第三章、研究方法

儘管時間序列模型能夠根據觀測值的序列提供良好的預測，然而若是希望納入更多資訊，像是颱風天數、假日天數等，進而更完整的解釋遊客人數的變化，一般研究常用的ARIMA模型便無法滿足我們的需求。因此本研究採用 Forecasting: principles and practice (Rob J Hyndman, George Athanasopoulos, 2013) 第九章的動態迴歸模型(Dynamic Regression Model)，來描繪2008年開放陸客來台以及2011年開放陸客自由行此二政策以及颱風天數、假日天數等因素對於遊客人數變化之間的關係。

由於動態迴歸模型(Dynamic Regression Model)是由複迴歸模型(Multiple Regression)以及差分整合移動平均自迴歸模型(ARIMA)之混合模型。以下將簡要說明複迴歸模型及ARIMA模型，再說明動態迴歸模型及其建構程序，最後再說明模型選取之判別方法。

一、複迴歸(Multiple Regression)

(1)模型

$$y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}+e_i \ , \ i=1,\cdots,n$$

y為反應變數，x為解釋變數，β為估計參數，ei為殘差。
β可以視為每個解釋變數x對於y之變異產生了多少貢獻。一般而言，此模型必須符合以下假設：

(1) 所有解釋變數x不得線性相依(linearly dependent)

(2)$E(e_i)=0 \ ; E(e_i^2)=\sigma^2 \ ; E(e_i e_{\not i})=0$

(3) $e_i \sim i.i.d.\mathcal{N}(0,\sigma^2)$

$$y_i=\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}+e_i,=1,\cdots,n$$

(2) 參數估計

為了使配適之迴歸線距離所有觀察值有最小距離，意即：

$$SSE=\sum(y_i-\hat{y}_i)^2=\sum e_i^2 \text{具有最小值}$$

因此我們可以針對k個β進行偏微分，並令各個式子為零，求解未知數即可得到β之估計值。將估計值代入，即可得以下預測式：

$$\hat{y}=\hat{\beta}_{1n}x_1+\cdots+\hat{\beta}_k x_k$$

而將真實資料的$(x_1,\cdots,x_n)$代入本式，所求得知$\hat{y}$為該方程式配適值(fitted value)，而$e_i = y_i - \hat{y}_i$則為殘差值(residuals)。

(3)參數推論

在殘差是常態分配$\mathcal{N}(0,\sigma^2)$的假設下，我們可以得知

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right), j = 1,\cdots,k$$

也就是

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \ \mathcal{N}(0,1), j = 1,\cdots,k$$

同時，經由標準誤的分配

$$\sum_{i=1}^{n} e_i^2 \Big/ \sigma_0^2 = (n-k)\ \hat{\sigma}^2 \ / \ \sigma_0^2 \sim x^2(n-k)$$

可以得知分母為標準誤得分配為t分配：

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim t x^2(n-k), j = 1,\cdots,k$$

得知了參數的分配，便可以對參數進行推論，虛無假設以及對立假設如下：

$$\mathcal{H}_0 : \beta_i = 0 \quad \mathcal{H}_{01} : \beta_i \neq 0$$

在設定好的α下(通常是0.05)，棄卻虛無假設表示該解釋變數之參數β在特定信心水準下不為零，意即該解釋變數x對於y之變異具有解釋力。一般在統計軟體會顯示其p-value，若是p-value小於設定的α，表示該解釋變數x顯著。

二、差分整合移動平均自迴歸模型(ARIMA)

由於主旨在於用ARIMA模型的概念去描繪具有時間序列特性的殘差，因此僅簡單介紹ARIMA模型。ARIMA模型具有自我相關的時間序列資料一種良好的描繪方式，可表示為：

$$\acute{y}_t = C + \emptyset_1 \acute{y}_{t-1} + \cdots + \emptyset_p \acute{y}_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t,$$

其中，$\acute{y}_t$ 為差分後的序列資料，通常可以把該時間序列資料的ARIMA模型用 y_t~ARIMA(p,d,q)表示，其中
p=自迴歸的項數;
d=差分次數(階數);
q=移動平均的項數.

將上述式子改寫，可得：

$$(1 - \emptyset_1 B - \cdots - \emptyset_1 B^p)(1 - B)^d y_t = C + (1 + \emptyset_1 B + \cdots + \emptyset_q B^q) e_t$$

在動態迴歸模型中將會使用上述ARIMA模型的方式來描繪殘差項。

三、動態迴歸模型(Dynamic Regression Model)

將原先的複迴歸模型的殘差改寫，可得：

$$y_t = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + n_i,$$
$$(1 - \emptyset_1 B)(1 - B) n_t = (1 + \emptyset_1 B)\, e_t,$$

此時在這樣的模型假設下，估計參數β並不是最好的估計法。由於參數的分配改變，檢定參數所使用的t test亦不能採用。以往選模型常使用的AIC亦不能用來判別模型好壞。

此外，採用這樣的模型必須先確定資料是定態(Stationary)，或者這些非定態變數(Non-stationary variables)具有共整合性(Co-integrated)。

由於β的估計已經不準確了，因此可以採用漸進的方式建構模型。通常是先假設非季節性資料的殘差是AR(2)或者季節性的殘差符合ARIMA(2,0,0)(1,0,0)m，這個模型能夠允許最多的自相關資訊，因此β的估計將在合理的範圍內。接著可以計算n_t，在嘗試幾個模型計算出n_t後，在進行整個模型的β估計。

整個模型建立的流程如下：

1. 選定解釋變數
2. 檢驗是否需要資料轉換->做Box-Cox 轉換
3. 檢驗是否定態 -> 非定態，差分
4. 配適複迴歸模型，並視資料的季節性，針對殘差配適AR(2)或是ARIMA(2,0,0)(1,0,0)m
5. 計算n_t，重複嘗試幾次找出最適合殘差的ARMA模型
6. 使用新的ARMA殘差模型，重新配適整個複迴歸模型。

7. 檢驗殘差

在按照這個模型建構流程後，所建構的模型符合我們一般的複迴歸模型假設。此時AIC為模型建立良好的判別標準。

四、判別標準

一般情況下，我們假設模型的殘差服從獨立常態分配，所以AIC可以表示為：

$$AIC = 2k - 2\ln(L) \quad AIC = 2k - n\ln(RSS/n)$$

其中，K是估計參數的數量，L是概似函數(Likelihood function)，n為樣本數。
在用AIC作為選模標準時，應選擇AIC最小的那一個，意即包含最少解釋變數，卻具有最好解釋力的模型。