論述 》統計 · 調查

運用統計技術優化大專 1 年級學生數預測模型

為應大專校院生源減少及大專轉型退場等問題，爰運用妥適統計技術，提升大專 1 年級學生人數預測確度，俾先期提供正確資訊，作為教育政策重要參考。

郭溫慈、曾仁人、金允文、程冠瑜（教育部統計處科員、專員、專案助理、專案助理）

壹、前言

教育爲國家發展的根基，學生人數規模及未來變動情形實係制定教育政策、資源配置及擘劃教育藍圖之重要參據。爲因應近年少子女化、生源減少及大專轉型退場等問題，教育部按年預測各階段學生人數，以即時掌握教育發展變化，提供教育部政策訂定及執行之參考。大專 1 年級因係國民教育階段結束後邁入高等教育之起點，其學生[1]人數預測結果備受各界關注，爰本文將以該類學生爲研究主體，運用分段迴歸（Segmented Regression）及時間序列（Time Series）等統計技術，優化大專新生數預測模型，提升預測確度及品質。

貳、現況分析

一、大專 1 年級學生人數預測結果

104 學年以前，因大專 1 年級學生所對應之出生時期尚未受少子女化現象衝擊，學生人數僅受生肖效應影響，約維持在 25 萬人至 27 萬人之水準；惟因出生率下降，自 105 學年起人數大致呈減少趨勢，依原預測方法，117 學年之人數更因虎年效應而降至 14.1 萬人之低點，較 108 學年劇減 3 成 5。

二、現行預測方法

目前大專 1 年級學生人數之預測方法如下：

大專 1 年級學生人數 = 上學年高級中等學校 3 年級學生人數 × 近 5 學年就學機會率平均，其中，當學年就學機會率 = 當學年大專 1 年級學生人數 ÷ 上學年高級中等學校 3 年級學生人數

高級中等學校 3 年級因仍屬國民教育階段，學生人數與對應之出生人口數相近，仍維持原預測方式；至就學機會率大致可反映預測主體與對應出生人口之差距，隱含部分高級中等學校應屆畢（修）業生因生涯規劃、家庭、經濟、健康或國際間流動等因素，並未於畢業後隨即進入國內大專教育體系；或部分高級中等學校非應屆畢（修）業生於畢業數年後回流教育體系等因素。而每年大專 1 年級學生均以就學機會率近 5 年平均作爲預測基準，係移動平均（Moving Average）法之應用實例，雖較簡易，惟觀察其預測結果（圖 1），76 學年以前因學生人數變化平緩，移動平均法預測能力尚佳，迄 80 至 95 學年因廣設大學致大專校數增加，就學機會率大增，學生人數變化較爲明顯，移動平均法延後反映現況之問題隨即浮現，且平均年數之選擇，亦有流於主觀之疑慮，爰改採統計技術中之模型配適概念，運用迴歸分析（Regression Analysis）和時間序列統計方法，針對就學機會率建立預測模型，以解決前揭問題。

參、模型選取

一、迴歸模型

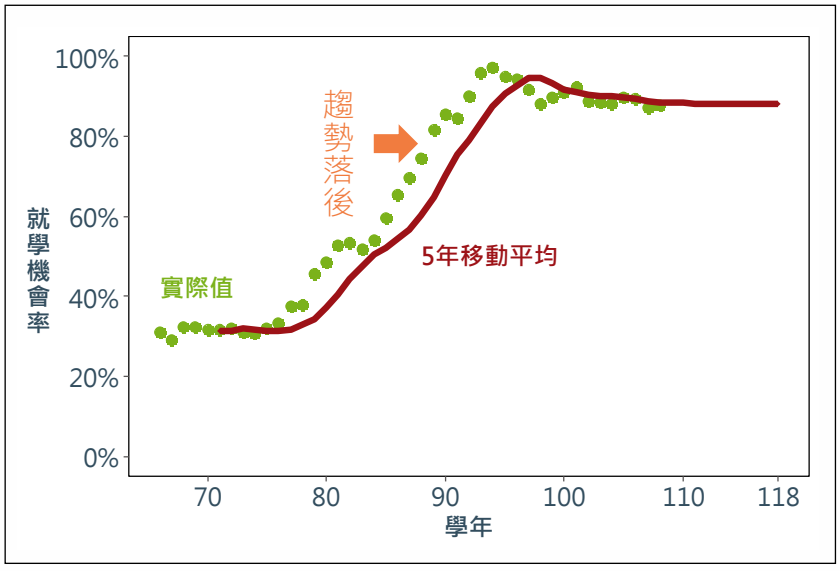
迴歸模型（Regression Model）主要用於解釋或預測變數之間的關係，即透過樣本建立因變數（Dependent Variable）及自變 數（Independent Variable）之函數模型，藉以探討變數成因及預測未來變動。本文使用迴歸模型如下：

（一）多項式迴歸模型（Polynomial Regression Model）通常用於曲線函數，可以包含兩個或兩個以上的預測變數，且每個預測變數都可有不同的乘冪。由大專 1 年級就學機會率圖發現有 2 個較明顯反曲點（圖 1），爰選取 3 次多項式迴歸模型如下：

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \beta_3 X_t^3 + \varepsilon_t \quad , \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

（二）分段迴歸模型（Segmented

圖 1　現行預測結果



資料來源：教育部統計處編製。

Regression Model）係指當因變數與自變數在不同範圍區間，具有不同斜率的迴歸模型，而大專 1 年級就學機會率與學年度確於不同範圍，存在不同線性關係（上頁圖 1），亦可選取 2 個分段點之迴歸模型如下：

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 (X_t - \varphi_1)_+ + \beta_3 (X_t - \varphi_2)_+ + \varepsilon_t \ ,$$
$$\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

$(X_t - \varphi_1)_+ = (X_t - \varphi_1) \times I(X_t \geq \varphi_1)$ ，其中

$$I(\cdot) = \begin{cases} 1, X_t \geq \varphi_1 \\ 0, X_t < \varphi_1 \end{cases}$$

$(X_t - \varphi_2)_+ = (X_t - \varphi_2) \times I(X_t \geq \varphi_2)$ ，其中

$$I(\cdot) = \begin{cases} 1, X_t \geq \varphi_2 \\ 0, X_t < \varphi_2 \end{cases}$$

二、時間序列模型

時間序列模型（Time Series Model）係指以時間順序型態蒐集的觀測值，可利用歷史時間序列資料來建置模型以預測未來的結果，而就學機會率係爲各學年度所蒐集大專 1 年級學生人數除以上學年高級中等學校 3 年級學生人數之比率，具時間順序特性，適合建立時間序列模型，本文採用模型如下：

（一）移動平均模型（Moving Average Model）係指現在到過去 q 期隨機誤差（Random Error）之加權平均，q 階移動平均模型（MA(q)）

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \ ,$$
$$\varepsilon_t \overset{iid}{\sim} WN(0, \sigma^2)$$

（二）自我迴歸模型（Auto-regressive Model）係指前期時間序列資料爲自變數之迴歸， p 階自我迴歸模型（AR(p)）

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \ ,$$
$$\varepsilon_t \overset{iid}{\sim} WN(0, \sigma^2)$$

（三）ARMA 模型（Auto-regressive Moving Average Model）係指結合 AR(p) 及 MA(q) 之模型，以 ARMA(p,q) 表示

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \ ,$$
$$\varepsilon_t \overset{iid}{\sim} WN(0, \sigma^2)$$

（四）ARIMA 模型（Auto-regressive Integrated Moving Average Model）係指結合差分、AR(p) 及 MA(q) 之模型，以 ARIMA(p,d,q) 表示

$$\Phi(L)(1-L)^d Y_t = \Psi(L)\varepsilon_t \ ,$$

其中 Φ(L) 爲 AR(p) 模型，Ψ(L) 爲 MA(q) 模型，d 爲差分次數，$\varepsilon_t \overset{iid}{\sim} WN(0, \sigma^2)$ 。

三、序列相關檢定

由於最適預測模型之殘差項應符合無序列相關特性，故本文採用 Ljung 與 Box（1978）提出的Q統計量以檢定殘差項：

$$Q = T(T+2) \sum_{i=1}^{n} \frac{\hat{\rho}_i^2}{T-i} \sim X^2(n) \ ,$$

其中 T 爲樣本數，$\hat{\rho}_i$ 表相距 i 期之自我相關係數。若預測模型之殘差項仍存序列相關，則須利用差分等方式重新配適模型，以消除序列相關特性。

四、預測評估指標

本作業先以 9 成資料（66 至 103 學年）建置模型，另 1

成（104 至 108 學年）則爲驗證資料，當確認所建置模型之殘差項已無序列相關特性，即以預測評估指標輔助觀察 95% 信賴區間預測圖形之方式，選定最適模型，其中預測評估指標數值愈小，模型預測精準度愈高：

（一）均方誤差（Mean-Square Error）

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

（二）平均絕對百分比誤差（Mean Absolute Percentage Error）

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

（三）平均絕對誤差（Mean Absolute Error）

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

本作業整體過程可以下圖 2 表示。

肆、模型配適

一、迴歸模型

（一）3 次多項式迴歸模型

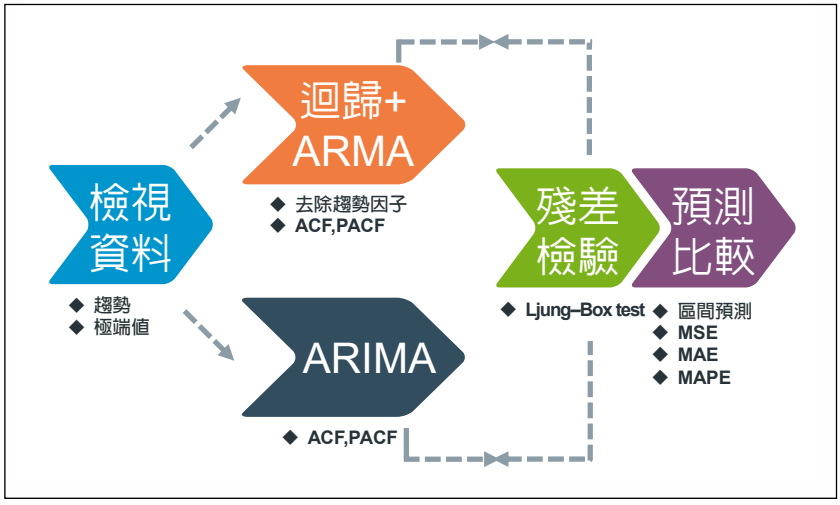
3 次多項式迴歸模型之 adjusted-R-square 雖高達 0.982，惟囿於模型最末轉折點具有因變數下降之特性，致就學機會率預測值於 4 學年間急速下滑至 6 成以下，明顯低於現況的近 9 成比率，爰先行捨棄 3 次多項式迴歸模型。

（二）分段迴歸模型

先行計算本模型之兩個分段點 76.270 和 93.669，配適結果 adjusted-R-square 高達 0.991，惟無論從殘差圖或 Ljung-Box 檢定均可發現殘差間存在序列相關，而由分段迴歸殘差的自我相關（ACF）和偏自我相關（PACF）圖形觀察發現，ACF 呈緩慢消失的樣態，而 PACF 則於落後 1 期後消失（該下頁圖 3 之分段迴歸），表示分段迴歸的殘差需建立 AR(1) 模型；惟爲不遺漏任何最適模型之可能性，再建立 MA(1) 模型。

建立分段迴歸搭配 AR(1) 或 MA(1) 之模型後，殘差圖即呈現隨機跳動的樣態； ACF 及 PACF 圖形均落在 95% 信賴區間內，顯示已無自我相關；Ljung-Box 檢定結果也顯示殘差間彼此獨立，故將兩個模型皆納入最適模型比較之中（第 79 頁表 1、下頁圖 3 之分段迴歸搭配 AR(1)、MA(1)）。

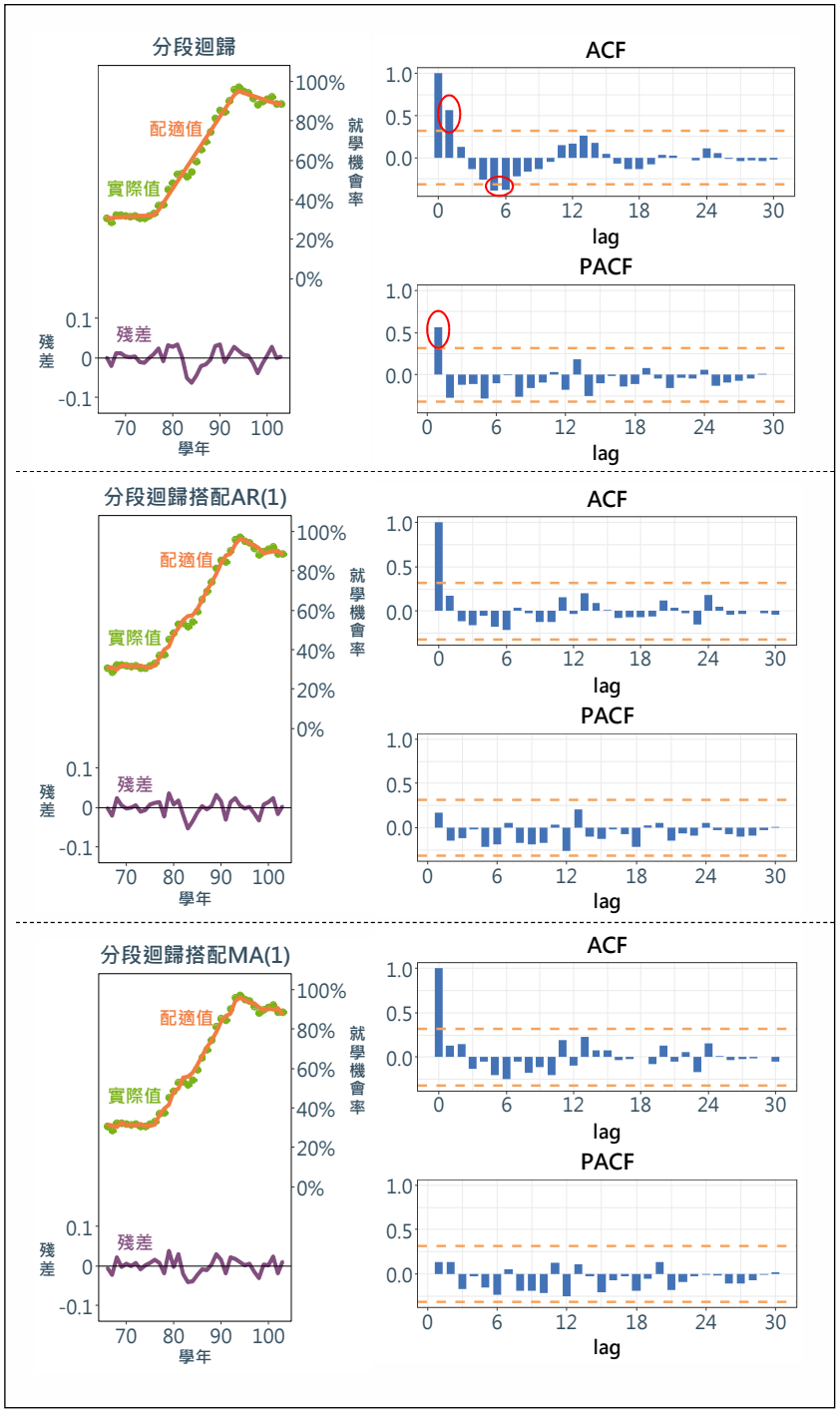
圖 2　建立模型流程



資料來源：教育部統計處編製。

論述 》統計 · 調查

圖 3　分段迴歸模型之殘差分析



資料來源：教育部統計處編製。

二、ARIMA 模型

由大專 1 年級就學機會率之 ACF 與 PACF 圖形觀察發現，ACF 呈現逐期緩慢消失樣態，表示就學機會率存在長期趨勢，須以一次差分的方式排除趨勢因子，使資料呈現定態（stationary），再進行後續的時間序列分析（下頁圖 4）。因 1 次差分後的就學機會率之 ACF 與 PACF 皆在落後 1 期後消失，故判斷需建立 ARIMA(0,1,1) 或 ARIMA(1,1,0) 模型。進一步觀察這兩個模型的殘差圖、ACF 與 PACF 圖形及 Ljung-Box 檢定結果後得知，殘差皆符合獨立的特性，爰納入後續模型比較（下頁表 1、第 80 頁圖 5 之 ARIMA(0,1,1)、ARIMA(1,1,0)）。

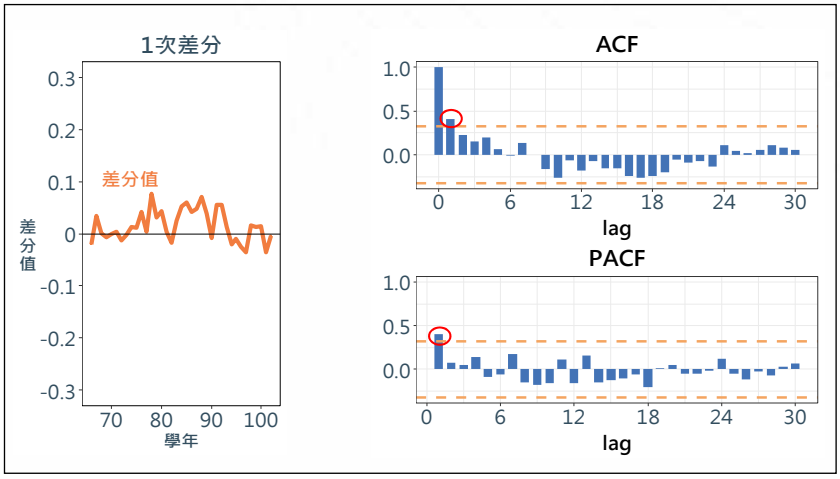
三、預測模型比較

模型配適完成後，再利用 104 至 108 學年驗證資料對上述 4 個模型及原先之 5 年移動平均模型進行區間、單點預測等精確度比較，以選擇最

適模型。由第 81 頁圖 6 可知分段迴歸分別搭配 AR(1) 及 MA(1) 之 2 個模型區間預測最窄，適合進行高、中、低區間預測的模型，惟由單點預測來看，預測就學機會率下滑速度較快，且低於實際就學機會率，MSE、MAPE 及 MAE 值明顯高於其他模型，單點預測能力最差。

至於 ARIMA(1,1,0) 與 ARIMA (0,1,1) 模型的預測區間雖寬，

圖 4　就學機會率 1 次差分之 ACF 與 PACF 圖



資料來源：教育部統計處編製。

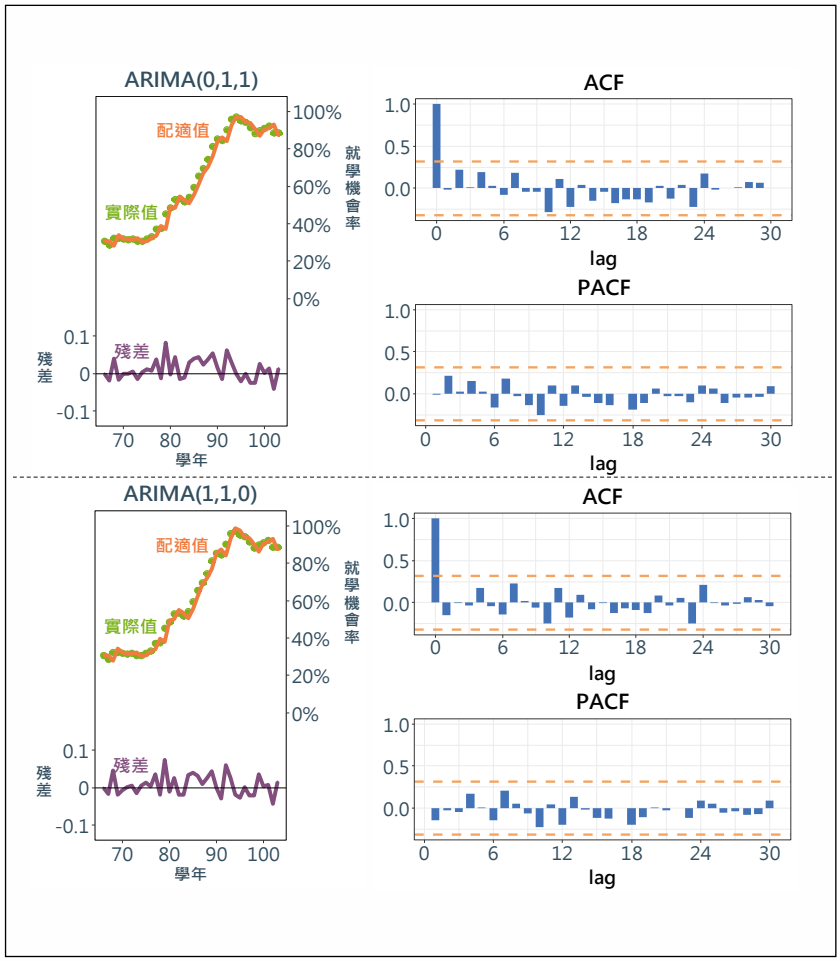
表 1　模型配適結果

預測模型	Variable	Estimate	Std. Error	p-value	Ljung–Box test（註）p-value
分段迴歸搭配 AR(1)	AR(1)	0.5589	0.1318	0.0000***	0.2708
	截矩	0.2504	0.2090	0.2309	
	學年 (X_t)	0.0009	0.0029	0.7631	
	$(X_t - \varphi_1)_+$	0.0353	0.0037	0.0000***	
	$(X_t - \varphi_2)_+$	-0.0437	0.0041	0.0000***	
	$I(X_t \geq \varphi_1)$	0.0077	0.0201	0.7041	
	$I(X_t \geq \varphi_2)$	-0.0044	0.0203	0.8298	
分段迴歸搭配 MA(1)	MA(1)	0.6242	0.1489	0.0000***	0.3949
	截矩	0.3408	0.1924	0.0766	
	學年 (X_t)	-0.0004	0.0027	0.8858	
	$(X_t - \varphi_1)_+$	0.0360	0.0031	0.0000***	
	$(X_t - \varphi_2)_+$	-0.0430	0.0034	0.0000***	
	$I(X_t \geq \varphi_1)$	0.0191	0.0213	0.3696	
	$I(X_t \geq \varphi_2)$	0.0012	0.0187	0.9486	
ARIMA(0,1,1)	MA(1)	0.4096	0.1259	0.0011***	0.9272
ARIMA(1,1,0)	AR(1)	0.5385	0.1347	0.0000***	0.3363

說　　明：*** 係指在顯著水準 α=0.05 時，p-value ＜ 0.05 有顯著差異。
註：虛無假設 (H_0)：殘差彼此獨立。
資料來源：教育部統計處編製。

論述 》統計 · 調查

圖 5　ARIMA 模型之殘差分析



資料來源：教育部統計處編製。

表 2　ARIMA (1,1,0) 模型

預測模型	Variable	Estimate	Std. Error	p-value	Ljung–Box test（註）p-value
ARIMA(1,1,0)	AR(1)	0.5296	0.1279	0.0000***	0.3498

說　　明：*** 係指在顯著水準 α=0.05 時，p-value ＜ 0.05 有顯著差異。
註：虛無假設 (H_0)：殘差彼此獨立。
資料來源：教育部統計處編製。

惟於近期預測仍較現行 5 年移動平均模型窄，顯示在區間預測精確度上相對優於 5 年移動平均模型；由單點預測能力做爲最適模型之評判標準觀察，ARIMA(1,1,0) 之 MSE、MAPE 及 MAE 值最小，表示單點預測精確度最佳。於區間、單點預測綜合考量下，最終選取 ARIMA(1,1,0) 做爲本文之最佳模型（下頁圖 6）。

四、最終配適模型

將全數資料（66 至 108 學年）納入 ARIMA(1,1,0) 模型中，再透過殘差圖、ACF、PACF 及 Ljung-Box 檢定，以確認殘差不存在序列相關現象後，可得本文之最終配適模型（表 2）。

伍、結語

運用統計技術解決估算及預測困難係統計工作者之重要使命，此一做法非但可排除作業中之主觀成分，降低誤差，亦可使整體作業具備理論

基礎。未來本部將運用本作業模式，逐步將統計理論導入各教育階段學生數及教師數之預測作業，期全面提升教育統計預測工作之品質，精進人員職能，並充分發揮教育統計資訊之「證據基礎決策（Evidence-Based Policy Making，EBPM）」之功能。

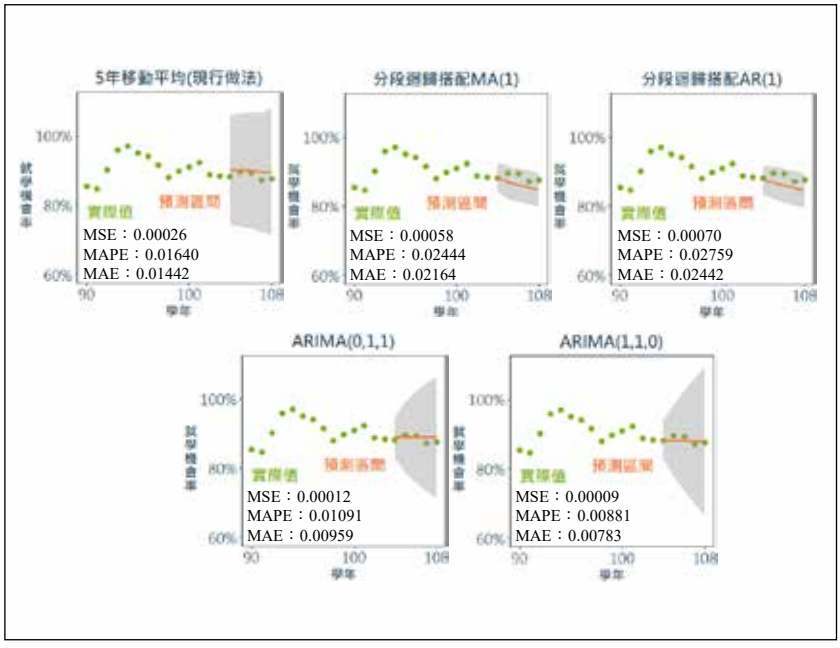
註釋

1. 大專 1 年級學生係指本國籍日間部、進修部大學四年制及二專之 1 年級學生，包含非應屆高級中等學校畢（修）業生回流、舊生復學等。

參考文獻

1. 吳宗正（1993），迴歸分析，臺北市：三民書局。
2. 陳旭昇（2013），時間序列分析：總體經濟與財務金融之應用，臺北市：東華書局。
3. 楊踐爲、李家豪、類惠貞（2007），應用時間序列分析法建構台灣證券市場之預測交易模型，中國管理評論國際學報，10 卷 3 期。
4. Muggeo, Vito M. R.（2008）. Segmented: An R package to Fit Regression Models with Broken-Line Relationships. R News, No.8(1), p. 20-25.❖

圖 6　各模型預測比較



資料來源：教育部統計處編製。