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

王政忠　　陳雲蘭
中央氣象局氣象預報中心

摘　要

為了改善傳統降水機率預報模型中的缺點，預報機率值會有超出正常範圍的狀況及模型的不適用性，並驗證最小絕對壓縮挑選機制(LASSO)於離散型反應變數的預報表現，本研究以實際天氣資料建立邏輯斯迴歸模型輔以最小絕對壓縮挑選機制(Logistic-LASSO)應用於逐日降水機率預報，分別利用白氏得分(Brier Score)、白氏技術得分(Brier Skill Score)等準則去校驗其預報狀況，並與傳統線性機率-前序選取模式(LPM-FS)、邏輯斯迴歸-前序選取模式(Logistic-FS)與傳統線性機率-最小絕對壓縮挑選機制(LPM-LASSO)三組模式相比較。

關鍵字：變數選取、降雨機率預報、LASSO、Logistic Regression、Linear Probability Model

一、前言

在天氣資料中建立迴歸預報模式時，對於降水機率預報，傳統建模方式將降水資料轉成二元化資料(binary data)，把二元化資料視為降水機率，搭配前向選取(Forward Selection)變數選取方法，與預報因子建立線性迴歸模式，在統計上稱為線性機率模型(Linear Probability Model)。美國 MDL 現行降水機率預報即是按照此方式建立預報，預報因子取點方式主要採取將大氣層場格點直接內插至測站經、緯度位置，迴歸方程採用區域化方程預報，同一個區域內的測站由同一方程建立降水機率預報。本局過去降水機率預報，亦採取傳統建模方式建模，預報因子取點方式同美國 MDL 做法，將測站附近四個網格點內插至測站經、緯度位置，為了降低預報因子間的共線性，在前向選取(forward selection)步驟加入高共線性預報因子刪除的判斷，即被選取預報因子間之相關係數高於某一設定值時將其捨棄，預報區域為台灣本島局屬 25 個測站，單站一個迴歸方程。

不過以線性機率模式建立降水機率模型，在模型上存在著不適用性的問題，主要缺點為所預報機率值會有超出正常範圍的狀況(大於 1 或小於 0)；另一方面，估計迴歸係數採用最小平方法時，即將反應變數 Y 其變異數(variance)假設為常數，但當反應變數 Y 為二元化變數時，其變異數隨著預報因子(X)其值而改變，而且反應變數其值僅有 0 與 1 兩個值，與常態分布相違背，因此以最小平方法來估計迴歸係數並不恰當。在統計上當面對反應變數為二元化資料，資料將會被視為類別資料(categorical data)來處理，常見的建模方式為邏輯斯迴歸模型(Logistic Regression Model)，以非線性的方式建模，改善了傳統降水機率模型的不適用性，讓迴歸係數係數估計更加合理(詳見第二章)。Applequist 等人在 2002 年發表一篇各種統計方法應用於定量降水機率預報(pQPF)比較的文章(Applequist，2002)，預報因子取點方式同美國 MDL，預報區域為美國中部與東部區域的 154 個測站，採用區域化迴歸方程，同一個區域內的測站將由同一方程建立預報。文章中應用的統計模式有線性迴歸模式(即線性機率模型)(含原始資料與轉換為主成分資料兩種做法)、邏輯斯迴歸(含原始資料與轉換為主成分資料兩種做法)、判別分析、類神經網路、分類系統共五種，並採取逐步方法(stepwise method)挑選預報因子，其結論指出利用轉換後的主成分資料建立邏輯斯迴歸模型具有較高的白氏技術得分(Brier Skill Score)。

為了改善傳統降水機率預報模型中的缺點，本研究將引進邏輯斯迴歸建立降水機率預報，另外在(王與陳，2009)文章指出利用最小絕對壓縮挑選機制(LASSO)於連續的氣象要素建立預報，如逐日平均溫度預報，預報表現與運算效率明顯優於逐步方法所建立的模型。因此在建模時除了引進邏輯斯迴歸模式外，並將利用最小絕對壓縮挑選機制(LASSO)對預報因子做挑選，以實際天氣資料建立 Logistic-LASSO 降水機率預報模式，依此模式進行逐日降水機率預報，分別利用白氏得分(Brier Score)、白氏技術得分(Brier Skill Score)等準則去校驗其預報狀況，並與傳統降水機率建模方式線性機率模式-前序選取(LPM-FS)等模式比較其配適階段(Fitting Period)與校驗階段(Verifying Period)的表現。

二、方法

為了探討降水機率預報模型在主模型假設差異(線性機率模型(LPM)、邏輯斯迴歸模型(Logistic))並搭配不同變數選取方法(Forward Selection、LASSO)其預報表現比較，本研究設定四組比較模型：

- LPM-FS：傳統降雨機率預報建模方式，如美國MDL、本局。
- Logistic-FS
- LPM-LASSO
- Logistic-LASSO：本研究引進的降水機率預報建模方式。

以下將分別對介紹主模型與變數選取方法的介紹。

當反應變數(Y)為一個二元化變數時，例如天氣降雨與否、疾病診斷是否有病，其中一類是研究者有興趣的結果稱為「成功」，記為1；另一類則稱為「失敗」，記為0，每個觀測值皆是兩種發生狀況其中之一，稱為百努利試驗(Bernoulli trial)。令P(Y=1)(當反應變數Y=1的機率)為 $\pi(x)$ ，則Y的期望值(E(Y))為P(Y=1)，Y的變異數(var(Y))為 $\pi(x)(1-\pi(x))$ 。以下將介紹兩個常見處理二元化資料的迴歸模式，分別是線性機率模式(Linear Probability Model)與邏輯斯迴歸模式(Logistic Regression Model)。以下為了簡化說明，以單一預報因子(x)為例。

2.1 線性機率模型(Linear Probability Model)

對於二元化的反應變數(Y)，解釋變數為 x ，其迴歸模式若為：

$$\pi(x)=\alpha+\beta x \quad (1)$$

稱為線性機率模式，即將反應變數Y=1視為發生機率為1；Y=0視為發生機率為0，其迴歸係數估計可以藉由傳統最小平方法(OLS)計算。從模式左式可知 $\pi(x)$ 機率值在0至1之間，但模式右式為線性函數取值範圍在整個實數線，因此當解釋變數x其值過大或過小時，預報機率將會出現超過機率正常範圍的狀況，為此模式的主要缺點。另一方面，當反應變數Y的變異數並非常數，其值隨著解釋變數值 x 而改變，此時最小平方法並非最佳估計方法，而經由最大概似估計法(MLE)得到的估計比最小平方法更具有有效性(MSE更小)，再加上二元化反應變數Y與常態分配背離狀況，因此最小平方法在此並不適用。

2.2 邏輯斯迴歸模型(Logistic Regression)

對於二元化的反應變數(Y)，解釋變數為 x ，其迴歸模式若為：

$$\pi(x)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \quad (2)$$

稱為邏輯斯迴歸模式，其勝算比(odds)形式為：

$$\frac{\pi(x)}{1-\pi(x)}=\exp(\alpha+\beta x) \quad (3)$$

再將勝算比(odds)取 log 以後，其線性關係為：

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=\alpha+\beta x \quad (4)$$

儘管 $\pi(x)$ 機率值在 0 至 1 之間，而(4)式左式 $\log\left(\pi(x)\big/1-\pi(x)\right)$ 可為任意實數值，與右式線性函數取值範圍一致，因此預報機率不會產生超過機率正常範圍的狀況，改善了線性機率模型的缺點。為何命名為「邏輯斯」迴歸模式？由於邏輯斯分布(logistic distribution)，當其參數 $\mu=0,\ s=1$ 時，其累積分布函數(cdf)為：

$$F(w)=e^w\big/1+e^w$$

當假設 $\pi(x)=F(\alpha+\beta x)$ ，經整理可得到(2)式結果。另外若 $F(w)$ 為標準常態分布的累積常態分布時，可得另一個迴歸模式為：

$$\pi(x)=\Phi(\alpha+\beta x)\ (\Phi^{-1}(\pi(x))=\alpha+\beta x)$$

稱為 Probit 迴歸模式，與邏輯斯迴規模式差異僅在 $\pi(x)$ 假設的累積分布函數不同。在迴歸係數估計方面，由於反應變數 Y(離散)與 $\alpha+\beta x$ (連續)並無直接的關係，無法採用最小平方法來估計，改用最大概似估計法來估計。

$$Y_i \sim Bernoulli(\pi(x_i)),\ \pi(x_i)=F_i=F(\alpha+\beta x_i)$$

其概似函數(likelihood function)為：

$$\begin{aligned}L(\alpha,\beta\mid y_1,\cdots y_n)&=f_{Y_1,\cdots,Y_n}(y_1,\cdots y_n\mid\alpha,\beta)\\&=\prod_{i=1}^{n}f_Y(y_i\mid\alpha,\beta)\\&=\prod_{i=1}^{n}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}\\&=\prod_{i=1}^{n}F_i^{y_i}(1-F_i)^{1-y_i}\end{aligned}$$

$$\xrightarrow{\text{取log}}\ \log L(\alpha,\beta\mid y_1,\cdots y_n)=\sum_{i=1}^{n}\left[\log(1-F_i)+y_i\log(\frac{F_i}{1-F_i})\right]$$

最大概似估計概念即選擇一組參數值使得概似函數最大，表示當 $(\alpha,\beta)=(\hat{\alpha},\hat{\beta})$ 時，同時出現n個觀測值($y_1,\cdots y_n$)的可能性最大。相當於要解決下列最佳化問題：

$$\underset{\alpha,\beta}{Max}\,L(\alpha,\beta\mid y_1,\cdots y_n)$$

在不影響最佳解情況下，將概似函數取對數函數以便利計算，最佳化問題變為：

$$\underset{\alpha,\beta}{Max}\ \log L(\alpha,\beta \mid y_1,\cdots y_n)$$

把對數概似函數分別對α 和 β 微分並設定為 0 後，將會得到兩個非線性方程式，由於無法求出封閉解(closed form solution)，藉由非線性方程式數值迭代方法可求到最佳解，如 Newton-Raphson 法，即可得到邏輯斯迴歸係數估計。

2.3 變數選取方法

前序選取(Forward Selection)方法，主要參考(陳、王與張，2009)的Forward Scheme，在每次選取過程利用判定係數或AIC準則決定哪些變數可以進入模型，並加以考慮變數間的共線性狀況，每個步驟加入共線性檢視準則，以降低已挑選變數間的共線性，唯一不同點在第一步驟並無利用P值準則去刪除變數，詳細步驟流程見(陳與王，2009)。

最小絕對壓縮挑選機制(Least absolute shrinkage and selection operator)由Tibshirani在1996年所提(Tibshirani，1996)，簡稱LASSO，其方法是將傳統最小平方估計加入一個l-norm的限制式，即：

$$\min_{\beta}\left\|Y-X\beta\right\|^2 \ \ subject\ \ to\ \ \left\|\beta\right\|=\sum_{j=1}^{d}\left|\beta_j\right|\le t$$

此限制式下，除了係數會被壓縮，限制區域會出現奇異點(singular point)，當極值發生在奇異點位置，某些β 中的元素將被壓縮到零，所以 LASSO 在估計的同時，也經由壓縮係數而做了變數選取，詳見(王與陳，2009)。上式為當模式為線性迴歸的狀況，若模式為邏輯斯迴歸時，即在其概似函數加入上式的限制式$\left\|\beta\right\|=\sum_{j=1}^{d}\left|\beta_j\right|\le t$ ，藉由數值方法即可求得迴歸係數估計。

三、校驗方法

3.1 白氏得分(Brier Score；BS)

白氏得分可用於驗證機率預報的準確性，其計算式如下：

$$BS=\frac{\sum_{i=1}^{n}\left(f_{p_i}-v_i\right)^2}{n}$$

f_{p_i} : 每個事件的預報機率

v_i ：v = 1，代表有降雨事件

v = 0，代表無降雨事件

其值介於0至1之間，越小代表整體預報越準確。

3.2 白氏技術得分(Brier Skill Score；BSS)

白氏得分技術得分用於驗證本研究迴歸模式與另一參考模式(通常比較參考對象為氣候持續法預報)其白氏得分的改善程度，其計算式如下：

$$SS=\frac{BS_c-BS_f}{BS_c}$$

BS_c : 校驗參考方法的BS, 此處與氣候持續法比較

BS_f : 本研究迴歸預報的BS

其值介於$-\infty$ 至1之間，其值為正時代表具有預報技術得分。

3.3 分類狀況

利用模式配適階段的預報機率與其對應事件發生狀況(0 或 1)，根據不同的機率臨界值可以繪製 ROC 曲線(Receiver Operating Characteristic curve)，即 X 軸為 FPR(False positive rate；同 False alarm rate)相對於 Y 軸為 TPR(True positive rate；同 Hit rate)的曲線，除了檢視模式配適階段的狀況，並要決定機率臨界值以提供校驗階段作分類，其最佳機率臨界值發生點即尋找當 ROC 曲線越靠近左上方位置的座標，使得 TPR 更高而 FPR 更低，可利用 *Youden index* 決定最佳機率臨界值，即：

$$Youden\ index=\underset{c}{Max}(TPR_{(c)}-FPR_{(c)})$$

其中 c 為不同的機率臨界值。藉由配適階段所決定的最佳機率臨界值，可進一步提供校驗階段作分類，即當預報機率大於等於最佳機率臨界值時，則判為降雨事件；反之，則判為非降雨事件。並製作 2x2 列聯表，同時檢視其 TPR、FPR、準確度(Accuracy)，校驗其降雨與非降雨事件分類狀況。

四、資料與分析結果

為了檢視邏輯斯迴歸(Logistic Regression)輔以最小絕對壓縮挑選機制(LASSO)(Logistic-LASSO)對於降雨機率預報能力，建立了本局屬六個測站(台北、基隆、花蓮、台南、台中、台東)整年度降雨機率預報，大致預報表現相近，但由於本研究著重於建模方法的

比較，選定反應變數(Y)與預報因子關係較佳的基隆站為主要預報目標，預報因子取自各層場的格點，其層場包括 H(高度)、T(溫度)、U(風場-U component)、V(風場-V component)、q 等基本場(field)，及從基本場所衍生的導出量場等，共 44 個層場，而候選的預報因子集合選取方式不同於以往單層僅取一點，除了各層場內插至測站上空單一格點並加入測站附近四點，即單一層場有五個預報因子，共 220 個候選因子，樣本分季方式以月為單位，採取預報當月資料並納入當月前後 15 日，由於資料長度的限制，樣本只有 3 年至 4 年的資料。為了檢視各模式穩定性，以交叉驗證方式(任取 2 年或 3 年預報另一年)並根據上一章設定的校驗方法與 LPM-FS、Logistic-FS、LPM-LASSO 三個模式比較其預報表現。

以預報基隆站 2 月降雨機率為例，從配適階段觀察其交叉驗證白氏得分圖，如圖 1，由邏輯斯迴歸所建立的模式，無論搭配何種變數選取方法，白氏得分均比其他兩個模式(LPM-FS、LPM-LASSO)低，顯示邏輯斯迴歸對於離散型預報目標建模確實有顯著改善，而以邏輯斯迴歸建模搭配前序選取變數選取法(Logistic-FS)表現最佳，白氏得分均可低於 0.3，而 LPM-FS 表現較差。從校驗階段來觀察，如圖 2，在配適階段表現最佳的 Logistic-FS，在此階段反而表現比其他三個模式差(白氏得分均較高)，尤其以預報 2010 年 2 月差異最明顯，其白氏得分超過了 0.4，而利用 LASSO 挑選變數的兩個模式白氏得分皆在 0.3。再進一步與氣候持續法比較，如圖 3，Logistic-FS 其白氏技術得分在 2007 年與 2010 年出現了負分與零分的狀況，表示在這兩年預報能力不佳，而 Logistic-LASSO 與 LPM-LASSO 其白氏技術得分均為正分，在 2010 年可明顯展現模式運用 LASSO 挑選變數後改善了降雨機率預報的能力。

從配適階段的 ROC 曲線觀察，如圖 4，若曲線下的面積(Area under curve, AUC)越大代表此模型越具有分類能力，logistic-FS 的 ROC 曲線下面積最大，其次是 LPM-LASSO 與 Logistic-LASSO，而 LPM_FS 的 ROC 曲線下面積最小，表示此模式其分類能力較差。藉由 *Youden index* 決定機率臨界值後，校驗各年分類預報狀況，同時檢視 TPR(即真實有下雨，經由模式判斷有雨比率)、FPR(即真實無下雨，經由模式誤判為有降雨比率)與準確率(Accuracy，經由模式正確判斷的比率)，如圖 5，觀察 2010 年 2 月降雨分類結果，Logistic-FS 分類傾向於有雨類別，使得 TPR(TPR=1)與 FPR 皆偏高，間接影響了準確度；LPM-FS 其分類傾向於無雨類別，使得 TPR 與 FPR(FPR=0)皆偏低，準確度也因而變低； LPM-LASSO 與 Logistic-LASSO 其分類狀況在此年表現較佳。若從整體交叉驗證來看，主要發現 Logistic-FS 其分類預報結果較不穩定且準確度較低。

進一步觀察基隆站 1 月至 12 月各月降雨機率預報交叉驗證白氏得分比較圖，如圖 6，其配適階段還是以 Logistic-FS 表現最好，Logistic-LASSO 其次，而 LPM-FS 較差；從校驗階段可發現，如圖 7，Logistic-FS 在某些月份白氏得分異於其他三個模型出現過高的狀況，變的相當不穩定，而 Logistic-LASSO 與 LPM-LASSO 表現相近，且其白氏得分略低於 LPM-FS，統整兩個階段來看，Logistic-LASSO 模型其各月降雨機率預報表現較佳。

五、結論

以邏輯斯迴歸建立降雨機率預報經實證確實改善了線性機率模型在統計上種種不適用性，並讓機率預報值落於合理的範圍。由本研究分析可發現，若以邏輯斯迴歸搭配前序選取來挑選預報因子，可在配適階段配出最佳模型，但在校驗預報階段卻顯得很不穩定，可能原因在於邏輯斯迴歸對於機率預報為較佳的配模模型，且以前序選取(Forward Selection)挑選變數也是以配模最佳化的前提來挑預報因子，兩者搭配起來出現了過分配適的狀況，造成預報階段不穩定的狀況。但以邏輯斯迴歸搭配最小絕對壓縮挑選機制(LASSO)來挑選預報因子，其挑選變數過程具有壓縮迴歸係數估計量作用，此作用可降低迴歸係數估計變異，讓所建的模型具有容忍預報階段資料改變的能力，使得整體預報更趨於穩定。

若期望降水機率預報模式在統計上能具有其合理性，則邏輯斯迴歸模式為較佳的降水機率建模模型，另外變數選取方法的搭配也是影響整體預報結果的關鍵，若以預報為目的，搭配最小絕對壓縮挑選機制(LASSO)來挑選預報因子，可避免模型過度配適的問題，讓模式更具有彈性，因此，邏輯斯迴歸輔以最小絕對壓縮挑選機制應用於降水機率預報是值得我們期待的工具。

六、參考文獻

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. Wea. Forecasting, 17, 783–799.

Bröcker, J., 2009: Regularized Logistic Models For Probabilistic Forecasting and Diagnostics. Monthly Weather Review, 2009 early online release.

Kumar, A., P. Maini, and S. V. Singh, 1999: An Operational Model for Forecasting Probability of Precipitation and Yes/No Forecast. Wea. Forecasting, 14, 38-48.

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

王政忠、陳雲蘭，2009：最小絕對壓縮挑選機制(LASSO)於天氣分析迴歸預報的應用。天氣分析與預報研討會論文彙編，中央氣象局，314-319

陳雲蘭、王政忠與張琬玉，2009：統計迴歸模式季內時間取樣差異測試。九十八年度中央氣象局自行研發計畫成果報告第 CWB 98-1A-03 號

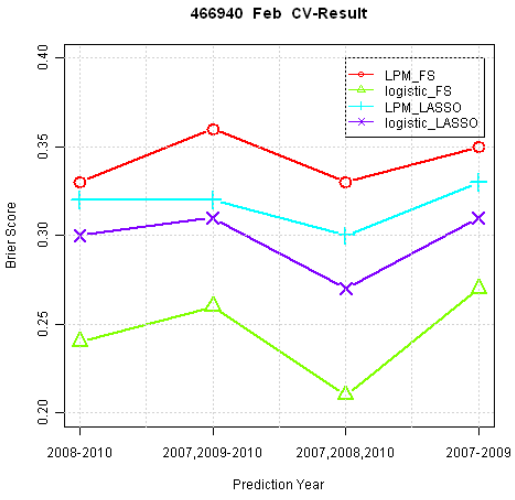

圖 1：基隆站 2 月各方法交叉驗證白氏得分圖(配適階段)

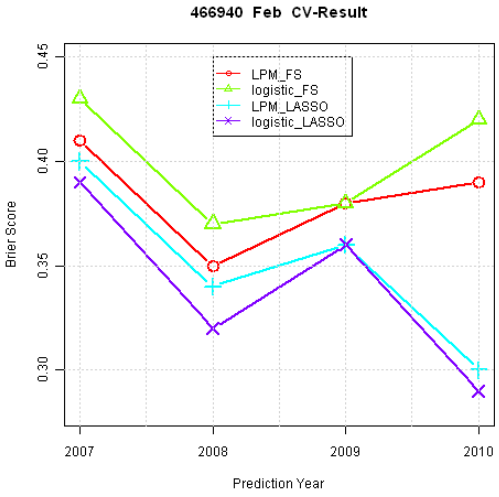

圖 2：基隆站 2 月各方法交叉驗證白氏得分圖(校驗階段)

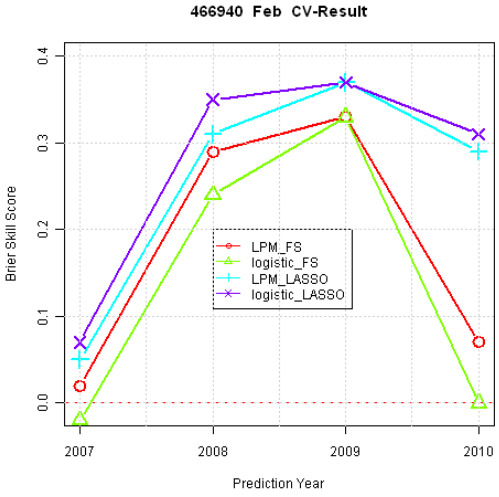

圖 3：基隆站 2 月各方法交叉驗證白氏技術得分圖(校驗階段)

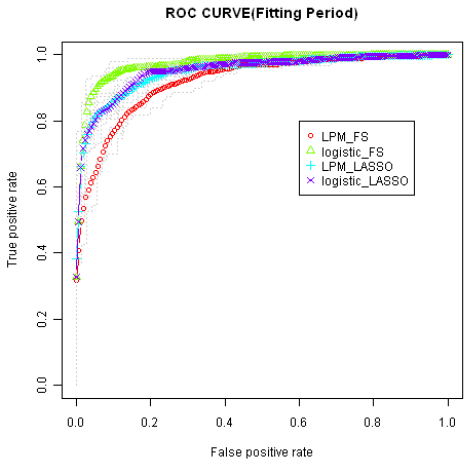

圖 4：基隆站 2 月各方法交叉驗證 ROC 曲線(配適階段)

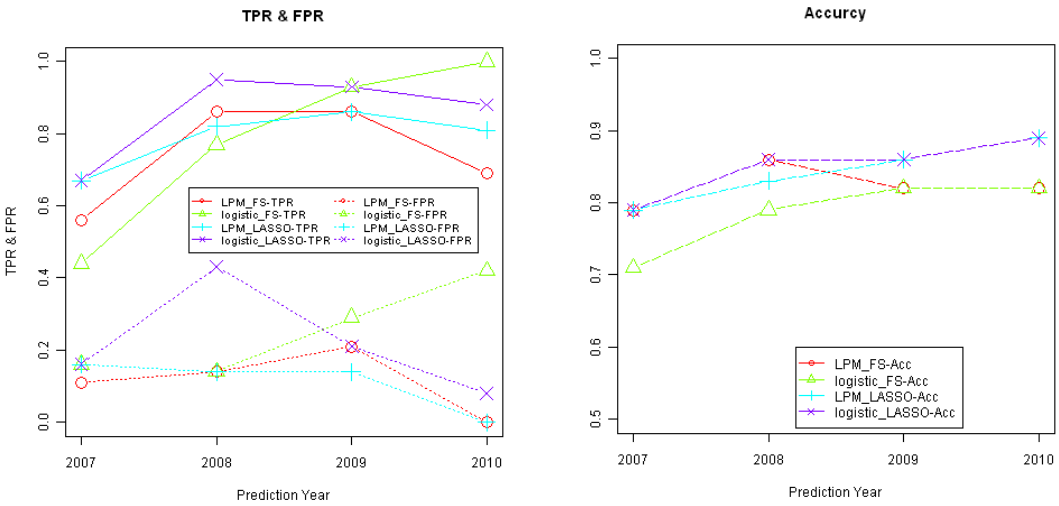

圖5：基隆站 2 月 POP 預報各方法分類狀況校驗比較圖(圖左上：TPR；圖左下：FPR。圖右：Accuracy)

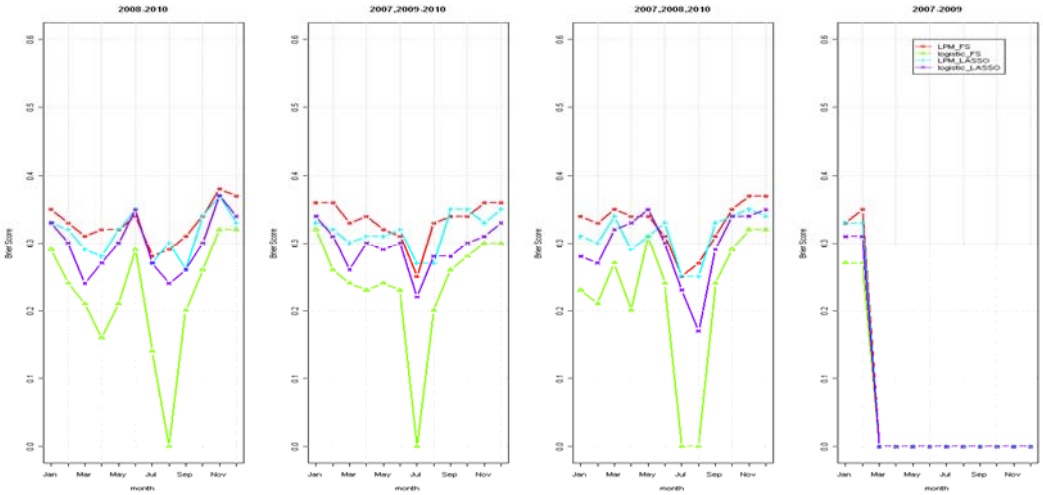

圖6：基隆站POP預報之白氏得分交叉驗證比較圖(三年配模資料)(配適階段)

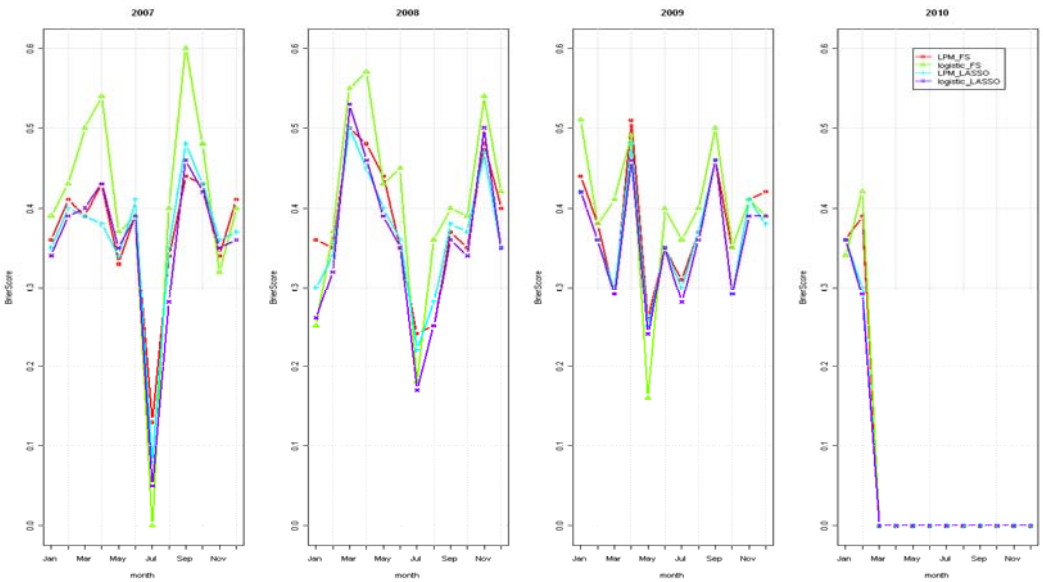

圖7：基隆站POP預報之白氏得分交叉驗證比較圖(一年校驗資料)(校驗階段)