$ 的聯合機率跟時間點 t 的變異數有關, 關係式如引理 2.1.1 ，此證明請參考附錄 A; 而由引理 2.1.1，可證明出任意兩時間點診斷試驗結果的共變異數如定理 2.1.2。

引理 **2.1.1.** 假設 σ_t^2 為診斷試驗結果 Y_t 的變異數, 則任意兩時間點 Y_t 與 Y_u 事件皆發生的機率為

$$\Pr(Y_t = 1, Y_u = 1) = p_t p_u + \sigma_t^2 \prod_{m=t}^{u-1} \frac{\alpha_{m,m+1}}{1+\alpha_{m,m+1}}, \quad \forall\, 1 \le t < u \le k$$

定理 **2.1.2.** 假設 σ_t^2 為診斷試驗結果 Y_t 的變異數, 則任意兩時間點 Y_t 與 Y_u 的共變異數為

$$\mathrm{Cov}(Y_t, Y_u) = \sigma_t^2 \prod_{m=t}^{u-1} \frac{\alpha_{m,m+1}}{1+\alpha_{m,m+1}}, \quad \forall\, 1 \le t < u \le k$$

Proof. 由共變異數之定義可得

$$\begin{aligned}
\mathrm{Cov}(Y_t, Y_u) &= \mathrm{E}(Y_t Y_u) - \mathrm{E}(Y_t)\mathrm{E}(Y_u) \\
&= \sum_{y_t} \sum_{y_u} y_t y_u f(y_t, y_u) - p_t p_u \\
&= \Pr(Y_t = 1, Y_u = 1) - p_t p_u \\
&= p_t p_u + \sigma_t^2 \prod_{m=t}^{u-1} \frac{\alpha_{m,m+1}}{1 + \alpha_{m,m+1}} - p_t p_u \\
&= \sigma_t^2 \prod_{m=t}^{u-1} \frac{\alpha_{m,m+1}}{1 + \alpha_{m,m+1}}
\end{aligned}$$

□

NTPU

第 3 章
二元長期追蹤資料之平均反應模型

3.1 邊際模型與 MLE

在已知隨機變數分配的情況下, 邊際模型將可藉由最大概似估計法來估計參數。假設二元隨機變數服從第 2 章所提的多變量二元分配 (2.1), 故在 GLM 之邏吉斯模型架構下, 已知時間點 t 試驗結果, 時間點 $t+1$ 之平均反應模型為

$$\text{logit}\,[\text{E}(Y_{i,t+1}=1\,|\,Y_{it}=y_{it})]=\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}+\gamma_t\,y_{it},\ \ t=1,\cdots,k-1, \qquad (3.1)$$

其中, γ_t 係配合聯合機率密度函數中的一階馬可夫鏈假設, 用以解釋時間 t 與 $t+1$ 間關係之參數。此時, 條件機率可表示如下:

$$f(y_{i,t+1}\,|\,y_{it}\,;\boldsymbol{x}_{i,t+1})=\frac{\exp[y_{i,t+1}(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}+\gamma_t\,y_{it})]}{1+\exp(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}+\gamma_t\,y_{it})},\ \ y_{it},\ y_{i,t+1}=0,1, \qquad (3.2)$$

而第 i 個試驗對象 k 次試驗結果的聯合機率模型可表示如下:

$$\begin{aligned}
&f(y_{i1},\ldots,y_{ik}\,;\boldsymbol{x})\\
&\quad=f(y_{i1};\boldsymbol{x}_{i1})f(y_{i2}\,|\,y_{i1}\,;\boldsymbol{x}_{i2})f(y_{i3}\,|\,y_{i2}\,;\boldsymbol{x}_{i3})\cdots f(y_{ik}\,|\,y_{i,k-1}\,;\boldsymbol{x}_{ik})\\
&\quad=f(y_{i1};\boldsymbol{x}_{i1})\prod_{t=1}^{k-1}f(y_{i,t+1}\,|\,y_{it}\,;\boldsymbol{x}_{i,t+1})\\
&\quad=\frac{\exp[y_1(\boldsymbol{x}'_{i1}\boldsymbol{\beta})]}{1+\exp(\boldsymbol{x}'_{i1}\boldsymbol{\beta})}\prod_{t=1}^{k-1}\frac{\exp[y_{i,t+1}(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}+\gamma_t\,y_{it})]}{1+\exp(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}+\gamma_t\,y_{it})},\ \forall t=1,\cdots,k-1\ \circ
\end{aligned}$$

以 $\boldsymbol{\theta}_l$ 代表概似函數中所有參數, $\boldsymbol{\theta}_l = (\beta_1, \cdots, \beta_p, \gamma_1, \cdots, \gamma_{k-1})'$, 則此隨機樣本之對數概似函數為

$$\begin{aligned}
\ell &= \log L(\boldsymbol{\theta}_l \,|\, \boldsymbol{y}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}_l \,|\, \boldsymbol{y}_i) \\
&= \sum_{i=1}^{n} \log \left[f(y_{i1}; \boldsymbol{x}_{i1}) \prod_{t=1}^{k-1} f(y_{i,t+1} \,|\, y_{it}\,; \boldsymbol{x}_{i,t+1}) \right] \\
&= \sum_{i=1}^{n} \left\{ \log f(y_{i1}; \boldsymbol{x}_{i1}) + \sum_{t=1}^{k-1} \log f(y_{i,t+1} \,|\, y_{it}\,; \boldsymbol{x}_{i,t+1}) \right\} \\
&= \sum_{i=1}^{n} \left\{ \log \left(\frac{\exp[y_{i1}(\boldsymbol{x}'_{i1}\boldsymbol{\beta})]}{1+\exp(\boldsymbol{x}'_{i1}\boldsymbol{\beta})} \right) + \sum_{t=1}^{k-1} \log \left(\frac{\exp[y_{i,t+1}(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta} + \gamma_t\, y_{it})]}{1+\exp(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta} + \gamma_t\, y_{it})} \right) \right\},
\end{aligned} \tag{3.3}$$

最後, 藉由最大概似估計法估計模型中參數, 及費雪訊息矩陣之反矩陣作為參數的共變異數矩陣, 並分別以 $\hat{\boldsymbol{\theta}}_l$ 及 $\mathrm{Cov}(\hat{\boldsymbol{\theta}}_l)$ 表示。其數學推導過程同前面所述, 在此僅列示對數概似函數中各參數一階偏微分結果如下:

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \left[\frac{y_{i1}\, x_{i1j} + x_{i1j}\,(y_{i1}-1)\exp(\boldsymbol{x}'_{i1}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'_{i1}\boldsymbol{\beta})} \right. \\
&\qquad \left. + \sum_{t=1}^{k-1} \frac{y_{i,t+1}\, x_{itj} + x_{itj}\,(y_{i,t+1}-1)\,\exp(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta} + \gamma_t\, y_{it})}{1+\exp(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta} + \gamma_t\, y_{it})} \right], \quad j = 1, \ldots, p, \\
\frac{\partial \ell}{\partial \gamma_s} &= \frac{\partial}{\partial \gamma_s} \left\{ \sum_{i=1}^{n} \log \frac{\exp[\, y_{i,s+1}(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta} + \gamma_s\, y_{is})\,]}{1+\exp(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta} + \gamma_s\, y_{is})} \right\} \\
&= \sum_{i=1}^{n} \frac{y_{is}\, y_{i,s+1} + y_{is}\,(y_{i,s+1}-1)\,\exp(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta} + \gamma_s\, y_{is})}{1+\exp(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta} + \gamma_s\, y_{is})}, \quad s = 1, \ldots, k-1,
\end{aligned}$$

二階偏微分結果如下:

$$\frac{\partial}{\partial\beta_l}\frac{\partial}{\partial\beta_j}\ell = -\sum_{i=1}^{n}\left\{\frac{x_{i1l}\,x_{i1j}\exp(\boldsymbol{x}'_{i1}\boldsymbol{\beta})}{[1+\exp(\boldsymbol{x}'_{i1}\boldsymbol{\beta})]^2} + \sum_{t=1}^{k-1}\frac{x_{itl}\,x_{itj}\exp(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}+\gamma_t\,y_{it})}{[1+\exp(\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}+\gamma_t\,y_{it})]^2}\right\},$$

$$l, j = 1, \ldots, p,$$

$$\frac{\partial}{\partial\beta_j}\frac{\partial}{\partial\gamma_s}\ell = -\sum_{i=1}^{n}\frac{x_{i,s+1,j}\,y_{is}\exp(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta}+\gamma_s\,y_{is})}{[1+\exp(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta}+\gamma_s\,y_{is})]^2},\ j = 1, \ldots, p; s = 1, \ldots, k-1,$$

$$\frac{\partial^2\ell}{\partial\gamma_s^2} = -\sum_{i=1}^{n}\frac{y_{is}^2\exp(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta}+\gamma_s\,y_{is})}{[1+\exp(\boldsymbol{x}'_{i,s+1}\boldsymbol{\beta}+\gamma_s\,y_{is})]^2},\quad s = 1, \ldots, k-1\text{。}$$

在不失一般性假設下, MLE 具有一致性 (Consistency)。此外, 當 Cov($\hat{\boldsymbol{\theta}}_l$) 係藉由費雪訊息矩陣的反矩陣估計時, 此共變異數估計值為估計誤差的下界 (即 Cramér-Rao Lower Bound), 故根據 MLE 近似有效性 (Asymptotic Efficiency) 性質 (Casella & Berger, 1990), 本論文在多變量二元分配假設下所建構之邊際模型, 其 MLE 在大樣本假設下近似多變量常態分配, 亦即

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_l) \to \text{MVN}[\mathbf{0}, \text{Cov}(\hat{\boldsymbol{\theta}}_l)]\ \text{。}$$

NTPU

第 4 章
模擬分析

本章將藉由不同的參數假設, 模擬各種情境下之二元長期追蹤資料, 以分析第 2 章所提的多變量二元分配特性及其參數估計式表現; 及比較第 3 章所提的兩種平均反應模型之參數估計表現。此部份的數值分析係在各種參數假設下模擬 10,000 次, 並比較以下四項指標:

1. 平均值 (Sampling Mean of Estimator): 參數估計值之平均值,
2. 估計式標準誤 (Sampling Standard Error of Estimator, 簡稱 SSE): 參數估計值之標準差,
3. 平均標準誤 (Sampling Mean of Standard Error of Estimator, 簡稱 SEE): 參數標準差估計值之平均值,
4. 覆蓋率 (Coverage Probability, 簡稱 CP): 10,000 組隨機樣本參數估計結果之 95% 信賴區間包含真實參數比率。

以下逐一說明模擬參數假設及其分析結果。

4.1 多變量二元分配之模擬

本節假設重覆試驗次數增加為 5 次, 以模擬各種長期追蹤資料試驗機率間可能存在的變化趨勢及試驗間關係, 分析多變量二元分配之參數估計表現。在

重覆試驗次數 $k = 5$ 時，多變量聯合機率質量函數可表示為

$$f(y_1, \ldots, y_5) = p_1^{y_1}(1-p_1)^{1-y_1} \prod_{t=1}^{4} (\tilde{\mathrm{P}}_{y_t,1}^{t,t+1})^{y_{t+1}} (1 - \tilde{\mathrm{P}}_{y_t,1}^{t,t+1})^{1-y_{t+1}},$$

函數中參數包括各次試驗事件發生機率 $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5)$, 及試驗間關係 $\boldsymbol{\rho} = (\rho_{12}, \rho_{23}, \rho_{34}, \rho_{45})$。

在說明模擬結果前, 先在 $\boldsymbol{\rho}$ 與 $\boldsymbol{p}$ 是否相同兩大架構下, 隨機產生幾組不同參數假設下的隨機樣本 ($n = 100$), 以便了解模擬分析資料之特性, 但由於多變量資料無法以圖型呈現, 故改為統計樣本資料所有可能試驗結果組合之次數分配。亦即在重覆試驗次數 $k = 5$ 的假設下, 樣本值之所有可能組合共有 $2^5 = 32$ 種, 在此記錄樣本資料中每一種可能試驗結果之發生人數, 並比較不同參數假設下次數分配的差異。表 4.1 為次數統計結果, 表中參數設定如下:

1. 相同 $\boldsymbol{\rho}$, 不同 $\boldsymbol{p}$: 假設試驗間之相關係數相同, $\boldsymbol{\rho} = (0.5, 0.5, 0.5, 0.5)$, 觀察不同機率趨勢的影響, 其中五組 $\boldsymbol{p}$ 值設定如下:

   A : 相同, $\boldsymbol{p} = (0.3, 0.3, 0.3, 0.3, 0.3)$,

   B : 遞增, $\boldsymbol{p} = (0.7, 0.75, 0.8, 0.85, 0.9)$,

   C : 遞減, $\boldsymbol{p} = (0.8, 0.75, 0.7, 0.65, 0.6)$,

   D : 先增後減 $\boldsymbol{p} = (0.4, 0.45, 0.5, 0.45, 0.4)$,

   E : 先減後增, $\boldsymbol{p} = (0.6, 0.55, 0.5, 0.55, 0.6)$。

2. 相同 $\boldsymbol{p}$, 不同 $\boldsymbol{\rho}$: 假設事件發生機率相同, $\boldsymbol{p} = (0.3, 0.3, 0.3, 0.3, 0.3)$, 觀察各種試驗間相關係數大小的差異, 其中四組 $\boldsymbol{\rho}$ 值設定如下:

   F : 低, $\boldsymbol{\rho} = (0.4, 0.3, 0.2, 0.1)$,

   G : 中, $\boldsymbol{\rho} = (0.65, 0.6, 0.55, 0.5)$,

   H : 高, $\boldsymbol{\rho} = (0.85, 0.8, 0.75, 0.7)$,

I : 相同, $\boldsymbol{\rho} = (0.75, 0.75, 0.75, 0.75)$。

以結果 1 為例, 5 個試驗時間點所關心之事件皆未發生的人數, 在 100 個樣本資料中, 假設試驗間之相關係數相同, 機率變化趨勢假設為 A 時有 36 人, B 時有 2 人, C 時有 2 人, D 時有 22 人, E 時有 8 人; 假設機率相同, 相關係數大小假設為 F 時有 22 人, G 時為 39 人, H 時為 43 人, I 時為 52 人; 其餘情況依此類推。

表 4.1: 重覆試驗次數 $k = 5$, 假設 $\boldsymbol{\rho} = (0.5, 0.5, 0.5, 0.5)$ 在不同試驗機率趨勢下, 及 $\boldsymbol{p} = (0.3, 0.3, 0.3, 0.3, 0.3)$ 在不同試驗間相關係數大小下, 樣本資料於各種試驗結果可能值組合之人數統計 (樣本數為 100)。

結果	試驗結果可能值					次數統計								
						相同 $\boldsymbol{\rho}$, 不同 $\boldsymbol{p}^*$					相同 $\boldsymbol{p}$, 不同 $\boldsymbol{\rho}^*$			
	y_1	y_2	y_3	y_4	y_5	A	B	C	D	E	F	G	H	I
1	0	0	0	0	0	36	2	2	22	8	22	39	43	52
2	0	0	0	0	1	11	2	2	5	5	9	10	5	5
3	0	0	0	1	0	2	0	1	4	1	6	4	2	0
4	0	0	0	1	1	4	6	3	5	7	4	5	4	3
5	0	0	1	0	0	0	0	0	2	0	3	1	1	2
6	0	0	1	0	1	0	1	0	0	0	3	1	0	0
7	0	0	1	1	0	1	0	1	6	2	2	1	2	2
8	0	0	1	1	1	2	6	4	5	6	2	4	2	7
9	0	1	0	0	0	2	0	1	2	1	3	1	0	0
10	0	1	0	0	1	0	0	0	1	0	3	0	0	0
11	0	1	0	1	0	0	0	0	0	0	3	0	0	0
12	0	1	0	1	1	1	0	0	2	1	0	0	0	0
13	0	1	1	0	0	1	1	0	3	2	3	2	0	0
14	0	1	1	0	1	0	0	0	1	0	1	1	0	0
15	0	1	1	1	0	1	0	1	1	0	4	1	0	0
16	0	1	1	1	1	6	16	3	3	6	2	2	1	4
17	1	0	0	0	0	9	1	3	3	10	7	6	3	5
18	1	0	0	0	1	0	1	1	1	4	0	2	1	0
19	1	0	0	1	0	0	0	0	0	1	2	1	0	0
20	1	0	0	1	1	0	1	1	0	4	1	0	1	1
21	1	0	1	0	0	1	0	1	0	1	2	0	0	0
22	1	0	1	0	1	0	1	0	0	0	0	0	0	0
23	1	0	1	1	0	3	0	1	1	0	0	0	0	0
24	1	0	1	1	1	0	2	2	0	3	1	0	0	0
25	1	1	0	0	0	6	1	8	4	3	10	2	5	3
26	1	1	0	0	1	2	3	3	2	2	1	0	0	1
27	1	1	0	1	0	0	1	1	0	2	2	0	0	0
28	1	1	0	1	1	0	2	6	0	4	1	0	1	0
29	1	1	1	0	0	3	1	5	3	4	1	7	7	1
30	1	1	1	0	1	1	1	1	1	2	0	1	0	0
31	1	1	1	1	0	3	4	9	8	4	1	5	7	4
32	1	1	1	1	1	5	47	40	15	17	1	4	15	10

* 詳細設定請參考本文第 14 頁。

4.2 平均反應模型之模擬

本節將藉由模擬不同型態的樣本資料, 比較第 3 章所提的兩種平均反應模型參數估計方式, 在不同的共變異數結構及樣本大小假設下之表現。在此假設重覆試驗次數 $k = 5$, 模擬產生來自以下兩種模型的資料:

1. 情境一: 本論文所提之平均反應模型如 (3.1), 可表示為

$$\text{logit}\left[\text{E}(Y_{i,t+1} = 1 \mid \boldsymbol{x}_{i,t+1}, Y_{it} = y_{it})\right] = \boldsymbol{x}'_{i,t+1}\boldsymbol{\beta} + \gamma_t\, y_{it}, \quad t = 1, \cdots, 4;$$

2. 情境二: 一般平均反應模型

$$\text{logit}\left[\text{E}(Y_{i,t+1} = 1 \mid \boldsymbol{x}_{i,t+1})\right] = \boldsymbol{x}'_{i,t+1}\boldsymbol{\beta}, \quad t = 1, \cdots, 4;$$

模型中解釋變數之設定包括連續型態的量測時間 (T) 及二元型態的試驗組別 (G); 其中, 量測時間係以 0 代表第一次量測, 可能值為 0-4, 並假設解釋變數間存在交互作用 (Interaction), 故解釋變數之線性模型為

$$\boldsymbol{x}'_{i,t+1}\boldsymbol{\beta} = \beta_1 + \beta_2\,\text{G}_i + \beta_3\,\text{T}_{i,t+1} + \beta_4\,\text{G}_i \times \text{T}_{i,t+1}, \qquad i = 1, \cdots, n, \ t = 1, \cdots, 4;$$

其中, β_1 為截距項 (Intercept), 參數 $\boldsymbol{\beta}$ 的模擬設定為 $(\beta_1, \beta_2, \beta_3, \beta_4)$ = (-1.23, 0.14, 0.5, 1.2)。

本論文建構在多變量二元分配假設下之邊際模型係藉由參數 $\boldsymbol{\gamma}$ 來衡量同一研究對象各次試驗結果間可能存在之關聯, 故模型中參數除了 $\boldsymbol{\beta}$ 外, 還有 $\boldsymbol{\gamma}$ 的部分; 兩者皆以最大概似估計法進行參數估計, 並利用 3.1 節所列示之最大概似函數二階微分結果計算觀測的費雪訊息矩陣之反矩陣, 作為參數估計值的漸近共變異數矩陣, 並以 Σ_{FI} 表示。而以 GEE 進行參數估計之邊際模型則是藉由不同的工作矩陣假設來將上述個體內關聯納入考量, 本論文以 $\boldsymbol{\lambda}$ 表示, 矩陣中參數個數與共變異數結構假設有關; 此節模擬以

對數勝算比方式來設定工作矩陣, 矩陣型態包括試驗時間點之間的共變異數相同、Toeplitz 及無結構三種, 並分別以 Σ_{GE}、Σ_{GT} 及 Σ_{GF} 表示, 以 GEE 進行參數估計之模擬, 係藉由 SAS 中的 GENMOD 程序, 而 MLE 則藉由 MATLAB 的 FMINSEARCH 函數 (Lagarias, Reeds, Wright and Wright, 1998) 計算 MLE。

首先, 以 $\boldsymbol{\gamma} = (0.3, 0.25, 0.2, 0.15)$, $n = 250$ 為例, 比較各種方法之估計結果, 表 4.2 為此情境下各種方法解釋變數參數模擬結果。其中, 平均值為 10,000 組隨機樣本參數估計值之平均值; SSE 為 10,000 組隨機樣本參數估計值之標準差; SEE 為 10,000 組隨機樣本參數標準差之平均值; CP 為 10,000 組參數估計結果之 95% 信賴區間包含真實參數比率。比較表 4.2 中之 SSE 及 SEE 可知, 四種參數估計方式的 SSE 差異不大, 彼此差距在小數位第 3 位; 而 Σ_{FI} 的 SEE 是四種方法中最小的, 但 SEE 與 SSE 差距為最大, 故其 CP 表現較差。

表 4.2: 假設 $\boldsymbol{\gamma} = (0.3, 0.25, 0.2, 0.15)$, $n = 250$ 時, 各種方法之平均反應模型解釋變數參數估計結果比較。

參數	真實值	估計指標	MLE	GEE		
			Σ_{FI}	Σ_{GE}	Σ_{GT}	Σ_{GF}
$\hat{\beta}_1$	-1.23	平均值	-1.2376	-1.2041	-1.2049	-1.2052
(截距項)		SSE	0.1667	0.1631	0.1633	0.1643
		SEE	0.0711	0.1610	0.1607	0.1599
		CP	0.6058	0.9414	0.9412	0.9382
$\hat{\beta}_2$	0.14	平均值	0.1315	0.1058	0.1067	0.1070
(G)		SSE	0.2451	0.2484	0.2486	0.2493
		SEE	0.1272	0.2469	0.2467	0.2460
		CP	0.6911	0.9487	0.9474	0.9460
$\hat{\beta}_3$	0.50	平均值	0.5056	0.5213	0.5215	0.5216
(T)		SSE	0.0790	0.0654	0.0655	0.0659
		SEE	0.0330	0.0644	0.0643	0.0639
		CP	0.5972	0.9390	0.9373	0.9344
$\hat{\beta}_4$	1.20	平均值	1.2249	1.2671	1.2669	1.2667
(T $\times$ G)		SSE	0.1799	0.1760	0.1761	0.1764
		SEE	0.1093	0.1706	0.1705	0.1702
		CP	0.7720	0.9386	0.9397	0.9382

NTPU

第 5 章

結論與建議

本論文在第 2 章完整介紹多變量二元聯合機率分配及其特性, 第 3 章則是多變量二元分配假設下之長期追蹤資料模型的部分, 同時推導參數之最大概似函數估計式, 及概似函數二階偏微分結果以計算費雪訊息矩陣之反矩陣作為參數標準差; 由於上述估計式沒有封閉解, 故本研究以 MATLAB 進行大量統計模擬, 探討多變量二元分配及長期追蹤資料邊際模型之參數估計式在各種可能資料型態上的表現, 確認參數估計結果之大樣本性質。由 4.1 的模擬結果可知, 多變量二元分配不管是在何種假設條件下, 其參數估計值及標準差的估計結果在樣本數 $n = 100$ 後極為穩定, 偏誤趨近於 0, 且在多數模擬情境之 SEE 與 SSE 差異不大, $\boldsymbol{p}$ 及 $\boldsymbol{\rho}$ 的 CP 落在 90% 至 95% 間; 但在某些兩次試驗結果為高度相關, 或相關係數假設接近其上限之模擬中, 僅有 $\boldsymbol{\rho}$ 的 CP 可接近 95% 的期望水準, 但 $\boldsymbol{p}$ 之 SEE 會隨著樣本數增加而遞減速度較快, SEE 明顯低於 SSE, 致 CP 表現極差。

目前文獻上有許多連續型態的多變量聯合分配, 卻較少有類別型態的多變量聯合分配, 使得多變量類別型資料在統計分析上受到些限制。本論文藉由 Biswas 和 Hwang (2002) [2] 所提之二元二項分配中的條件機率公式, 建構多變量二元聯合機率分配並將之運用在長期追蹤資料邊際模型中; 在未來, 若能有其他更適合用以衡量類別型資料間關聯的方法, 則本論文所提之多變量二元分配建構方式應可延伸至更多的類別型式資料適用分配。

NTPU

參考文獻

[1] A. Agresti. *Categorical data analysis*. John Wiley & Sons, 2002.

[2] A. Biswas and J. S. Hwang. A new bivariate binomial distribution. *Statistics & probability letters*, 60(2):231–240, 2002.

[3] G. Casella and R. L. Berger. *Statistical inference*. Duxbury Press Belmont, CA, 1990.

[4] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied longitudinal analysis*. John Wiley & Sons, 2012.

[5] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147, 1998.

[6] K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[7] M. Pinsky and S. Karlin. *An introduction to stochastic modeling*. Academic Press, 2010.

[8] SAS Institute Inc. The SAS system, version 9.3. *Cary, North Carolina*, 2014. URL http://www.sas.com.

[9] The MathWorks Inc. MATLAB – the language of technical computing, version 7.1. *The MathWorks, Inc., Natick, Massachuseets*, 2014. URL http://www.mathworks.com/products/matlab/.

附 錄
多變量二元分配任意兩時間點 $Y_t = 1$ 與 $Y_u = 1$ 之聯合機率公式證明

引理 2.1.1 與證明

假設 $Y_t = 1$ 與 $Y_{t+1} = 1$ 的條件機率為

$$\mathrm{P}_{11}^{t,t+1} = \Pr(Y_{t+1} = 1 | Y_t = 1) = p_{t+1} + \frac{\alpha_{t,t+1}}{1 + \alpha_{t,t+1}}(1 - p_t)$$

使用數學歸納法 (Mathematical Induction) 證明任意兩時間點 t 與 u 的聯合機率, 皆可寫成下式:

$$\Pr(Y_t = 1, Y_u = 1) = p_t p_u + \sigma_t^2 \prod_{m=t}^{u-1} \frac{\alpha_{m,m+1}}{1 + \alpha_{m,m+1}}, \quad \forall\, 1 \le t < u \le k.$$

Proof. 當 $u = t + 1$ 時,

$$\Pr(Y_{t+1} = 1 \,|\, Y_t = 1) = p_{t+1} + \frac{\alpha_{t,t+1}}{1 + \alpha_{t,t+1}}(1 - p_t)$$

故

$$\begin{aligned}
&\Pr(Y_t = 1, Y_{t+1} = 1) \\
&\quad = \Pr(Y_{t+1} = 1 \,|\, Y_t = 1)\, \Pr(Y_t = 1) \\
&\quad = \left[p_{t+1} + \frac{\alpha_{t,t+1}}{1 + \alpha_{t,t+1}}(1 - p_t) \right] p_t \\
&\quad = p_t\, p_{t+1} + \sigma_t^2 \frac{\alpha_{t,t+1}}{1 + \alpha_{t,t+1}}
\end{aligned}$$

(略) □

著作權聲明

論文題目：建構多變量二元聯合機率分配並應用於長期追蹤資料模型

論文頁數：85 頁

系所組別：統計學系

研究生：葉麗芬

指導教授：黃怡婷

畢業年月：一〇三年七月