使用虛擬變數探討統計預報之資料分季策略

郭孟坤

資拓宏宇國際股份有限公司

摘要

現行局內處理不同季節之統計預報時，一般以不同季節之資料分別配適其所需之統計模型，於預報時再分別以所屬之季節之模型產製所需之統計預報。本研究藉由使用虛擬變數(Dummy Variable)，輔以 Stepwise Selection 以及 AIC 等變數篩選之準則，將月份視爲變數，探討月份與各氣象變數之不同形式之交互作用(Interaction)對於應用線性迴歸(Linear Regression)預測日間最高溫(Tmax)以及應用邏輯斯迴歸(Logistic Regression)預測降雨機率(Probability of Precipitation, PoP) 在配適階段(Training Phase)及校驗階段(Testing Phase, Forecasting Phase)之預報能力，並針對 Tmax 探討逐月配適資料之分季策略於配適與校驗之表現。

研究結果顯示，若將月份視爲變數並強迫其進入模型，考慮各個氣象變數與月份間交互作用之變數選取策略，於配適階段可得到相對於未考慮月份時較好之結果，但對於校驗階段之預報結果並沒有實質之幫助。若分月配適模型，在配適階段可以得到最佳之結果，但在校驗階段之結果卻不如預期。此外，分月配適模型時若每個月前後各增加 15 天進入模型，可有效提升模型穩定性，對日間最高溫而言，將預測變數之個數限制爲 10 個亦可降低模型之誤差，但在校驗階段仍不及允許加入月份爲模型截距項之模型策略。因此，在考量模型之模型意義、預測效果、模型配適之時間成本、作業方便性之前提下，未來若需處理分月之問題，使用虛擬變數亦不失爲一可行之方案，但可不需考慮月份與其他氣象變數之交互作用。

關鍵字：Dummy Variable, Interaction, Linear Regression, Logistic Regression, Stepwise Selection, AIC.

一、前言

時至今日，使用統計方法進行數值天氣預報(Numerical Weather Prediction, NWP)之統計後處理(Statistically Postprocessing)已是一種有效降低 NWP 誤差之途徑，目前統計後處理在配適資料時主要分成 Perfect Prog(PP)、Model Output Statistics(MOS) 和 Reanalysis(RAN)三類，其中 PP 在建模時使用觀測資料建立線性模型，於預報時再以數值模式輸出之值代入模型取得統計預報值；MOS 使用數值模式之輸出建立線性模型，藉此掌握不同數值模式之特徵；RAN 則是藉由建立線性模型，以觀測資料預測重分析資料，之後再以重分析資料與模式輸出資料建立線性模型，藉此產生統計預報值，有關統計後處理之進一步資訊可參考 Marzban et al(2006)。

目前中央氣象局亦已有對於統計預報之相關研究及預報產品，如陳重功與羅存文(2010)之 MOS 測站定量降水預報指引、FIFOW 計畫下之測站與格點統計預報、王政忠與陳雲蘭(2010)使用邏輯斯迴歸(Logistic Regression)模型輔以最小絕對壓縮挑選機制(LASSO)預報降水機率以及陳雲蘭(2010)等人之季內時間取樣測試等。大抵上，在處理不同時間區段時，皆將不同

時間區段各自配適出模型，於預報時再選取各個時間區段之模型產製出預報值。舉例來說，若選擇以月份做爲分季之策略時，於配適模型時需每個月各自配適出一個模型，必要時再加上前後各十五天之資料以增加模型穩定度。

就作業上來說，過多組統計模型除了增加模型配適所需之時間外，亦增加作業系統之複雜度，在本文中我們引入虛擬變數(Dummy Variable)之概念，嘗試將月份視爲變數，探討不同資料配適策略於預報時之成效。在實際操作時，我們使用 Stepwise Selection 以及 AIC 等變數篩選之準則，探討月份與各氣象變數之不同形式之交互作用(Interaction)對於應用線性迴歸(Linear Regression)預測日間最高溫(Tmax)以及應用邏輯斯迴歸(Logistic Regression)預測降雨機率(Probability of Precipitation, PoP) 在配適階段(Training Phase)及校驗階段(Testing Phase, Forecasting Phase)之預報能力，並針對 Tmax 探討逐月配適資料之分季策略於配適與校驗之表現。

二、研究方法

（一）線性迴歸(Linear Regression)

若 n 爲樣本數，p 爲預測變數個數，令 $Y = X\beta + \varepsilon$，其中：

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \mathrm{M} \\ y_n \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \mathrm{M} \\ \beta_p \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \mathrm{M} \\ \varepsilon_n \end{bmatrix},$$

$$X = \begin{bmatrix} X_1^{'} \\ X_2^{'} \\ \mathrm{M} \\ X_n^{'} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \Lambda & x_{1p} \\ 1 & x_{21} & x_{22} & \Lambda & x_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ 1 & x_{n1} & x_{n2} & \Lambda & x_{np} \end{bmatrix}$$

若 $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2), i = 1,\Lambda\ ,n$，一般可使用最小平方估計式 $\hat{\beta} = (X'X)^{-1}XY$ 估計迴歸係數，其迴歸係數之變異數爲 $Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$。

（二）邏輯斯迴歸(Logistic Regression)

若現在 Y 爲二元之反應變數，其事件發生之機率爲 π，假設 $Y_i \overset{iid}{\sim} Bin(1,\pi)$，若我們想要建構出 π 與預測變數 $x_1, x_2, \Lambda\ , x_p$ 之間之關係，亦即 $\pi = \pi(x_1, x_2, \Lambda\ , x_p)$，可將 π 寫爲：

$$\pi_i = \pi(X_i^{'}) = \frac{\exp(X_i^{'}\beta)}{1+\exp(X_i^{'}\beta)}, i = 1,\Lambda\ n$$

移項之後則爲：

$$\ln\left(\frac{\pi(X_i^{'})}{1-\pi(X_i^{'})}\right) = X_i^{'}\beta, i = 1,\Lambda\ ,n$$

在估計參數時，通常採用最大概似估計法(Maximum Likelihood Estimation)估計參數，限於篇幅在此並不詳細介紹。

（三）因子(Factor)與虛擬變數(Dummy Variable)

在進行資料分析時，資料型態一般可分爲屬量(quantitative)與屬質(qualitative)兩大類，如溫度、雨量之類可以實際數據量化之資料皆屬於屬量之資料，而諸如性別、月份等無法以一組數字表現其所代表之強度資訊者，則爲屬質之資料，亦稱爲類別資料。

當一個有兩個以上分類之類別資料被拿來當作預測變數(Predictor)時，即稱之爲因子。因子可藉由使用虛擬變數之方式引入模型，舉例來說，若我們想以虛擬變數來表示男生和女生，可令：

$$d = \begin{cases} 1, male \\ 0, otherwise(\ female) \end{cases}$$

若一個含有 k 類之因子欲納入一個含有截距項之線性模型，則需設立 k-1 個虛擬變數以代表不同之分類，一般常見之變異數分析(ANOVA)亦是虛擬變數之一種應用。

（四）交互作用(Interaction)

在迴歸式中包含兩個以上預測變數之乘積項即稱之爲交互作用，交互作用可以幫助我們探討兩個以上之預測變數對反應變數(predictand)之聯合作用。以

下以一個簡單例子來講解一個屬質的預測變數與一個屬量之預測變數可能之交互作用。若我們現在欲運用身高 x 與性別 d 來預測體重 Y，身高與性別間之交互作用所形成之體重之 mean function 可以有以下四種型式：

模型一：男生與女生有共同之截距與斜率

$$E(Y \mid x) = \beta_0 + \beta_1 x$$

模型二：男生與女生各自有不同之截距與斜率

$$E(Y \mid x,d) = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 \cdot x \cdot d$$

當 $d = 1$ (male)時，其 mean function 爲：

$$E(Y \mid x, d = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x$$

而當 $d = 0$ (female)時，其 mean function 爲：

$$E(Y \mid x, d = 1) = \beta_0 + \beta_1 x$$

模型三：男生與女生有各自有不同之截距

$$E(Y \mid x,d) = \beta_0 + \beta_1 x + \beta_2 d$$

當 $d = 1$ (male)時，其 mean function 爲：

$$E(Y \mid x, d = 1) = (\beta_0 + \beta_2) + \beta_1 x$$

而當 $d = 0$ (female)時，其 mean function 爲：

$$E(Y \mid x, d = 1) = \beta_0 + \beta_1 x$$

模型四：男生與女生各自有不同之斜率

$$E(Y \mid x,d) = \beta_0 + \beta_1 x + \beta_3 \cdot x \cdot d$$

當 $d = 1$ (male)時，其 mean function 爲：

$$E(Y \mid x, d = 1) = \beta_0 + (\beta_1 + \beta_3)x$$

而當 $d = 0$ (female)時，其 mean function 爲：

$$E(Y \mid x, d = 1) = \beta_0 + \beta_1 x$$

（五）變數篩選(Variable Selection)

變數篩選我們主要使用逐步選取(Stepwise Selection)搭配 AIC(Akaiki Information Criterion)，AIC 之計算方式如下：

$$AIC = 2p - 2 \cdot \ln(L)$$

其中 p 爲變數個數，L 爲模型之概似函數(likelihood function)最大值之估計值，由式子中可看出，AIC 不僅反映出模型之適合度(Goodness of Fit)，對於選取過多參數導致過度配適(Overfitting)之情形亦給予對應之懲罰，原則上最小的 AIC 即代表最適合之模型。

在進行逐步選取時，首先須決定起始之模型，可依照對於預測變數之了解設定若對於模型參數沒有特別的偏好可選取加入一個參數時表現狀況最好之參數作爲起始參數，之後在每個階段則比較加入一個參數及減去一個參數 AIC 最小之模型做爲候選模型，並挑選候選模型中 AIC 最小之模型做爲下一個模型，重複此步驟，直至選取到增加或減少參數皆無法使 AIC 變小之模型爲止。

三、校驗準則

（一）屬質反應變數

1.平均絕對誤差(Mean Absolute Error, MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right|$$

其中 f_i 爲預測值，y_i 爲真實值。

2.均方根差(Root Mean Square Error, RMSE)

$$RMSE = \sqrt{\frac{\sum(f_i - y_i)^2}{n}}$$

其中 f_i 爲預測值，y_i 爲真實值。

（二）屬量反應變數

1.ROC(Receiver Operating Characteristic) Curve

令 p 表示 positive，n 表示 negative，對於一個二元分類(binary classfication)可定義：

True Positive(TP)：預測爲 p 其真實值也爲 p。

True Negative(TN)：預測爲 n 其真實值也爲 n。

False Positive(FP)：預測爲 p 其真實值爲 n。

False Negative(FN)：預測爲 n 其真實值也爲 p。

True Positive Rate(TPR, Sensitivity)：

TPR=TP/P=TP/(TP+FN)

False positive Rate(FPR)：

FPR=FP/N=FP/(FP+TN)

Accuracy(ACC)：

ACC=(TP+TN)/(P+N)

一個 ROC 曲線即是以 FPR 爲橫軸，TPR 爲縱軸之圖形，45 度線代表的是無效之分類，原則上一個良好的分類器之 ROC 曲線會越靠近圖形之左上角。

在決定分類之臨界點時，可利用 Youden Index：

$$Youden\,\text{index} = \underset{c}{Max}(TPR_c - FPR_c)$$

當我們認爲犯 FP 和 FN 所需付出之代價相同時，其在 ROC 曲線上所代表的即是最靠左側且斜率爲 1 之切點所代表之臨界機率值。

2.白氏得分(Brier Score, BS)

$$BS = \frac{1}{n}\sum_{t=1}^{n}(f_t - o_t)^2$$

其中 f_t 爲模型預測出來之機率值，當實際事件有發生時 o_t 爲 1，反之則爲 0，越小之 BS 代表所預報之機率值越傾向實際之事件。

四、研究設計與結果

本研究之資料使用台北站之日間最高溫與雨量之觀測資料，以及 NFS 模式初始時間爲 00，內插至台北上空之初始場，在雨量部分若有下雨視爲降雨，若沒有下雨或雨跡(Trace)則視爲沒有下雨。本研究試著從幾個面向來探討月份對於建立模型時之影響，首先我們將月份視爲因子，比較月份與其餘變數之四種可能之交互作用型態在配適與校驗階段之表現。另一方面，我們亦比較各個月份分別配適其迴歸模型於配適與校驗階段之結果。

（一）將月份視爲變數

1.以線性迴歸預測日間最高溫(Tmax)

此處以 2005~2007 之資料配適模型，並以 2008 年之資料校驗，共分爲以下四類模型策略，其中 Initial Model 表示模型之初始模型，Scope 表示於變數篩選階段允許納入之變數：

模型一：各月份有相同之截距與斜率

Initial Model:Tmax~925hpa 位溫

Scope: NFS 模式變數

模型二：各月份有不同之截距與斜率

Initial Model:Tmax~925hpa 位溫+月份

Scope: NFS 模式變數、月份、月份與 NFS 模式變數之交互作用

模型三：各月份有不同之截距

Initial Model:Tmax~925hpa 位溫+月份

Scope: NFS 模式變數、月份

模型四：各月份有不同之斜率

Initial Model:Tmax~925hpa 位溫

Scope: NFS 模式變數、月份與 NFS 模式變數之交互作用

各模型配適階段與校驗階段之 MAE 以及 RMSE 整理如下表：

Model	Phase	MAE	RMSE
Model 1	Training	1.209	1.5649
	Testing	1.2697	1.635
Model 2	Training	1.1017	1.4025
	Testing	1.3086	1.8782
Model 3	Training	1.2313	1.5789
	Testing	1.2782	1.6553
Model 4	Training	1.046	1.3486
	Testing	1.2991	1.7942

允許模型納入月份與其他變數之交互作用項時，於配適階段可得到較好之結果，但於校驗階段其結果反而不如未加入交互作用項之模型。值得一提的是，雖然模型三之初始模型含有月份，但經過逐步選取之後的模型並沒有包含月份，代表月份在變數篩選的過程中自然地被淘汰掉了，亦即對於預測日間最高溫來說，在考慮月份因子但不考慮月份與其他氣象變數之交互作用之前提下，其他氣象變數已足以解釋與日間最高溫之間之線性關係。

綜合此階段之結果，可發現較複雜之模型（模型二與模型四）僅在配適階段有較好之表現，於校驗時反而不如相對較簡單之模型（模型一與模型三）。考慮配適模型所需之時間成本以及成效之考量下，未來在配適模型時，若欲考慮月份之影響，可不需考慮月份與氣象變數之交互作用，亦即使用模型三之策略。

2.以邏輯斯迴歸預測降雨機率

此處以 2005~2007 之資料配適模型，並以 2008 年之資料校驗，模型策略如同之前以線性迴歸預測 Tmax，僅將迴歸模型替換爲邏輯斯迴歸，而初始之預測變數替換爲 500hpa 之下平均相對濕度，各模型於各階段之 Accuracy 以及 Brier Score 整理如下表：

Model	Phase	Accuracy	Brier Score
Model 1	Training	0.769	0.154
	Testing	0.765	0.346
Model 2	Training	0.786	0.146
	Testing	0.751	0.357
Model 3	Training	0.784	0.15

	Testing	0.765	0.349
Model 4	Training	0.79	0.143
	Testing	0.742	0.356

如同以線性迴歸預測日間最高溫一樣，允許模型納入月份與其他變數之交互作用項僅在配適階段有較好結果，於校驗階段反而不如未加入交互作用項之模型，觀察 ROC 曲線亦有相同之結論。

（二）逐月配適模型預測日間最高溫

此處我們針對日間最高溫，使用 2005~2007 之資料，並以 2008 年之資料校驗，各月份初始之預測變數爲 925hpa 位溫，各模型於各階段之 MAE 以及 RMSE 整理如下表：

Model	Phase	MAE	RMSE
逐月配適模型	Training	0.8732	1.1409
	Testing	1.6589	2.2312
逐月配適模型(加上前後 15 天)	Training	1.0211	1.3161
	Testing	1.4165	1.9461

在配適階段時，逐月配適模型可得到最好的結果，但在校驗階段時其預報能力卻大幅下滑，代表逐月配適模型過度反映 2005~2007 年間各個月份中各氣象變數所無法捕捉日間最高溫之隨機誤差，如此之模型並不適合拿來預測。而考慮逐月配適模型並加上前後 15 天在配適階段雖不及不加上前後 15 天之結果，但在校驗階段之表現卻遠優於不加上前後 15 天之結果（MSE 與 RMSE 皆減少約 15%），顯示逐月配適模型並加上前後 15 天之策略的確可有效增加模型之預測能力。

若進一步比較逐月配適模型以及將月份視爲變數之結果，可發現逐月配適模型並加上前後 15 天之策略於配適階段皆優於將月份視爲變數之四種模型策略，但於校驗階段卻不及將月份視爲變數之結果，將其與模型意義最接近（但校驗結果最差）的模型二比較，其結果仍略遜一籌。若是與模型三之校驗結果比較，將月份視爲變數再進行變數篩選之模型，MAE 可較逐月配適模型加上前後 15 天降低約 10%，RMSE 則可降低約 15%，降低之程度和逐月配適模型與加上前後 15 天再配適模型相當，而 RMSE 降低幅度大於 MAE，代表將月份視爲變數之策略較不容易給予誤差相對而言較大之預報。

（三）更多交叉驗證

此處我們比較逐月配適並加上前後 15 天(逐月 ± 15)與允許加入月份但不允許月份與其他變數之交互作用(model 3)兩種模型，取 2005~2008 之資料，以任意三年建模來預報另外一年，比較其預報之結果，各模型以檢定結果最顯著之變數做爲第一個加入之變數，變數篩選方法除了逐步選取(stepwise)之外，亦加入前序選取但只取十個變數(forward 10)、前序選取(forward)另外兩種變數選取之方法，各年 MAE 以及 RMSE 最小之預報標上底線後，結果如下：

year	Model	variable selection	MAE	RMSE
2005	逐月 ± 15	forward 10	1.59	2.0737
		forward	1.6017	2.0841
		stepwise	1.5392	2.0075
	model 3	forward 10	1.3966	1.8448
		forward	**1.3337**	**1.8274**
		stepwise	**1.3337**	**1.8274**
2006	逐月 ± 15	forward 10	1.5628	2.1036
		forward	1.5959	2.1581
		stepwise	1.5651	2.0807
	model 3	forward 10	1.5445	1.9729
		forward	1.4691	1.8921
		stepwise	**1.4503**	**1.8744**
2007	逐月 ± 15	forward 10	1.3346	1.7704
		forward	1.3993	1.8458
		stepwise	1.3728	1.8229
	model 3	forward 10	1.3104	1.748
		forward	**1.2592**	**1.6448**
		stepwise	1.2628	1.6463
2008	逐月 ± 15	forward 10	1.2947	1.7656
		forward	1.3555	1.8774
		stepwise	1.4054	1.9029
	model 3	forward 10	**1.2818**	1.693
		forward	1.2978	1.6936
		stepwise	1.2982	**1.6734**

若只觀察逐月 ± 15 之結果，可發現相較於前序選取但不限制變數個數，前序選取但只選取十個變數

可同時降低 MAE 與 RMSE，表示在進行前序選取預測日間最高溫時，限制變數個數爲十個之情形下，可有效抑制過度配適之情形；但若是看 2005 之結果，限制選取十個變數之結果反而不如逐步選取之結果，表示在預測本年時其建模過程中逐步選取即可有效地在所有預測變數中選取較爲適切之預測變數組。

而觀察 model 3 之結果，可發現在 2005、2006、2007 年之預報結果中，限制變數之個數並無法有效降低預報之誤差，表示限制只選取十個變數對於預報這三年之結果是不足的，增加變數之數目有助於提升預報的結果。而看 2008 之預報結果，雖然只選取十個變數之 MAE 較小，若比較 RMSE，則是逐步選取之結果較好。綜觀 model 3 之結果，可發現在不限制變數個數之情形之下，forward 與 stepwise 之結果並無太大差異，而限制變數個數並未確實有效提升模型之穩定性。

綜觀逐月±15 與 model 3 之結果，可以發現在 2005、2006、2007 年之結果中，model 3 之 MAE 與 RMSE 皆可較逐月±15 之結果降低約 10%，2008 年 model 3 與逐月±15 之差距雖較小，但 model 3 之 RMSE 相較於逐月±15 仍可降低約 7%。整體看來，使用 model 3 但不限制變數個數之模型策略優於逐月±15 並限制選取十個變數之策略。

五、結論

研究結果顯示，若將月份視爲變數並強迫其進入模型，考慮各個氣象變數與月份間交互作用之變數選取策略，於配適階段可得到相對於未考慮月份時較好之結果，但對於校驗階段之預報結果並沒有實質之幫助。若分月配適模型，在配適階段可以得到最佳之結果，但在校驗階段卻會得到最差之預報。此外，分月配適模型時若每個月前後各增加 15 天進入模型，可有效提升模型穩定性，對日間最高溫而言，將預測變數之個數限制爲 10 個亦可降低模型之誤差，但在校驗階段之結果卻不如預期。

若考慮加入月份做爲模型之截距項，可發現在不限制變數個數之情形之下，forward 與 stepwise 之結果並無太大差異，而限制變數個數爲十個並未確實有效提升模型之穩定性。因此，在考量模型之模型意義、預測效果、模型配適之時間成本、作業方便性之前提下，未來若需處理分月之問題，使用虛擬變數亦不失爲一可行之方案，但可不需考慮月份與其他氣象變數之交互作用。

六、參考文獻

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。天氣分析與預報研討會論文彙編(99)，中央氣象局，台北市，236-241

陳重功、羅存文，2010：中央氣象局降水統計預報之探討。天氣分析與預報研討會論文彙編(99)，中央氣象局，台北市，226-230

陳雲蘭、王政忠、張琬玉，2010：統計迴歸模式季內時間取樣差異測試。天氣分析與預報研討會論文彙編(99)，中央氣象局，台北市，264-269

Marzban, C., Sandgathe, S., Kalnay E., 2006: MOS, Perfect Prog, and Reanalysis. Monthly Weather Review Vol.134, 657-663

Weisberg, S., 2005: Applied Linear Regression. John Wiley & Sons, New Jersey

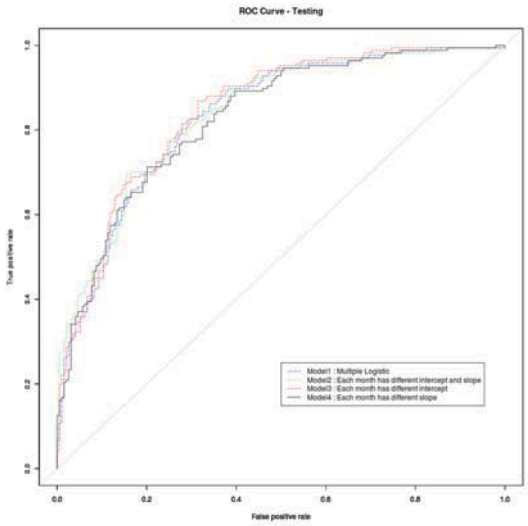

圖 1：ROC Curve – Testing