「發展鄉鎮逐時天氣預報」計畫之高解析度統計預報技術研究(2)
-引用順序型邏輯斯迴歸模型進行雲量多類別預報試驗

王政忠　　陳雲蘭
中央氣象局氣象預報中心

摘　要

傳統對於雲量多類別預報採取線性迴歸建模，從各類別迴歸方程計算得到其機率預報後，進一步決定各類別的機率門檻值，最後再根據機率門檻值決定類別預報。而雲量類別實屬於具有順序性的反應變數，傳統做法在建模忽略了其順序特性，且以線性迴歸建立機率預報，在王與陳(2010)研究中已指出其不適用性。爲了改善傳統雲量多類別預報的缺點，本研究將引進在統計上處理順序型反應變數的順序型邏輯斯迴歸模型，以實際資料建立雲量多類別預報試驗，並與傳統雲量建模方式比較其預報成效。

關鍵字：雲量多類別預報、順序型邏輯斯迴歸、比率勝算模型

一、前言

對於迴歸預報模型建立，以統計角度來看，線性迴歸模型(Linear Regression)一般是用來處理連續型的反應變數，如溫度、相對溼度等；當面對類別型之二元化變數(binary variable)，如降雨事件與否之機率預報，在王與陳(2010)的研究中發現無論在理論上與實際資料的驗證，邏輯斯迴歸(Logistic Regression)爲較合理的預報模型。另一方面，變數選取方法在建模時亦扮演著重要角色，在過去的研究發現建模時搭配連續型變數挑選方法之最小絕對壓縮挑選機制(LASSO)在溫度、降雨機率預報試驗，其預報表現均比傳統變數選取方法-前序選取(Forward Selection)佳。爲了延續過去研究，本研究設定雲量類別爲主要預報目標，除了將兩分類預報推廣到多類別預報，引進多類別預報之統計模型，並進一步探究搭配最小絕對壓縮挑選機制(LASSO)對於預報表現的改善。

傳統對於雲量多類別預報，以美國氣象作業單位爲例(Weiss, 2001)，先將雲量觀測數值分成四類(Clear、Scattered、Broken、Overcast)類別型變數，按照各類別將反應變數轉成二元化變數，以線性機率模型(Linear Probability Model; LPM)搭配前序選取(Forward Selection)同步建立(develop simultaneously)四組迴歸方程(詳見第二章)，由四組迴歸方程得到各類別的機率預報後，再根據機率門檻值決定預報類別。

然而雲量類別變數屬於順序型變數(ordinal variable)，傳統建模流程一開始將其視爲名義型變數(nominal variable)，於建模時又將其視爲連續型變數，最後再依據機率門檻值將其調整成順序型變數。此建模方式，忽略其反應變數的順序性，經調整後才成爲順序類別，似乎並不合理；其次，以線性迴歸建立機率預報，在王與陳(2010)的研究中已指出其不適用性，且由 4 組迴歸方程得到各類別機率，必須經過事後正規化調整，其總和才會爲 1；最後爲了決定順序類別，還必須先找出機率門檻值，才可進行分類。整個建模流程可見存在許多不合理處，且常須經過事後調整才可使用。若想在建模時保持原始資料的順序性，讓建模流程趨於合理性，在統計上是否有模型可以處理順序型變數，且可簡化其建模流程？

爲了簡化雲量多類別預報建模流程，讓整個建模流程具有合理完整性，本研究引進統計上對於處理順序類別變數最常見的模型，順序型邏輯斯迴歸模型中的比率勝算模型(Proportional Odds Model; POM)建模，以實際資料建立雲量多類別預報試驗，並與傳統雲量建模方式比較其預報成效。

二、方法介紹

以下介紹傳統對於雲量類別預報的流程，及以建模合理性爲出發點的順序性邏輯斯迴歸模型。

2.1 傳統模型-線性機率模型

傳統對於雲量類別預報的流程可分成五個步驟：

1. **觀測資料轉換：**將雲量觀測值依準則分成四類(Clear、Scattered、Broken、Overcast)，仿照二元化變數(降水事件)做法，若預報目標類別爲Clear類，將屬於Clear類的樣本標記爲1，非Clear類的樣本標記爲0，依此類推將各類樣本轉成二元化變數。

2. **迴歸模型建模：**模型採用線性機率模型搭配前序選取(Forward Selection)挑選預報因子，在考慮影響雲量各類別預報因子必須一致下，以同步方式(simultaneously)分別建立4組迴歸方程，預報因子最多挑選18個。

3. **類別機率正規化：**從4組迴歸方程可得到各類別預報機率，其機率總和不會爲1，且會出現超出合理機率範圍，必須經正規化調整後才可使用。

4. **決定累積機率門檻值：**爲了決定預報類別，此步驟必須利用建模資料決定累積機率門檻值，決定方式有很多種，本研究先以ROC曲線決定各類別最佳機率切點，再將各類別最佳機率切點正規化後，當作判別預報類別的累積機率門檻值。

5. **決定類別預報：**從4類依序檢視，從Clear類開始，若此類機率超過其累積機率門檻值則判定預報類別爲Clear，否則將此類機率累積至下一組，若此累積機率超過其累積機率門檻值則判定預報類別爲Scattered，輸出類別預報。

2.2 新引進模型-比率勝算模型

在統計上，對於屬於順序類別的雲量預報，最常見的統計模型爲比率勝算模型，除了如傳統雲量預報流程的第 1 步驟及第 2 步驟是必須的，其餘步驟皆可經由合適的模型假設即可在建模後直接輸出，而不需任何手動調整動作。以下將針對第 2 步驟迴歸模型建模，說明比率勝算模型的由來。

假設反應變數 Y 有 J 個具有順序性的類別，爲了保留反應變數 Y 其順序性，以累積機率建構順序型邏輯斯迴歸模型。累積至第 j 組之機率即：

$$P(Y \le j \mid x) = \pi_1(x) + \cdots + \pi_j(x),\ \ j = 1, \ldots, J \qquad (1)$$

其中$\pi_j(x) = P(Y = j \mid x)$

將累積機率 $P(Y \le j \mid x)$ 其勝算比(odds)取對數後，可得到 J－1 個對數累積勝算(log cumulative odds)：

$$\log\left(\frac{P(Y \le j \mid x)}{1 - P(Y \le j \mid x)}\right) = \log\left(\frac{\pi_1(x) + \ldots + \pi_j(x)}{\pi_{j+1}(x) + \ldots + \pi_J(x)}\right),\ j = 1, \ldots, J-1 \quad (2)$$

上式每個對數累積勝算皆使用到 J 個順序性類別資訊，且可視爲由第 1 組累積至第 j 組爲一類別，而由第 j+1 組至第 J 組爲另一類別，則此模型與二元化邏輯斯迴歸概念相同，但更好的做法，可將 J－1 個對數累積勝算視爲單一精簡模型，同時利用 J－1 個結果，即：

$$\log\left(\frac{P(Y \le j \mid x)}{1 - P(Y \le j \mid x)}\right) = \alpha_j + \beta' x,\ j = 1, \ldots, J-1 \quad (3)$$

稱爲比率勝算模型，其中每個對數累積勝算有其個別的 α_j，但具有相同的迴歸係數 β。在固定 x 情況下，由於 $P(Y \le j \mid x)$ 隨著 j 而遞增，α_j 亦隨 j 而遞增。由於此模式對於每個類別具有相同的 β，不可單獨個別去估計，必須採用最大概似估計法(Maximum Likelihood Estimate)同步估計，其概似函數爲：

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(x_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(P(Y \le j \mid x_i) - P(Y \le j-1 \mid x_i)\right)^{y_{ij}}\right]$$
$$= \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(\frac{e^{\alpha_j + \beta' x_i}}{1 + e^{\alpha_j + \beta' x_i}} - \frac{e^{\alpha_{j-1} + \beta' x_i}}{1 + e^{\alpha_{j-1} + \beta' x_i}}\right)^{y_{ij}}\right] \qquad (4)$$

將上式視爲 $(\{\alpha_j\}, \beta)$ 的函數，藉由數值方法求使得概似函數爲最大的解，即爲迴歸係數估計。

由(3)式，當 x 其值分別爲 x_1 與 x_2 ，則：

$$\log\left(\frac{P(Y \le j \mid x_1)}{1 - P(Y \le j \mid x_1)}\right) - \log\left(\frac{P(Y \le j \mid x_2)}{1 - P(Y \le j \mid x_2)}\right)$$
$$= \log\frac{P(Y \le j \mid x_1)/P(Y > j \mid x_1)}{P(Y \le j \mid x_2)/P(Y > j \mid x_2)}$$
$$= \beta'(x_1 - x_2)$$
$$\Rightarrow \quad \frac{P(Y \le j \mid x_1)/P(Y > j \mid x_1)}{P(Y \le j \mid x_2)/P(Y > j \mid x_2)} = e^{\beta'(x_1 - x_2)} \qquad (5)$$

對於所有類別，當 $x = x_1$ 其勝算爲 $x = x_2$ 其勝算的 $e^{\beta'(x_1 - x_2)}$ 倍，可發現此比率關係與類別無關，因此McCullagh(1980)稱其爲比率勝算模型，各類別累積機率圖呈現互相平行的形式，見(圖 1)。

在比率勝算模型中，爲了建構類別門檻，引進潛在變數概念，即：

$Y^* = \beta' x + \varepsilon = \eta(x) + \varepsilon$ ，其中 ε 服從邏輯斯分配，其累積分布函數爲 G。雖然無法直接觀察此連續變數 Y^* ，但是依連續門檻值：

$$-\infty < \alpha_1 < \cdots < \alpha_{J-1} < \infty \qquad (6)$$

將反應變數 Y 分成 J 個類別，即：

$$Y = \begin{cases} 1 & \text{for} \quad \mathrm{Y}^* \le \alpha_1 \\ 2 & \text{for} \quad \alpha_1 < \mathrm{Y}^* \le \alpha_2 \\ \vdots & \\ J-1 & \text{for} \quad \alpha_{J-2} < \mathrm{Y}^* \le \alpha_{J-1} \\ J & \text{for} \quad \alpha_{J-1} < \mathrm{Y}^* \end{cases} \qquad (7)$$

則 Y 的累積機率分布：

$$\begin{aligned} P(Y \le j \mid x) &= P(Y^* \le \alpha_j \mid x) \\ &= P(\eta(x) + \varepsilon \le \alpha_j \mid x) \\ &= P(\varepsilon \le \alpha_j - \eta(x) \mid x) \\ &= P(\varepsilon \le \alpha_j - \beta' x \mid x) \\ &= G(\alpha_j - \beta' x) \\ &= \frac{\exp(\alpha_j - \beta' x)}{1 + \exp(\alpha_j - \beta' x)} \end{aligned} \qquad (8)$$

將上式取反函數，即可得到(3)式結果。藉由潛在變數，使得比率勝算模型輸出後，即具有其門檻值，不需再經事後處理步驟，見(圖 2)。

三、試驗設計

本研究原本試驗設計除了進行傳統雲量預報模型與比率勝算模型比較，並進一步搭配 LASSO 來挑選預報因子是否能進一步改善預報，但由於比率勝算模型搭配 LASSO 的程式尚在撰寫階段，改以直接對建模模型比較。

首先將雲量觀測值分成三類，並挑選與雲量觀測值高度相關的預報因子 15 個，固定預報因子個數不再做變數挑選，直接建立傳統雲量模型與比率勝算模型，對六個代表測站(台北、基隆、花蓮、台中、台南、台東)進行一月與二月雲量預報，並以交叉驗證方式檢視其預報表現。

四、分析結果

由於雲量多類別預報，其分類結果是根據各類別預報機率經門檻判斷而得到，所以在校驗時，我們先以 RPS(Ranked Probability Score)進行機率預報準確度校驗；其次，爲了檢視分類狀況，以分類預報準確率(Accuracy)校驗實際正確類別預報的比率；最後，採用具有懲罰性預報錯誤類別效果的 GSS(Gerrity Skill Score)準則校驗，將以台北站一月雲量類別預報爲例，逐一分析其校驗結果。

對於機率預報校驗，在配適階段(圖 3)，兩者模型表現一致，可見傳統雲量預報模型與比率勝算模型其 RPS 均落於 0 至 0.1 之間；而在預報階段(圖 4)，亦發現兩者模型表現相近，兩個模型其 RPS 均低於 0.2。

進一步對其分類預報準確率校驗，在配適階段(圖 5)，可以發現兩者準確率均超過八成，比率勝算模型其準確率略高於傳統雲量預報模型；在預報階段(圖 6)，除 2007 外，比率勝算模型其準確率亦略高於傳統雲量預報模型。在類別預報準確校驗下，採用比率勝算模型具有較高預報準確率。

最後校驗類別預報能力，以 GSS 準則校驗，GSS 依據實際類別與預報類別結果，透過得分矩陣對於命中給予加分，未命中給予減分懲罰，則 GSS 大於 0 代表具有分類預報能力技術得分，小於 0 則代表其預報技術低於隨機預報。在配適階段(圖 7)，傳統雲量預報模型其 GSS 可達八成以上，而比率勝算模型略低於八成；在預報階段(圖 8)，在 2007、2008 年可明顯看出，比率勝算模型表現較差。因此，在類別預報能力較驗下，採用比率勝算模型沒有明顯改善 GSS。

在其他測站以及不同月份雲量預報試驗中，在 RPS 準則校驗下，比率勝算模型與傳統雲量預報模型表現亦趨近一致；而以預報類別準確率校驗，比率勝算模型表現比傳統模型佳；但在 GSS 校驗下，比率勝算模型表現略差。

五、小結

對於屬於順序類別雲量預報，引進順序型邏輯斯迴歸模型-比率勝算模型，在建模時以累積機率建構反應變數的順序性，而傳統模型其建模過程，將反應變數從名義類別變數轉成連續變數後，最後才轉爲順序變數，比率勝算模型改善了傳統建模對於反應變數的不合理的調整。另一方面，採用比率勝算模型所輸出的各類別機率即可落於合理機率範圍，不需經過如傳統模型的正規化調整步驟。而在影響雲量各類別的因子必須一致的假設下，傳統建模需經過同步處理才可達成，比率勝算模在模型假設即涵蓋了此效果，詳細比較見(表 1)。

本研究所引進的比率勝算模型在初步試驗並不期待其預報表現達到最佳，但若以簡化雲量多類別預報建模流程，且讓整個建模流程趨於合理完整性爲目的下，比率勝算模型的確是符合我們期待的模型。

另外，本研究由於程式尚在撰寫階段，尚未搭配變數挑選方法去挑選預報因子，讓模型自行去選擇對於模型最有貢獻的因子，未來將繼續將比率勝算模型搭配最小絕對挑選機制進行雲量類別預報試驗，期望在模型合理性及變數選取方法的改善下，可增進雲量預報的準確穩定性。

六、參考文獻

Agresti, A, 1996: An Introduction to Categorical Data Analysis, New York: Wiley.

McCullagh, P. 1980. Regression models for ordinal data. J. Roy. Statist. Soc. Ser. B 42: 109_142.

Tibshirani, R. 1996: Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288.

Weiss, M., 2001: AVN-based MOS ceiling height and total sky cover guidance for the contiguous United States, Alaska,Hawaii, and Puerto Rico. NWS Technical

Procedures Bulletin No. 483, NOAA, U.S. Department of Commerce, 22 pp.

王政忠、陳雲蘭，2010：邏輯斯迴歸(Logistic Regression)模型輔以最小絕對壓縮挑選機制(LASSO)於降水機率預報之應用。九十九年天氣分析與預報研討會論文彙編，中央氣象局，236-241

表1：比率勝算模型與傳統模型比較表

	比率勝算模型	傳統雲量預報模型
反應變數特性	順序型變數	名義型變數→連續型變數→順序型變數
類別機率範圍	均在合理機率範圍	可能超出合理機率範圍
如何達到影響各類別變數一致，機率總合爲 1	經由模型假設	經同步(simultaneously)建模
迴歸係數特徵	常數項即潛在變數的門檻值隨類別改變；限制 β 不隨著類別而改變	β 與常數項均隨類別而改變
各類別累積機率圖	呈現平行，較具解釋性	呈現不平行
順序類別決定方式	利用潛在變數的門檻值決定順序性類別	計算累積機率，檢視是否超過機率門檻值，再決定順序性類別
門檻值	即爲迴歸係數 $\alpha_1,\cdots,\alpha_{J-1}$	$P_1,\cdots,P_{J-1}$ 須經事後決定

圖2：比率勝算模型其潛在變數示意圖(Agresti,1996)

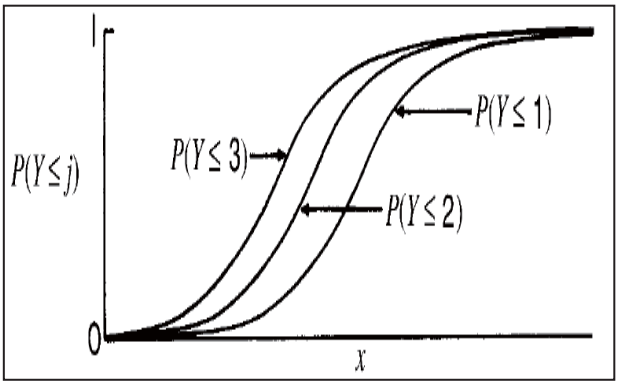
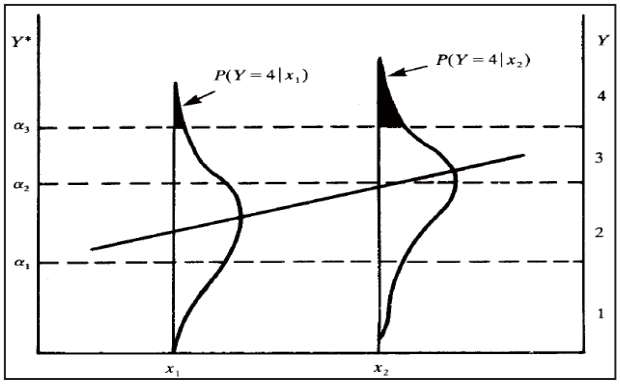


圖1：比率勝算模型其各類別累積機率圖，呈現平行(Agresti,1996)



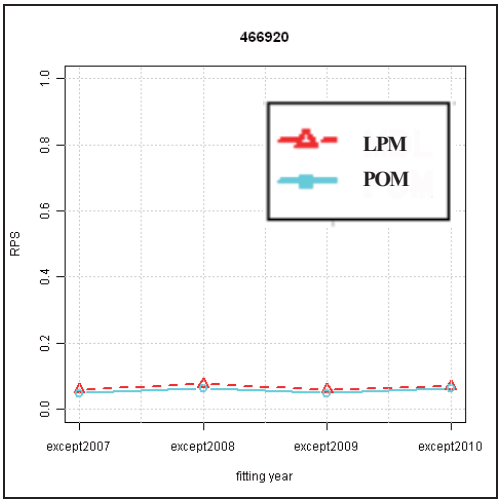


圖 3：台北站 1 月雲量類別預報交叉驗證圖，以 RPS 校驗(配適階段)(其中實線三角形爲傳統模型(LPM)；虛線圓形爲新引進模型

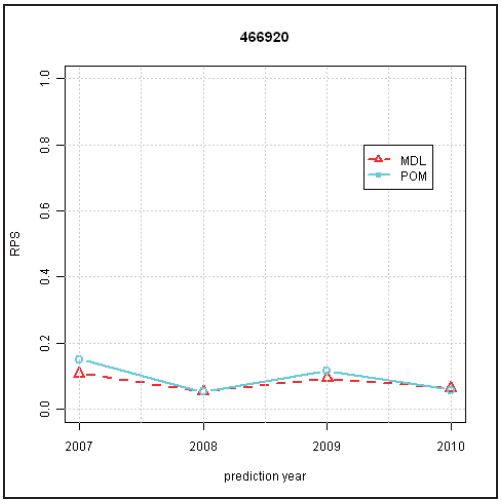


圖4：同圖3，改爲預報階段

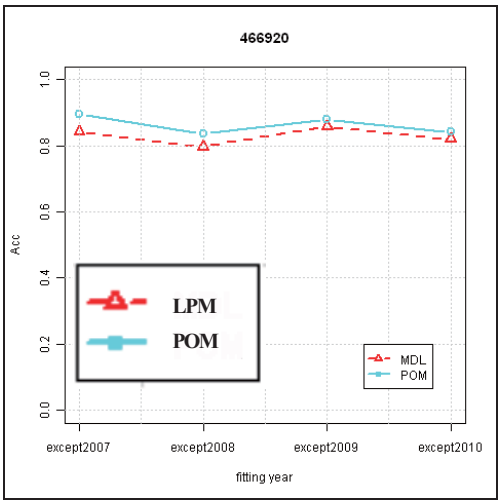


圖5：同圖3，改以分類準確率(Acc)校驗

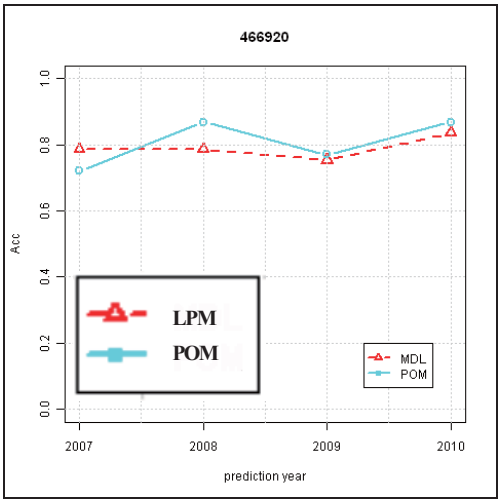


圖6：同圖5，改爲預報階段

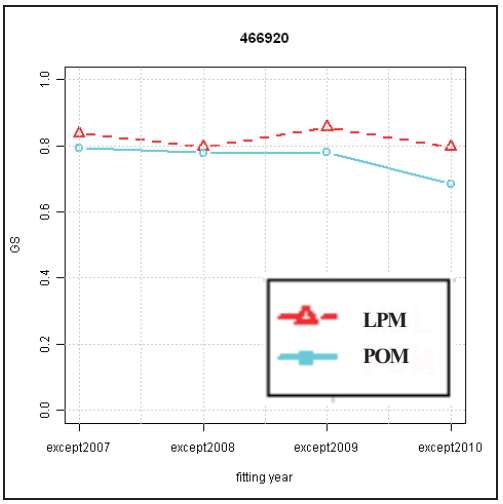


圖7：同圖3，改以GSS校驗

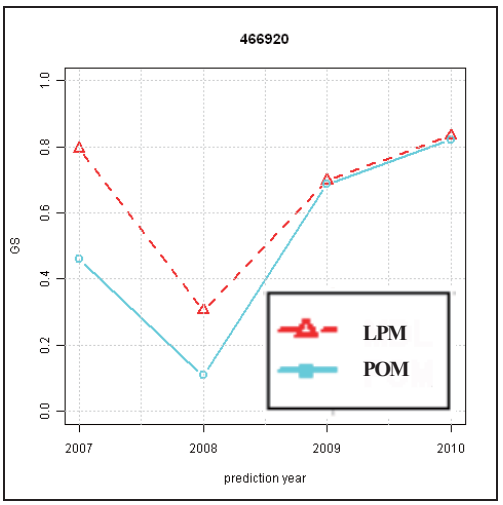


圖8：同圖7，改爲預報階段