Time Series Analysis

葉昱廷

2019 年 6 月 29 日

資料介紹

機動車輛及道路交通事故-月資料

1. 資料來源：中華民國統計資訊網-總體統計資料庫
   http://statdb.dgbas.gov.tw/pxweb/Dialog/statfile9L.asp

2. 資料時間：2000 年 1 月到 2019 年 2 月，共 230 個觀測值。

3. 資料地區：全台灣

4. 資料綜觀：以截圖(圖一)呈現

5. 重要欄位：肇事總件數、A1 類件數、A2 類件數

   (1) 肇事總件數：當月機動車輛交通事見總發生件數，亦即 A1 類件數及 A2 類件數之總和

   (2) A1 類件數：1999 年之前定義為「重大交通事故」之件數，其中「重大交通事故」係指現場死亡人數在 3 人以上，或死傷人數在 10 人以上，或受傷人數在 15 人以上者；2000 年後，其定義改為「造成當場或 24 小時內死亡之事故」。

   (3) A2 類件數：2000 年後，A2 類指造成受傷或超過 24 小時死亡之交通事故。

機動車輛及道路交通事故-月 依 期間, 種類 與 指標

	原始值														
	機動車輛數(輛) - 合計	機動車輛數(輛) - 汽車 (不含軍車)	機動車輛數(輛) - 機車	肇事總件數(件)	肇事率(件/萬輛)	A1類件數(件)	A1類死亡人數(人)	A1類受傷人數(人)	A2類件數(件)	A2類受傷人數(人)	A1類肇事件數按肇事原因(件) - 汽(機、慢)車駕駛人	A1類肇事件數按肇事原因(件) - 機件	A1類肇事件數按肇事原因(件) - 行人(或乘客)	A1類肇事件數按肇事原因(件) - 交通管制(設施)	A1類肇事件數按肇事原因(件) - 其他
2000M01	16,375,878	5,384,785	10,991,093	3,893	2.38	228	242	110	3,665	4,680	224	2	1	0	1
2000M02	16,423,704	5,395,403	11,028,301	3,086	1.88	198	218	102	2,888	3,745	194	1	2	1	0
2000M03	16,458,347	5,392,777	11,065,570	3,525	2.14	270	293	138	3,255	4,230	267	2	1	0	0
2000M04	16,532,431	5,406,194	11,126,237	3,753	2.28	244	259	139	3,509	4,543	234	4	6	0	0
2000M05	16,560,218	5,421,514	11,138,704	4,400	2.66	267	275	110	4,133	5,465	265	1	0	1	0
2000M06	16,618,046	5,442,371	11,175,675	4,480	2.70	300	311	152	4,180	5,436	293	2	5	0	0
2000M07	16,679,251	5,460,837	11,218,414	4,555	2.74	263	282	148	4,292	5,728	259	2	2	0	0
2000M08	16,771,059	5,512,973	11,258,086	4,691	2.80	267	286	124	4,424	5,804	261	2	3	1	0
2000M09	16,845,887	5,538,202	11,307,685	4,581	2.73	264	272	122	4,317	5,726	260	2	2	0	0
2000M10	16,919,426	5,563,793	11,355,633	5,313	3.15	296	310	128	5,017	6,703	290	3	2	1	0
2000M11	16,968,352	5,578,243	11,390,109	5,126	3.03	290	306	108	4,836	6,346	282	3	5	0	0
2000M12	17,022,689	5,599,517	11,423,172	5,549	3.26	320	334	160	5,229	6,948	312	2	6	0	0
2001M01	17,074,074	5,627,706	11,446,368	5,439	3.19	339	356	148	5,100	6,630	328	4	7	0	0
2001M02	17,121,351	5,643,186	11,478,165	4,573	2.67	253	274	102	4,320	5,553	247	2	4	0	0
2001M03	17,171,744	5,658,980	11,512,764	5,513	3.22	271	290	133	5,242	6,776	265	2	4	0	0
2001M04	17,189,428	5,662,231	11,527,197	5,261	3.06	252	278	161	5,009	6,479	244	2	6	0	0
2001M05	17,228,532	5,678,132	11,550,400	5,497	3.19	274	293	135	5,223	6,712	263	3	8	0	0
2001M06	17,269,620	5,692,999	11,576,621	5,124	2.97	239	251	129	4,885	6,399	234	1	4	0	0
2001M07	17,312,924	5,705,860	11,607,064	5,318	3.08	242	258	97	5,076	6,661	234	4	3	0	1

圖一、機動車輛及道路交通事故-月 資料截圖

資料處理流程

一、探索性資料分析(Exploratory Data Analysis):

　　一、Ts-plot：Overview the whole data and check trend and outliers

　　二、Box-plot：Check seasonal effect and its variance

二、排除趨勢效應(De-Trend) :

　　一、Determinist Trend ：Linear model

　　二、Stochastic Trend ：Difference

三、排除季節效應(De-Seasonality):

　　一、Determinist Seasonality：Seasonal model with dummy variable

　　二、Stochastic Seasonality：Difference with lag

四、配適模型(Modeling)

　　一、ARMA

　　二、SARIMA

　　三、Regression with SARIMA

五、模型診斷(Diagnose)

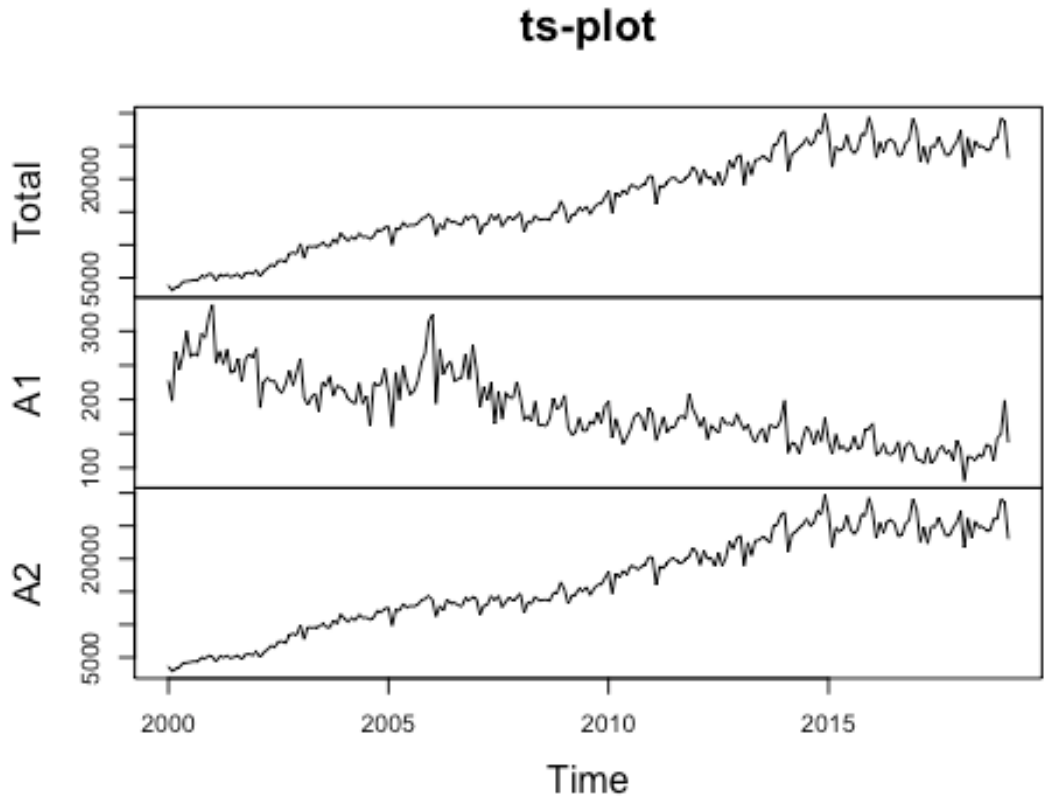
　　一、Residual Analysis

　　二、Outliers、White Noise

六、模型預測(Prediction)

一、 探索性資料分析

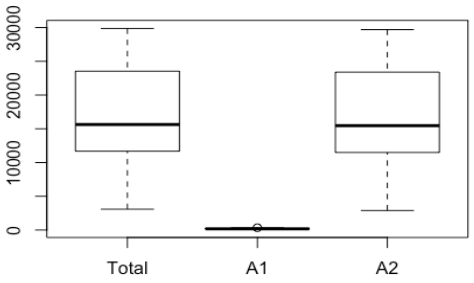
1. Ts-plot:
   根據總肇事件數、A1 類件數、A2 類件數的 Ts-plot (圖二)所示，總肇事件數的趨勢與 A2 類件數趨勢幾乎相符合並呈現隨時間增長的趨勢，反觀 A1 類件數則隨著時間增加而傾向遞減。



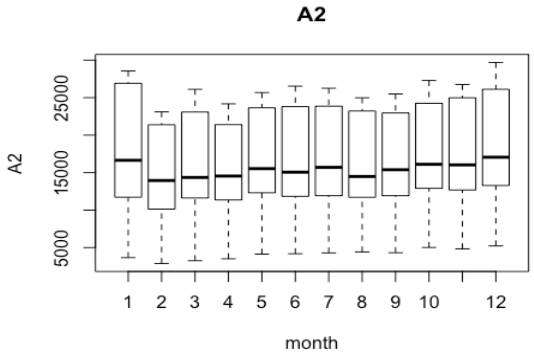
圖二、總肇事件數(Total)、A1 類件數(A1)、A2 類件數(A2)的 Ts-plot

2. Box-plot：
   觀察總肇事件數、A1 類件數、A2 類件數的 Ts-plot (圖三)，則可以發現到，A1 類的件數比起 A2 類少非常多，A2 類的件數分佈則幾乎等於總肇事件數，考慮到 A1 類代表當場或 24 小時內死亡的案件，條件較為嚴苛，可以合理解釋機動車總肇事件數大部分由條件較寬鬆的 A2 類組成。因此本篇研究對象鎖定在 A2 類的案件數，其結果也能解釋大部分總肇事案件數。 接著觀察 A2 類案件於每個月的件數分布(圖四)，以中位數的角度，一月、十二月有稍微偏高的傾向、二月份則有偏低的傾向；以四分位距(IQR)的角度來看，一月的 IQR 較寬、二月及四月的 IQR 較窄以外，其

餘月分的 IQR 長度差不多；整體來看，除了一月、二月、四月、十二月以外，其餘月份的分布情形看似差異不大。



圖三、總肇事件數(Total)、A1 類件數(A1)、A2 類件數(A2)的 Box-plot



圖四、A2 類案件於每個月的件數分布

二、 Determinist Method

本節將先介紹以 Determinist Method 處理時間數列的的趨勢(Trend)以及季節性(Seasonality)，後續再以 Stochastic Method 作為比較。

1. 趨勢效應(Determinist Trend)

(1) 線性迴歸模型：

由於 A2 案件數量呈現隨時間增長的趨勢，並且趨勢接近線性，這代表著 A2 案件數量隨著年份逐漸增長，如果參考原始資料，不難發現「機動車輛數」也隨著時間逐漸增加，越多的機動車數量便可能導致越多的交通事故發生，這部分解釋了趨勢效應的存在，因此配適該 A2 時間數列模型之時，可以先考慮趨勢帶來的影響。

由於該趨勢接近線性，以 Determinist Trend 的角度，可以用線性迴歸(Linear Regression)配適該趨勢情形，令 A2 案件數量(data_ts)作為被解釋變數，時間(Time)作為解釋變數，為求不偏估計將截距項納入模型，配適結果(報表一)呈現截距項及時間均有顯著影響(p-value 均小於 0,01)。

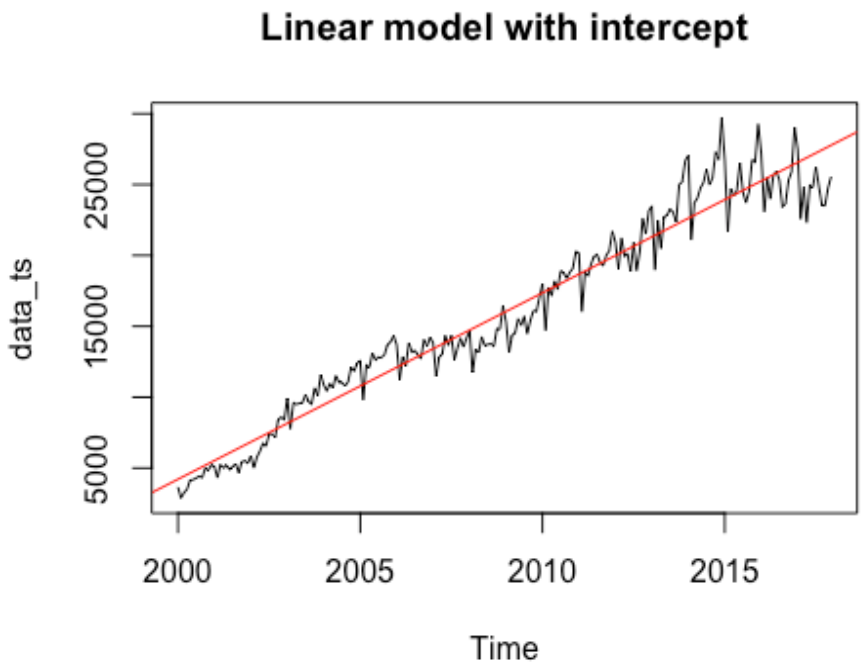
$$\text{趨勢模型：data\_ts} = (-2.626 \times 10^6) + (1.315 \times 10^3)Time$$

R-square=0.948，代表該線性性模型解釋變異比例可以達到 9.48 成，實際配適圖形(圖三)可以發現迴歸線確實能反映出資料的增長趨勢，但還不能顯示出資料的鋸齒狀的震盪情形，可以合理懷疑，除了趨勢(Trend)以外，還可能有季節性(Seasonality)的存在，因此繼續對排除趨勢(De-trend)後的殘差進行殘差分析(Residual Analysis)。

Linear Model with intercept

```
## Call:
## lm(formula = data_ts ~ Time)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4578.8  -989.7   -69.1   971.6  5865.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.626e+06  4.229e+04  -62.10   <2e-16 ***
## Time          1.315e+03  2.105e+01   62.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1608 on 214 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9478
## F-statistic:  3904 on 1 and 214 DF,  p-value: < 2.2e-16
```

報表一、Determinist Trend 線性迴歸配適結果



圖三、Determinist Trend 線性迴歸配適結果，迴歸線(紅)及原始資料(黑)

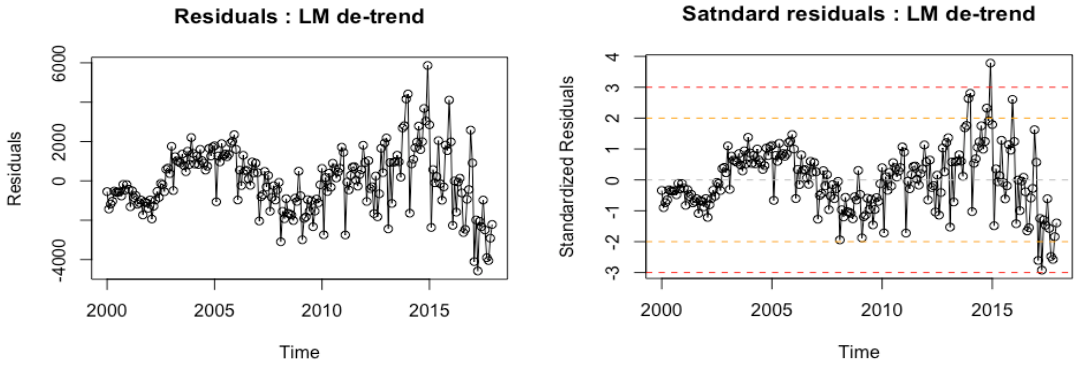
(2)殘差分析(Residual Analysis)：
取得線性模型的殘差之後，對殘差值(Residuals)及時間(Time)做散佈圖(圖四)，可以發現隨著時間增加，殘差的變異數有越來越大的趨勢，也顯示了殘差並不符合迴歸分析中變異數同質性(Homogeneity of variance)假設。

也因為上述原因，在標準化殘差分佈圖上，可以發現資料在 2013 年之後越來越多資料超過 2 個標準差，甚至在 2015 年超過 3 個標準差，越後期的資料離群值則越來越多。

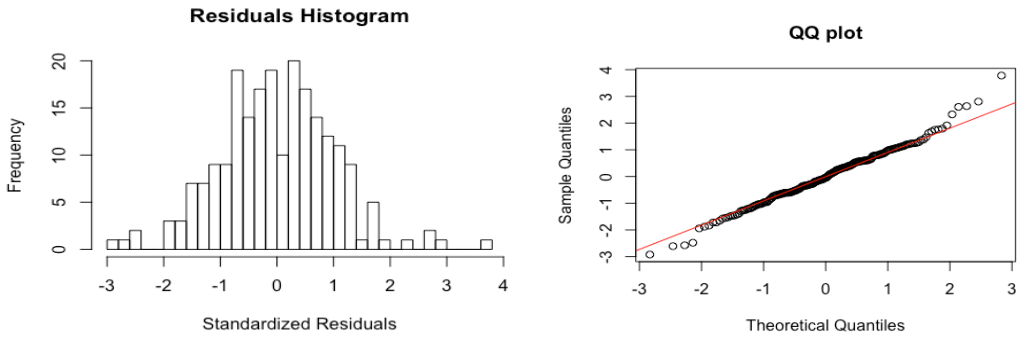
另外，殘差的直方圖(圖五)呈現鐘狀，QQ-plot 與 QQ-line 在中段處幾乎貼合，只有兩端的資料稍微偏離，並再以 K-S Test 檢定其是否符合常態分配假設，得 p-value=0.955，不拒絕常態分配(Normality)的假設。

再繪製殘差的 ACF(圖六)，可以發現從第一期(Lag1)殘差就有近 0.6 的自相關係數，並呈現拖尾(Tail-off)的情形，這代表當期的 A2 案件數量與前數期的 A2 案件數量是有相關的，也因此不符合迴歸分析中獨立性(Independent)的假設。

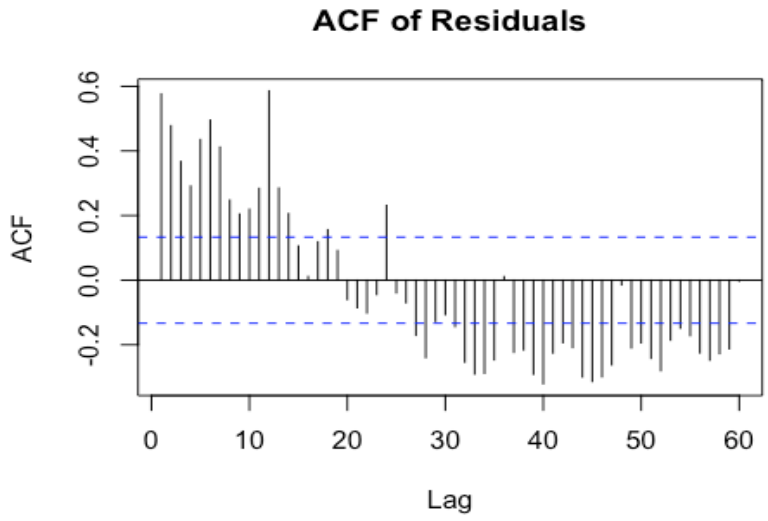
綜合以上，該筆資料不符合迴歸分析中的變異數同質性(Homogeneity of variance)假設、獨立性(Independent)的假設，雖然不拒絕常態分配(Normality)的假設，但從殘差分析的結果來看，證明了該筆資料不能完全以迴歸分析的結果來解釋，而是必須作為時間數列(Time Series)做更深入的探討。



圖四、線性迴歸模型下的殘差散佈圖(左)與標準化殘差散佈圖(右)



圖五、線性迴歸模型下的殘差直方圖(左)與 QQ-plot (右,紅線為 QQ-line)



圖六、線性迴歸模型下的殘差 ACF

2. 季節效應(Determinist-Seasonality)

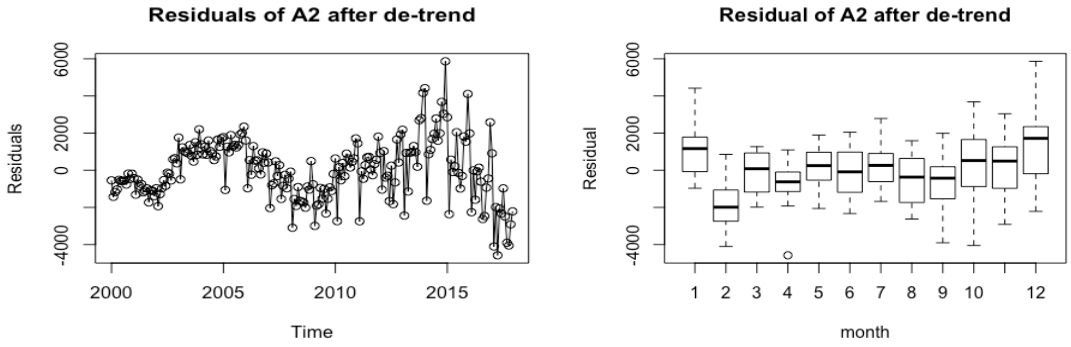
本節承接趨勢效應(Determinist Trend)的結果，將排除趨勢效應後的殘差(Residuals)繼續排除季節性(Seasonality)帶來的影響。

(1) 季節性模型

考慮到資料有週期性的震盪(圖七)，並且在二月份的數值相對偏低，而且每個月的分佈均不完全相同，相較於排除趨勢之前的分佈(圖四)已有落差，因此可以將季節性因子(月份)納入考量，本節採用將 12 個月份作為 12 個虛擬變數(Dummy Variables)來配適季節性模型，並非採用 11 個虛擬變數的原因是該模型不包含截距項(intercept)，包含截距項並採用 11 個虛擬變數的模型在配適值並沒有任何差異(補充)，因此本節採用前者建模。

季節性模型：$season\_data\_ts = \sum_{i=1}^{12} \gamma_i m_i, i = 1,2,\ldots,12$

其中γ_i的值如報表二所示



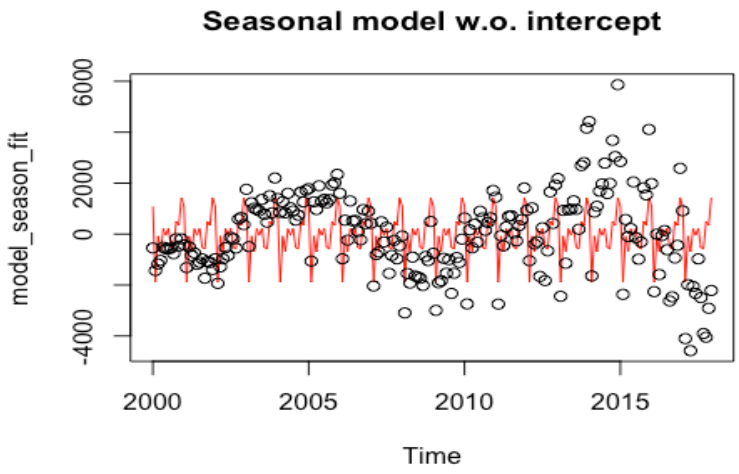
圖七、排除趨勢效應(De-Trend)後的殘差散佈圖(左)及 Box-plot(右)

Seasonal Model w.o. intercept

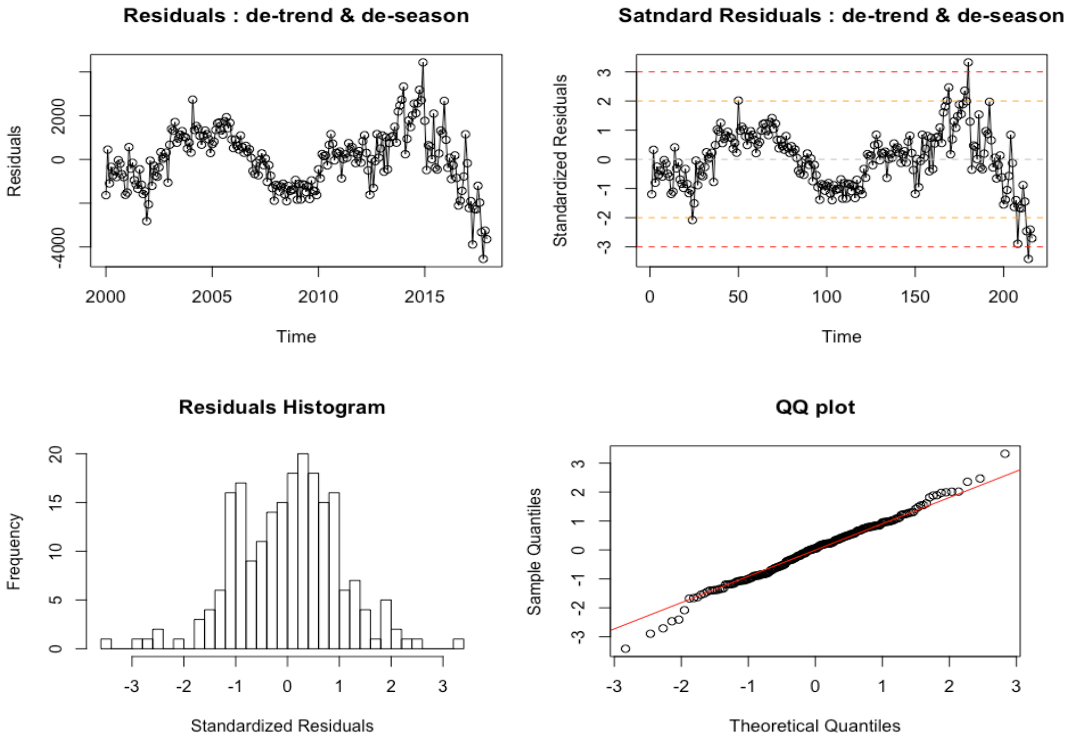
```
## Call:
## lm(formula = season_data_ts ~ mon - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4548.1  -976.3    73.0   927.8  4433.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## monJanuary    1086.50     331.48   3.278  0.00123 **
## monFebruary  -1879.83     331.48  -5.671 4.81e-08 ***
## monMarch       -78.44     331.48  -0.237  0.81318
## monApril      -691.49     331.48  -2.086  0.03821 *
## monMay         206.12     331.48   0.622  0.53476
## monJune        -46.82     331.48  -0.141  0.88780
## monJuly        232.18     331.48   0.700  0.48446
## monAugust     -520.93     331.48  -1.572  0.11761
## monSeptember  -576.88     331.48  -1.740  0.08331 .
## monOctober     498.68     331.48   1.504  0.13402
## monNovember    339.51     331.48   1.024  0.30693
## monDecember   1431.41     331.48   4.318 2.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1406 on 204 degrees of freedom
## Multiple R-squared:  0.2706, Adjusted R-squared:  0.2277
## F-statistic: 6.306 on 12 and 204 DF,  p-value: 1.87e-09
```

報表二、Determinist Seasonality 季節性模型配適結果

季節性模型的配適結果(圖八)並非完全貼合原始資料，然而季節性模型的目的在於考量了不同月份的平均值不全相同，在於解釋季節性因子帶來的影響，由於了解資料帶有自相關，因此剩餘的殘差可再由 ARMA 模型做最後配適，此時的殘差(圖九)的變異數仍在後期較大，但 K-S 檢定的 p-value=0.9165，仍不拒絕常態分配的假設。



圖八、Determinist Seasonality 季節性模型配適結果，配適線(紅)及原始資料(黑)



圖九、排除季節性效應(De-Seasonality)後的殘差分析圖

3. 配適模型(Modeling)

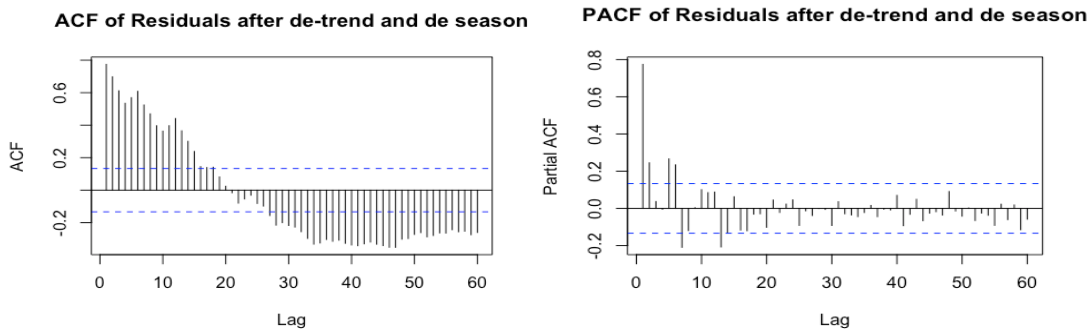
在排除趨勢(De-trend)跟季節性(De-seasonality)之後，由於先前發現了資料帶有自相關性，因此可再以 ARMA 模型或 Seasonal ARMA(SARMA)進行配適來解釋其自相關情形。

ARMA Model：$(1-\Sigma_i^{\boldsymbol{p}}\phi_i B_i)Y_t=(1-\Sigma_j^{\boldsymbol{q}}\theta_j B^j)e_t$

SARMA Model：$(1-\Sigma_i^{\boldsymbol{p}}\phi_i B^i)(1-\Sigma_j^{\boldsymbol{P}}\Phi_j B^{js})Y_t=(1-\Sigma_k^{\boldsymbol{q}}\theta_k B^k)(1-\Sigma_l^{\boldsymbol{Q}}\theta_l B^{ls})e_t$

(1) ACF 與 PACF

要決定 ARMA(p, q)模型中的參數 p 跟 q，可以先觀察資料的 ACF 跟 PACF 圖形(圖十)，拖尾(tail-off)的 ACF 跟截尾(cut-off)的 PACF 暗示了可能是偏向 AR 模型，ACF 從 lag 1 將近 0.8 的自相關係數隨 lag 期數增加逐漸下降，但在 lag 6 或 6 的倍數時會再有相對上升的凸起，因此模型配適上先考慮了 AR(1)以及 SARMA(p=1,q=0)(P=1,Q=0)[s=6]兩個模型。



圖十、排除趨勢效應(De-Trend)及季節性效應(De-Seasonality)後的 ACF 與 PACF

(2) AR(1)

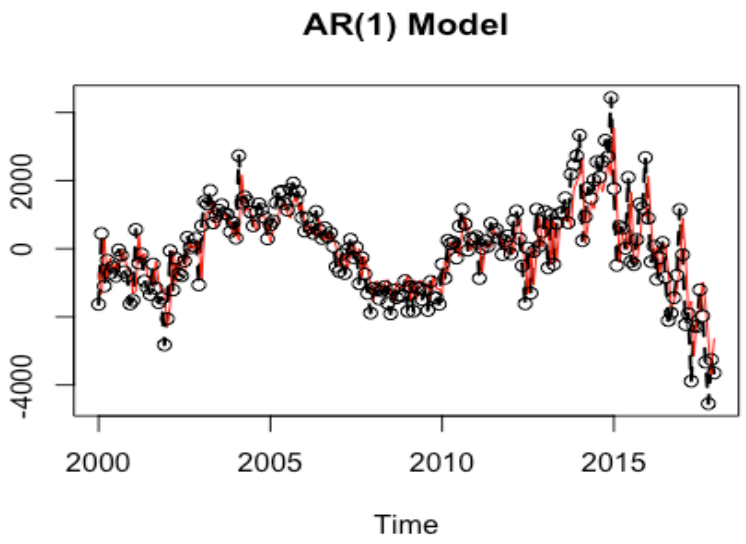
AR(1)Model：$(1-\phi_1 B_1)Y_t=e_t \implies Y_t=\phi_1 Y_{t-1}+e_t$

AR(1)模型的特點是當期的觀察值(Y_t)會和前一期的觀察值(Y_{t-1})有關，並且ϕ_1係數估計值為 0.8025(報表三)，配適結果(圖十一)可看出紅色配適線幾乎貼近原始資料，再觀察殘差分析(圖十二)發現最終殘差在前期跟後期離群值(Outlier)較多，變異數也較大，但 K-S 檢定 p-value=0.8086 仍不拒絕常態分配的假設。

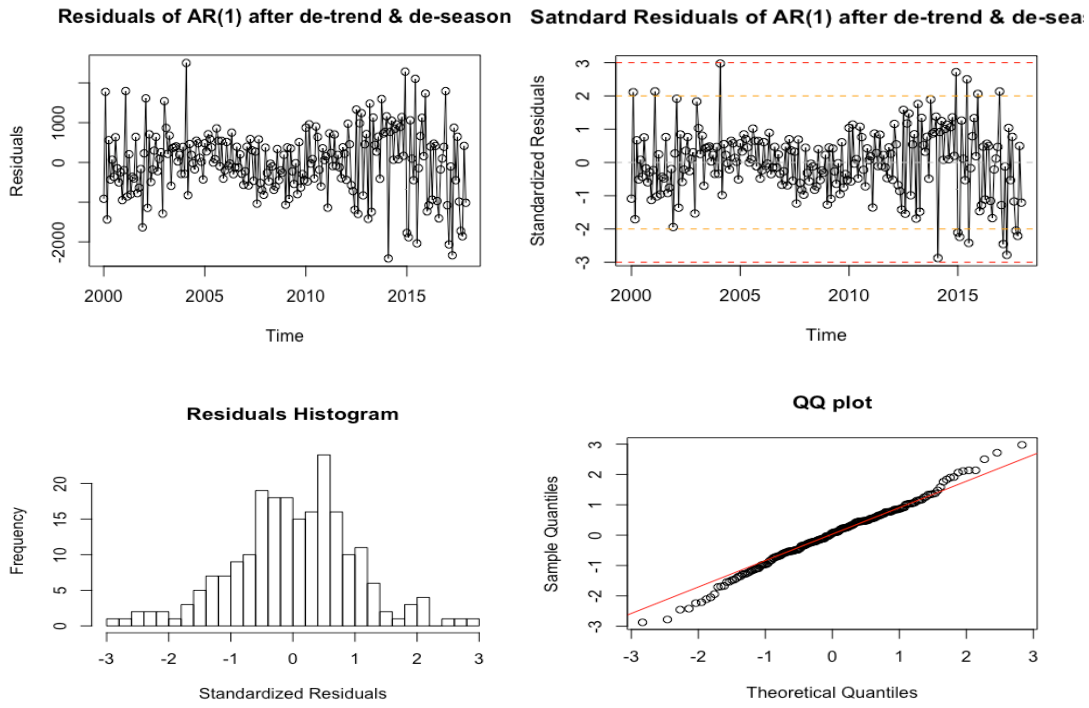
AR(1) Model

```
## Series: arma_data_ts
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
##          ar1      mean
##       0.8025  -96.8391
## s.e.  0.0420  284.3951
##
## sigma^2 estimated as 707878:  log likelihood=-1760.77
## AIC=3527.53   AICc=3527.64   BIC=3537.66
##
## Training set error measures:
##                    ME     RMSE      MAE    MPE     MAPE      MASE
## Training set 8.801504 837.4505 649.3579 40.606 230.7969 0.6335582
##                    ACF1
## Training set -0.2102008
```

報表三、AR(1)模型配適結果



圖十一、AR(1)模型配適結果，配適線(紅)及原始資料(黑)



圖十二、AR(1)模型的殘差分析圖

(3) SARMA(p=1,q=0)(P=1,Q=0)[s=6]

SARMA Model：$(1-\phi_1 B^1)(1-\Phi_1 B^6)Y_t = e_t$

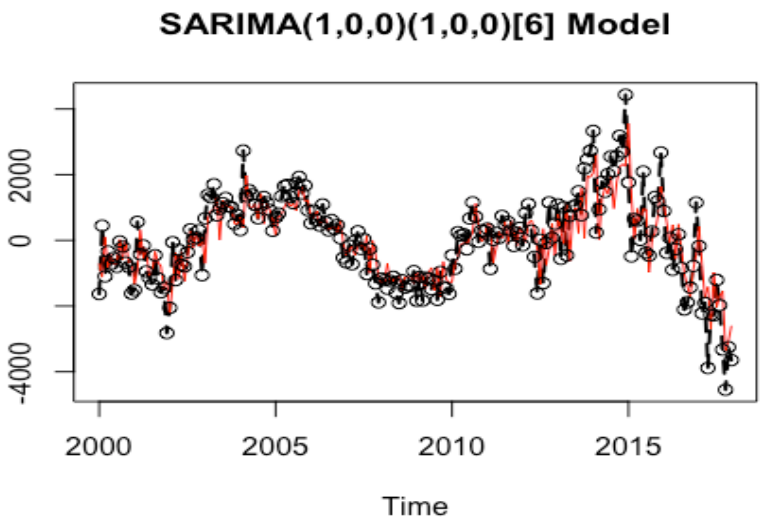
$\Longrightarrow Y_t = \phi_1 Y_{t-1} + \Phi_1 Y_{t-6} + \phi_1 \Phi_1 Y_{t-7} + e_t$

該模型的特點是當期的觀察值(Y_t)會和前一期、前六期、甚至前七期的觀察值(Y_{t-1}、Y_{t-6}、Y_{t-7})有關，並且ϕ_1和Φ_1係數估計值分別 0.7119 和 0.3859(報表四)，配適結果(圖十三)可看出紅色配適線也是幾乎貼近原始資料，再觀察殘差分析(圖十四)，和 AR(1)雷同，最終殘差在前期跟後期離群值(Outlier)較多，變異數也較大，K-S 檢定 p-value=0.5913，仍不拒絕常態分配的假設。

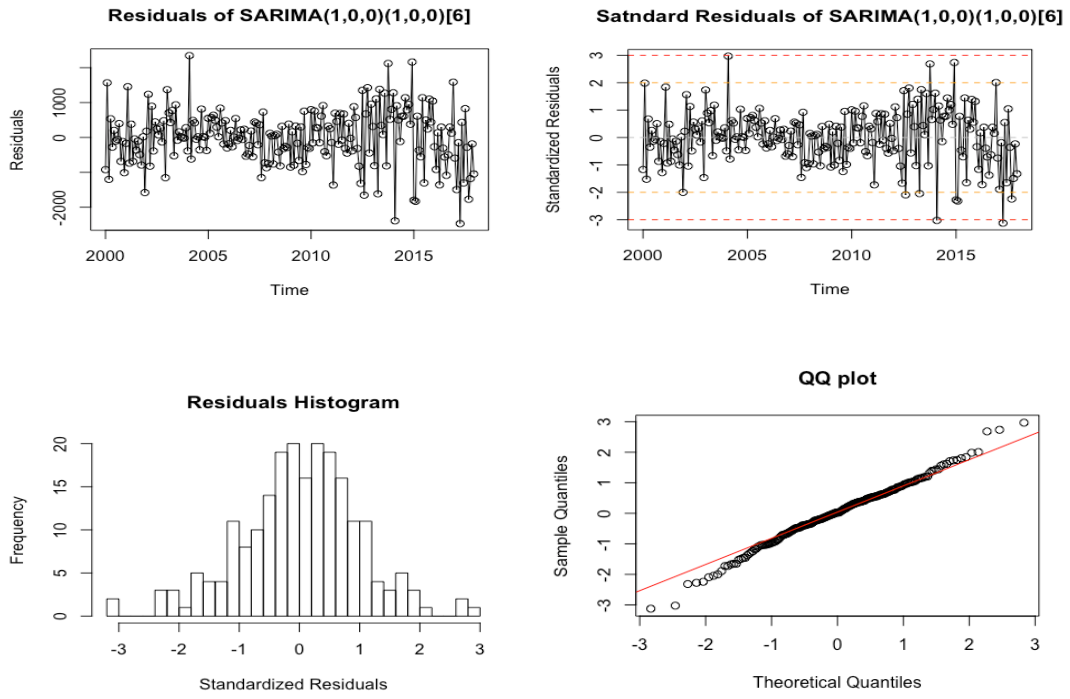
SARMA(1,0)(1,0)[6] MODEL

```
## Series: arma_data_ts
## ARIMA(1,0,0)(1,0,0)[6] with non-zero mean
##
## Coefficients:
##          ar1    sar1       mean
##       0.7119  0.3859  -132.1882
## s.e.  0.0540  0.0729   294.3333
##
## sigma^2 estimated as 625227:  log likelihood=-1747.22
## AIC=3502.45   AICc=3502.63   BIC=3515.95
##
## Training set error measures:
##                    ME     RMSE      MAE      MPE     MAPE      MASE
## Training set 9.343143 785.2024 606.2786 52.1924 211.0902 0.5915271
##                    ACF1
## Training set -0.203691
```

報表四、**SARMA(1,0)(1,0)[6]**模型配適結果



圖十三、AR(1)模型配適結果，配適線(紅)及原始資料(黑)



圖十四、AR(1)模型的殘差分析圖

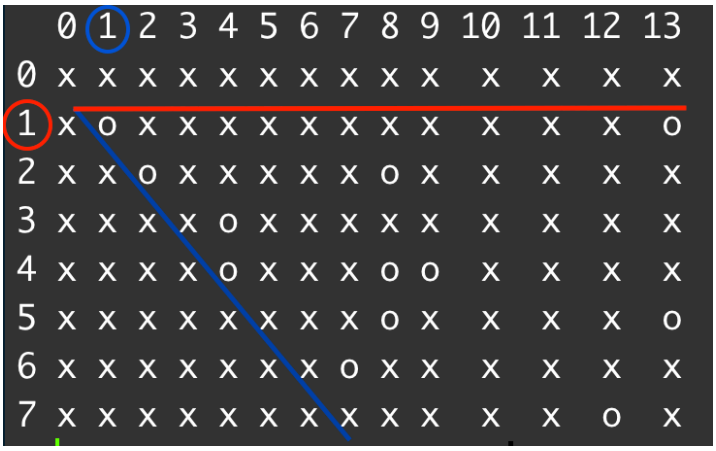
(3) EACF Method

蔡瑞雄於 1984 年提出推廣的自相關係數矩陣(EACF)，用來決定 ARMA 過程的 non-seasonal parameters (p, q)，其特徵是由符號「o」組成的三角形左上角頂點處，即為較好的(p, q)階數，然而並非每個資料都有明確的三角形(報表五)，R 軟體預設只要該階數的 EACF 的絕對值小於 0.1 就會顯示符號「o」，因此可以將該門檻值逐漸降低，在 EACF 的絕對值小於 0.01 的情況下(報表六)，試圖保留所有的符號「o」所需要的最精簡的三角形將是以(1, 1)為頂點，ARMA(1, 1)將作為候選的模型之一。

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	o	o	x	o	x	o	o	o	x	o	x	o	o
2	x	o	o	x	o	x	o	o	o	x	o	x	o	o
3	o	o	o	x	o	x	o	o	o	o	o	x	o	o
4	o	x	x	x	o	x	o	o	o	o	o	x	o	o
5	x	x	x	x	x	o	x	o	o	o	o	x	o	o
6	x	x	x	o	x	x	o	o	o	o	o	o	o	o
7	x	x	x	o	o	x	o	o	o	o	o	o	o	o

報表五、EACF 檢定結果，ACF 絕對值門檻：0.1



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	o	x	x	x	x	x	x	x	x	x	x	x	o
2	x	x	o	x	x	x	x	x	o	x	x	x	x	x
3	x	x	x	x	o	x	x	x	x	x	x	x	x	x
4	x	x	x	x	o	x	x	x	o	o	x	x	x	x
5	x	x	x	x	x	x	x	x	o	x	x	x	x	o
6	x	x	x	x	x	x	x	o	x	x	x	x	x	x
7	x	x	x	x	x	x	x	x	x	x	x	x	o	x

報表六、EACF 檢定結果，ACF 絕對值門檻：0.01

(3) auto.arima

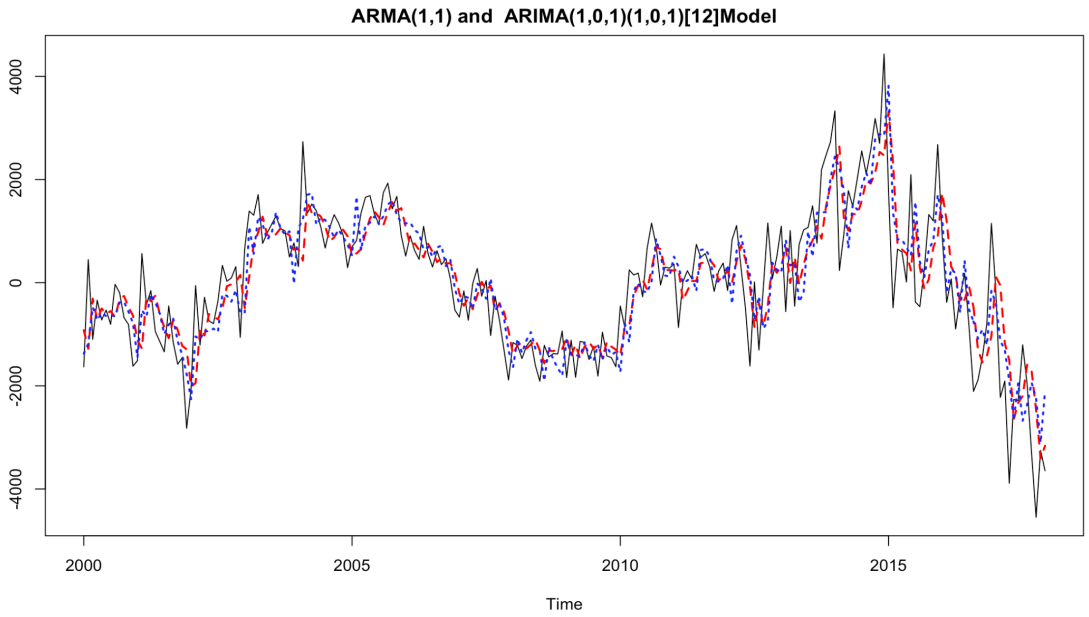
R 語言軟體的 forecast 套件中包含 auto.arima 函數，其作法會根據設定的起始參數(start.p, start,q, start.P, start.Q)以及參數最大值(max.p, max,q, max.P,max.Q)，由於 Determinist Method 透過線性迴歸與季節性模型排除趨勢跟季節性效應，因此設定參數時令 d 和 D 固定為 0。執行函數後，便會開始在設定的參數範圍內逐一建模並約略(approximately)估計每一個模型的 AIC 值，因此過程中會有許多候選模型(報表七)，鎖定在一定數量的模型之後，再精準(without approximations)計算 AIC 並找到局部(locally)最佳的參數後，此方法不能保證全域(globally)最小的 AIC，但在效率上卻是不錯的選擇，該方法篩選的局部最佳模型為 SARIMA(1,0,1)(1,0,1)[12] with zero mean(後續簡稱 SARMA(1,1)(1,1)[12])，同樣可以將此納入候選模型。

(4) arma(1, 1)與 SARMA(1,1)(1,1)[12]模型

透過 EACF 和 auto.arima 得到的候選模型分別為 arma(1, 1)與 SARMA(1,1)(1,1)[12]，模型配飾結果(圖十五)兩者都貼近原始資料的震盪情形，在離群值的預測上也都能維持趨勢並保守估計，其中 ARMA(1,1)的 AIC=3510.22，相較於 SARMA(1,1)(1,1)[12]的 AIC=3471.867 來得高一點但不會相差太多。

```
##  Fitting models using approximations to speed things up...
##
##  ARIMA(2,0,2)(1,0,1)[12] with non-zero mean : 3471.228
##  ARIMA(0,0,0)           with non-zero mean : 3736.151
##  ARIMA(1,0,0)(1,0,0)[12] with non-zero mean : 3500.003
##  ARIMA(0,0,1)(0,0,1)[12] with non-zero mean : 3593.304
##  ARIMA(0,0,0)           with zero mean     : 3734.114
##  ARIMA(2,0,2)(0,0,1)[12] with non-zero mean : 3482.919
##  ARIMA(2,0,2)(1,0,0)[12] with non-zero mean : 3477.057
##  ARIMA(2,0,2)(2,0,1)[12] with non-zero mean : 3481.449
##  ARIMA(2,0,2)(1,0,2)[12] with non-zero mean : 3473.296
##  ARIMA(2,0,2)           with non-zero mean : 3506.728
##              ……(中間過長省略)……
##  ARIMA(1,0,1)           with zero mean     : 3508.494
##  ARIMA(1,0,1)(0,0,2)[12] with zero mean     : 3476.181
##  ARIMA(1,0,1)(2,0,0)[12] with zero mean     : 3475.781
##  ARIMA(1,0,1)(2,0,2)[12] with zero mean     : 3478.018
##  ARIMA(0,0,1)(1,0,1)[12] with zero mean     : 3592.363
##  ARIMA(1,0,0)(1,0,1)[12] with zero mean     : 3491.253
##  ARIMA(2,0,1)(1,0,1)[12] with zero mean     : 3467.794
##  ARIMA(1,0,2)(1,0,1)[12] with zero mean     : 3468.353
##  ARIMA(0,0,0)(1,0,1)[12] with zero mean     : 3664.864
##  ARIMA(0,0,2)(1,0,1)[12] with zero mean     : 3543.6
##  ARIMA(2,0,0)(1,0,1)[12] with zero mean     : 3467.91
##  ARIMA(2,0,2)(1,0,1)[12] with zero mean     : 3469.932
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(1,0,1)(1,0,1)[12] with zero mean     : 3471.867
##
##  Best model: ARIMA(1,0,1)(1,0,1)[12] with zero mean
```

報表七、auto.arima 報表(trace = T 可追蹤建模過程)



圖十五、ARMA(1, 1)與 SARMA(1,1)(1,1)[12]模型配適圖

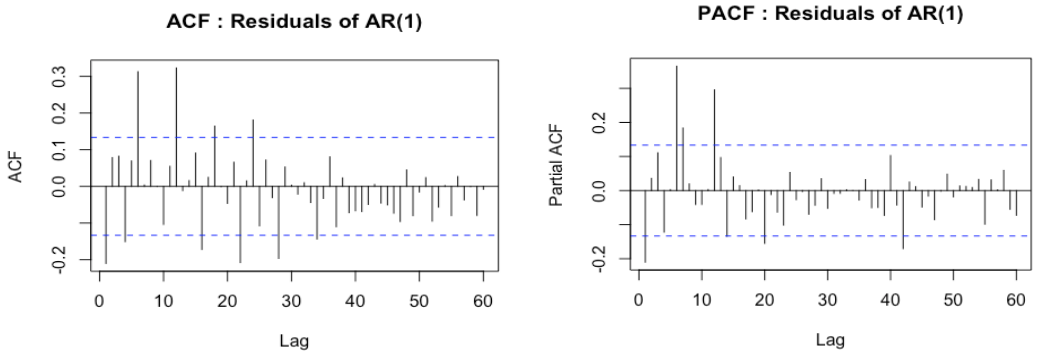
4. 模型診斷(Diagnose)

要區分模型優劣、診斷模型的好壞，可以進一步分析最終的殘差，以(S)ARIMA 的模型來說，最終的殘差的分配理論上必須是白噪音(White Noise)，亦即 $e_i \sim WN(0, \sigma_e^2),\ i = 1, 2, \ldots, t$ ，也就是最終殘差在相同時間間距的分佈上變異程度應相同；並且應符合獨立性，亦即$e_i \sqcup e_i\ for\ i \neq j$，這意味著殘差的 ACF 不應該有過大的值。

上述候選的模型共有 AR(1)、SARMA(1,0)(1,0)[6]、ARMA(1,1)、SARMA(1,1)(1,1)[12]

(1) AR(1)

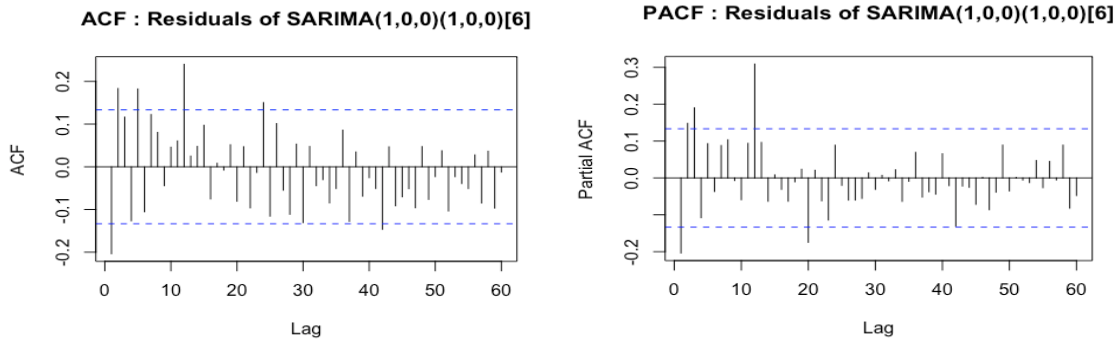
該模型的最終殘差的 ACF 與 PACF(圖十六)在 lag 1、lag 6 以及 lag 6 的倍數仍有一定的自相關係數，代表當期的觀察值與前六期的的觀察值有一定的相關程度，並且 AR(1)並無法考慮到這個情形，導致殘差不服從白噪音 White Noise，診斷結果 AR(1)並不適合配適該資料。



圖十六、AR(1)模型的最終殘差 ACF 與 PACF

(2) SARMA(1,0)(1,0)[6]

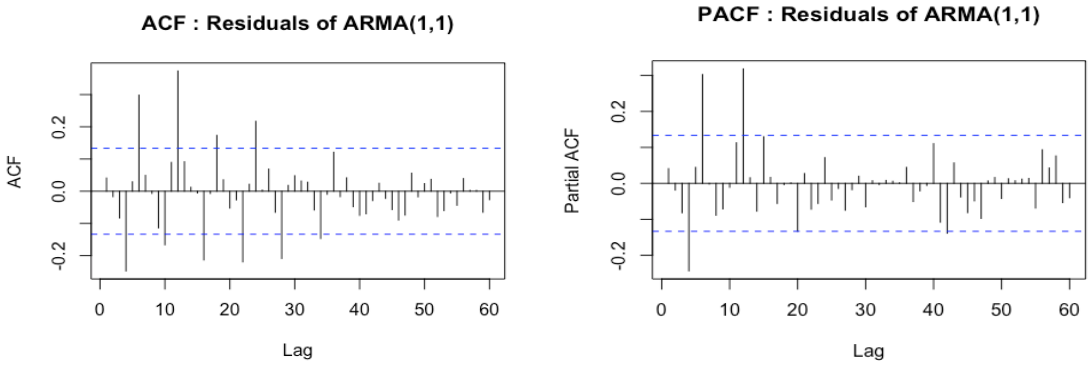
該模型的最終殘差的 ACF 與 PACF(圖十七)在 lag 1 以及 lag 12 的倍數仍有一定的自相關係數，代表當期的觀察值與前十二期的的觀察值有一定的相關程度，並且 SARMA(1,0)(1,0)[6]並無法考慮到這個情形，導致殘差不服從白噪音 White Noise，診斷結果 SARMA(1,0)(1,0)[6]並不適合配適該資料。



圖十七、SARMA(1,0)(1,0)[6]模型的最終殘差 ACF 與 PACF

(3) ARMA(1, 1)

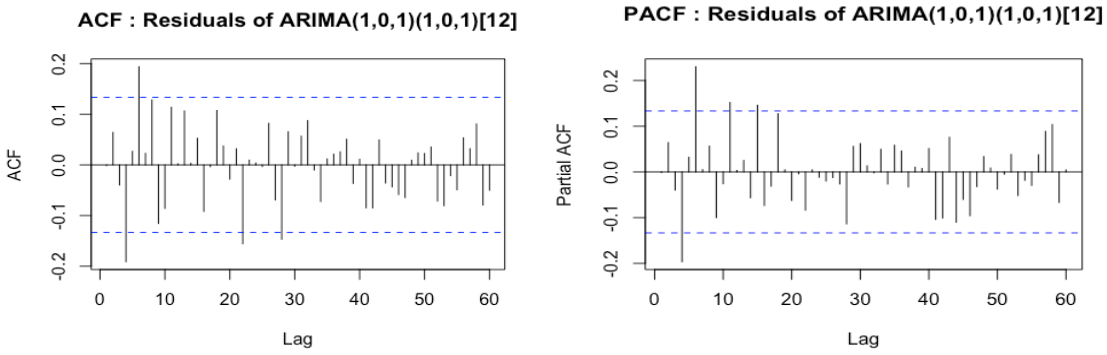
該模型的最終殘差的 ACF 與 PACF(圖十七)相較於 AR(1)跟 SARMA(1,0)(1,0)[6]在 lag1 的自相關係數降低了許多，這也意味著 MA(1)的加入可能有一定的必要性，然而在 lag 6 的倍數仍呈現不可忽視的自相關係數，代表當期的觀察值與前六期的的觀察值有一定的相關程度，並且 ARMA(1, 1)並無法考慮到這個情形，導致殘差不服從白噪音 White Noise，診斷結果 ARMA(1, 1)並不適合配適該資料。



圖十七、SARMA(1,0)(1,0)[6]模型的最終殘差 ACF 與 PACF

(4) SARMA(1,1)(1,1)[12]

該模型同時考量了 Non-seasonal 及 Seasonal 的 AR 及 MA 參數，從上述情況分析下來，是考量上最完整同時也是參數最多的一個模型，最終殘差的 ACF(圖十八)只有在 lag 6 仍超出兩個標準誤差，PACF(圖十八)則是在 lag 6、lag 12 仍有較高的偏字相關係數，代表當期的觀察值與前六期的的觀察值還存在著一定的相關程度，並且 SARMA(1,1)(1,1)[12]仍無法有效排除這個相關情形，雖然殘差已經大致服從白噪音 White Noise，但明顯的第六期自相關無法完全忽視，診斷結果該模型仍不夠適合配適該資料。



圖十八、SARMA(1,1)(1,1)[12]模型的最終殘差 ACF 與 PACF

三、Stochastic Method

從 Determinist Method 的結果發現在 lag 6 跟 lag 12 的自相關性光用季節性模型(Seasonal Model)處理可能不夠完善，導致最終殘差仍無法排除自相關性，因此本節將改用 Stochastic Method 處理時間數列的的趨勢(Trend)以及季節性(Seasonality)，並和 Determinist Method 作為比較。

1. 趨勢效應(Stochastic Trend)

Determinist Method 排除趨勢效應的方式是以線性迴歸模型配飾趨勢線，Stochastic Method 則是改以差分(Difference)的方式處理，經驗上一階差分後的資料就能使不平穩(non-stationary)的趨勢資料轉成平穩(stationary)的時間數列，少數會做到二階差分。

(1) Dicky Fuller Test

該檢定目的是為了確定資料否需進行 lag k 期的一階差分才能達到平穩(stationary)，該檢定假設該時間數列模型為：

$$Y_t = \alpha Y_{t-k} + e_t \implies Y_t - Y_{t-k} = (\alpha - 1)Y_{t-k} + e_t$$

亦即將資料進行 k 階差分後，若α等於 1，差分後資料會屬於白噪音(White Noise)，便可達到平穩(stationary)，因此該檢定相當於下列假設：

$$\begin{cases} \mathrm{H}_0: \alpha = 1 \\ \mathrm{H}_1: \alpha \neq 1 \end{cases}, while\ Y_t - Y_{t-k} = (\alpha - 1)Y_{t-k} + e_t,\ e_t \sim WN$$

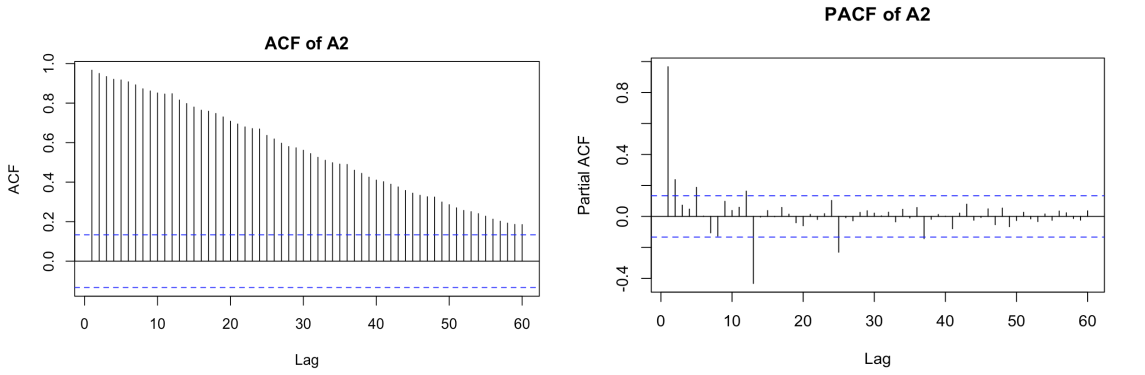
從 Determinist Method 的經驗可得，在 lag 1、lag 6、lag12 自相關性較高，因此針對 k=1, 6, 12 進行 Dicky Fuller Test(報表八)，檢定結果進行 lag 6 期或 lag 12 期的差分便能達到平穩；lag 1 期的 p-value=0.01 達到顯著，代表一階差分後仍不屬於白噪音。

```
##  Augmented Dickey-Fuller Test
##
## Dickey-Fuller = -5.156, Lag order = 1, p-value = 0.01
## alternative hypothesis: stationary
##  Augmented Dickey-Fuller Test
##
## Dickey-Fuller = -1.5549, Lag order = 6, p-value = 0.7627
## alternative hypothesis: stationary
##  Augmented Dickey-Fuller Test
##
## Dickey-Fuller = -0.56348, Lag order = 12, p-value = 0.9782
## alternative hypothesis: stationary
```

報表八、Dicky Fuller Test (lag = 1, 6, 12)

(2) ACF 與 PACF

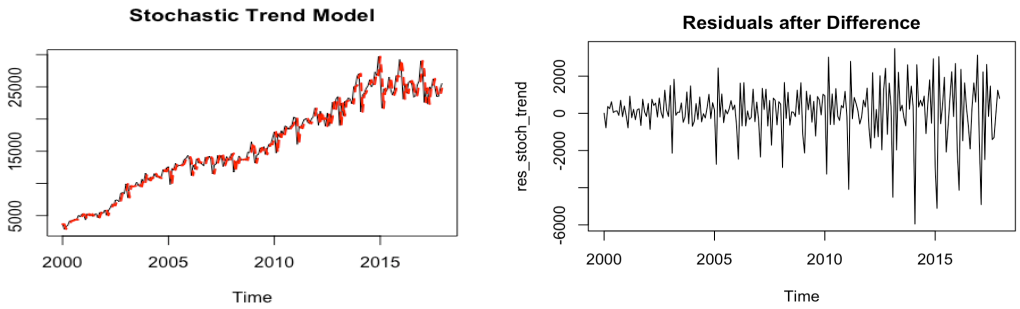
上述檢定方式是金融領域學者常用的技巧，然而統計上常以 ACF 與 PACF 做綜合判斷，從 A2 案件數量的原始資料 ACF 與 PACF(圖十九)能發現明顯拖尾(tail-off)的 ACF 跟在 lag 1 截尾(cut-off)的 PACF，這是 AR(1)的特徵，再加上 lag 1 的 ACF 係數將近 1.0，代表 AR(1)模型$Y_t = \phi Y_{t-1} + e_t$ 中的ϕ估計值幾乎等於 1.0（AR(1)模型的第一期自相關係數剛好等於係數ϕ），並發現每當 lag 12 或 12 的倍數時，ACF 會有較明顯的小突起，PACF 則在 lag 12 有明顯的偏自相關係數，這都代表著該資料適合對 lag1 做一階差分，是否再對 lag 12 做一階差分可以後續再考量。



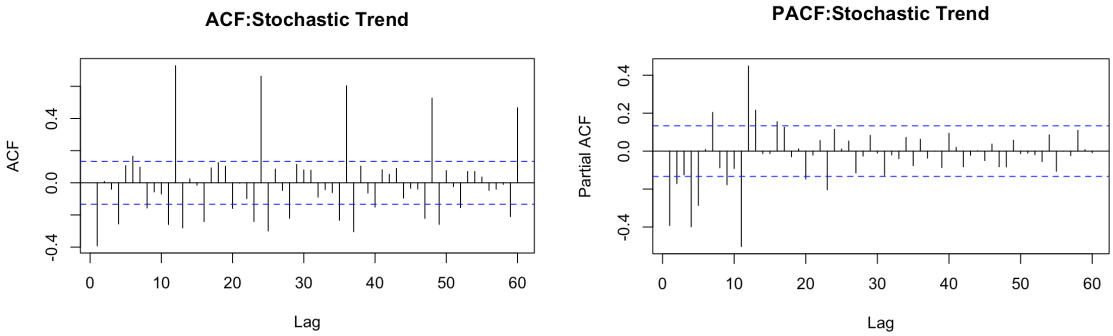
圖十九、A2 案件數量原始資料的 ACF 與 PACF

(3) Stochastic Trend

要實現 Stochastic Trend，可以對資料進行一階差分(Difference)，或是配適 ARIMA(0,1,0)模型並擷取模型殘差作為差分後的新資料，相較於採用線性迴歸的 Determinist Trend 僅呈現趨勢直線，採用 Difference 的 Stochastic Trend 更同時考慮了資料的震盪情形，差分後的資料也趨近平穩(圖二十)；取得殘差後分析 ACF 跟 PACF(圖二十一)已經簡化許多，呈現典型 Seasonal ARIMA 的形式，在 lag 12 有明顯的自相關，並且 ACF 在 lag 12 的倍數有逐漸拖尾(tail-off)的情形，要處理 lag 12 拖尾的問題，可以考慮進一步排除 Stochastic Seasonality。



圖二十、Stochastic Trend(Difference)模型配適圖(左)與殘差 ts-plot



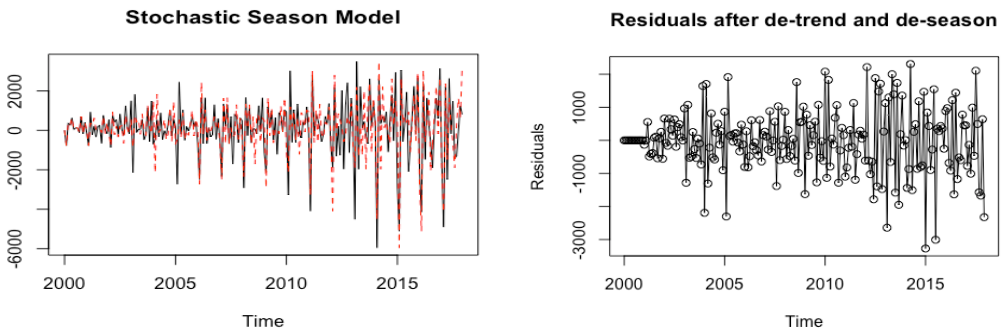
圖二十一、排除 Stochastic Trend(Difference)後的殘差 ACF 與 PACF

2. 季節性效應(Stochastic Seasonality)

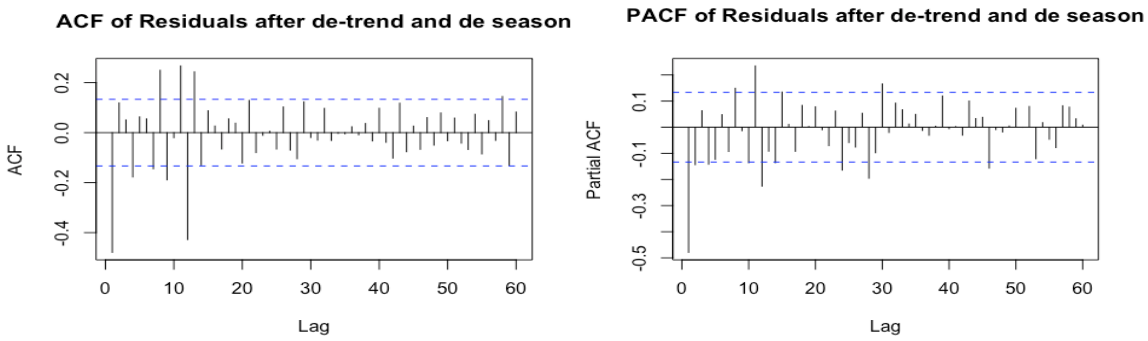
面對上述 lag 12 拖尾的問題有兩個可行的方法，其一是在後續配適模型(Modeling)時考慮加入 seasonal parameter P 且令 period = 12，但若該項係數過高，其實就可以直接考慮 Stochastic Seasonality，亦即針對 lag 12 再進行一次差分(Difference)，兩種方式都進行測試後，發現排除 Stochastic Seasonality 的步驟有其必要性。

(1) Stochastic Seasonality

除了以 lag 12 差分的方式以外，也可以配適 SARIMA(0,0,0)(0,1,0)[12]並擷取模型殘差作為排除季節效應後的新資料；雖然配適結果並非完全貼合資料(圖二十二)，但其目的在於消除季節性效應(Seasonality)的影響，因此不影響分析進行，另外殘差資料也呈現平穩(stationary)狀態，前面 12 個點呈現直線是因為 lag 12 的限制導致前 12 個觀察值無法進行差分。排除季節性效應後資料的 ACF 與 PACF(圖二十三)將會是後續建立模型(Modeling)的重要依據，從 ACF 可以發現在 lag1 及 lag 12 仍有超過 0.4 的負自相關係數，在 PACF 除了 lag 1 也有高度負相關係數外，lag 12 的倍數也有些微拖尾的情形，然而不論 ACF 或 PACF，在其他 lag 數也些無法忽略的正相關係數，這使得主觀判斷 SARIMA 參數難度較高。



圖二十二、Stochastic Seasonality 模型配適圖(左)與殘差 ts-plot



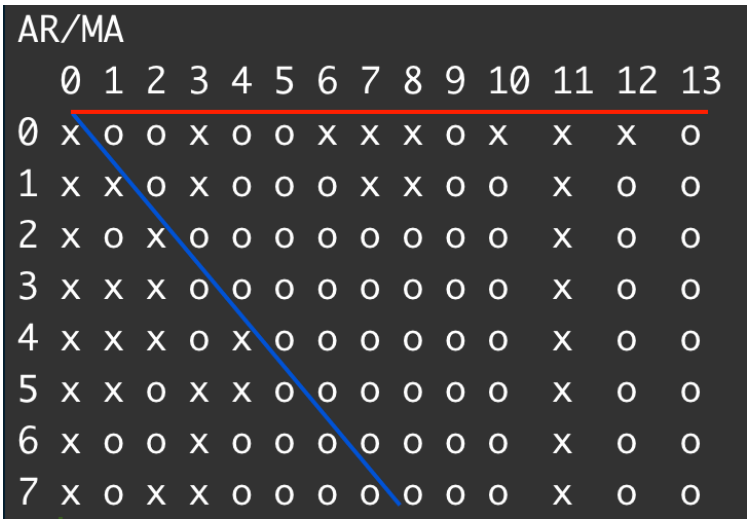
圖二十三、排除 Stochastic Seasonality 後的殘差 ACF 與 PACF

3. 模型配適(Modeling)

由於複雜的 ACF 及 PACF 導致 SARMA 參數難以評估，因此除了主觀猜測之外，也可以先借助 EACF 及 aito.arima 函數協助判斷參數。

(1)EACF

EACF 僅能判斷 ARMA 的 non-seasonal 參數，無法判斷 seasonal 參數，但仍然有一定的參考價值，從 EACF(報表十)可以發現可能的 ARMA 參數為 p=0 且 q = 1，也就是 MA(1)的模型，然而 MA(1)的模型理想是 ACF 在 lag 1 明顯截尾，lag2 以後自相關係數低，PACF 則應該要有拖尾的情況，然而實際圖形(圖二十三)和理想情況仍有些為落差，因此再以 auto.arima 方法進一步協助判斷，看是否能呼應 EACF 結果。



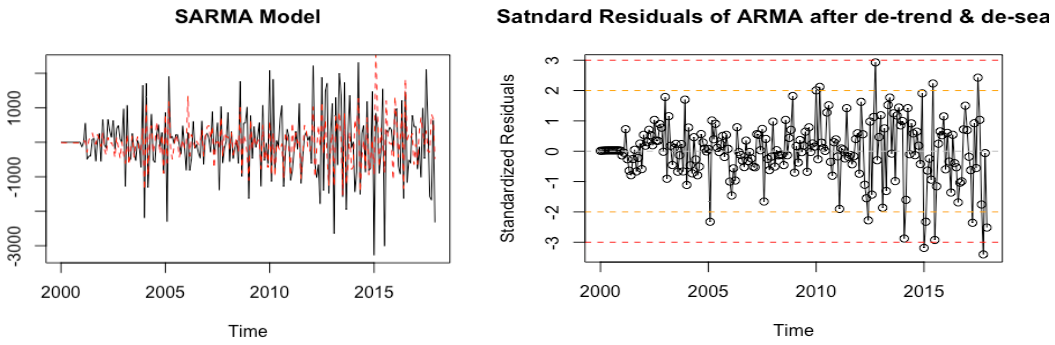
報表九、EACF 報表

(2) auto.arima

由於資料是承接 Stochastic Trend 及 Stochastic Seasonality，因此設定 auto.arima 參數時令 d 和 D 固定為 0，最後以最小 AIC 為篩選指標得候選模型為 SARMA(0, 1)(0, 1)[12]，AIC 為 3472.036(報表)，該模型同時考慮的 non seasonal MA parameter(q = 1)以及 seasonal MA parameter(Q=1)，也呼應了 EACF(報表九)中的 MA(1)的部分，以及 PACF(圖二十三)在 lag 12 偏高並且有些為拖尾的現象。模型配適圖形(圖二十四)的配飾效果雖不到完全貼合原始資料，殘差在時間後期也還有些離群值(圖二十四)，仍留到模型診斷(Diagnosis)以及預測(Prediction)在下結論。

```
##  Fitting models using approximations to speed things up...
##
##  ARIMA(2,0,2)(1,0,1)[12] with non-zero mean : 3488.884
##  ARIMA(0,0,0)            with non-zero mean : 3581.087
##  ARIMA(1,0,0)(1,0,0)[12] with non-zero mean : 3504.55
##  ARIMA(0,0,1)(0,0,1)[12] with non-zero mean : 3467.868
##  ARIMA(0,0,0)            with zero mean     : 3579.179
##  ARIMA(0,0,1)            with non-zero mean : 3522.59
##  ARIMA(0,0,1)(1,0,1)[12] with non-zero mean : 3482.222
##  ARIMA(0,0,1)(0,0,2)[12] with non-zero mean : 3469.939
##  ARIMA(0,0,1)(1,0,0)[12] with non-zero mean : 3500.84
##.              …… 中間部份省略……
##  ARIMA(0,0,1)(1,0,0)[12] with zero mean     : 3499.465
##  ARIMA(0,0,1)(1,0,2)[12] with zero mean     : Inf
##  ARIMA(0,0,0)(0,0,1)[12] with zero mean     : 3508.579
##  ARIMA(1,0,1)(0,0,1)[12] with zero mean     : 3469.923
##  ARIMA(0,0,2)(0,0,1)[12] with zero mean     : 3468.938
##  ARIMA(1,0,0)(0,0,1)[12] with zero mean     : 3471.799
##  ARIMA(1,0,2)(0,0,1)[12] with zero mean     : 3471.794
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(0,0,1)(0,0,1)[12] with zero mean     : 3472.036
##
##  Best model: ARIMA(0,0,1)(0,0,1)[12] with zero mean
```

報表十、auto.arima 報表

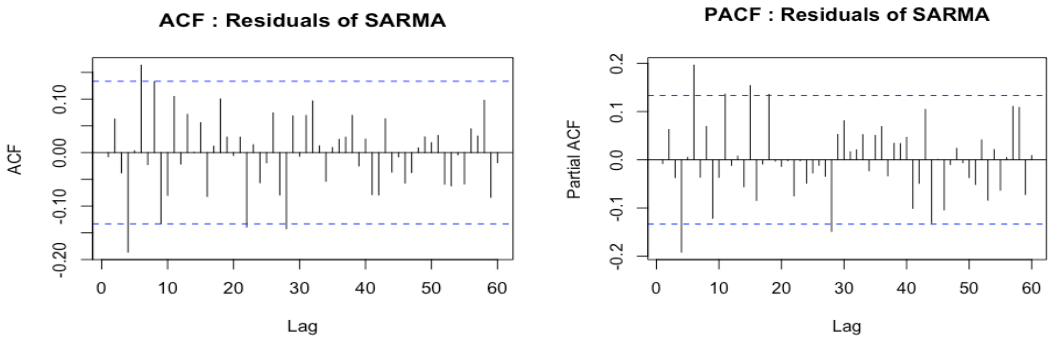


圖二十四、SARMA(0, 1)(0, 1)[12]的模型配適圖及標準化殘差圖

4. 模型診斷(Diagnosis)

(1) SARMA(0, 1)(0, 1)[12]

觀察模型最終殘差的 ACF 及 PACF(圖二十五)，理想情況是呈現白噪音 WhiteNoise，然而在 ACF 跟 PACF 圖中 lag 6 的自相關係數仍然有偏高的趨勢，甚至在 ACF 圖中 lag 6 仍有拖尾的情形，雖然相較於 Determinist Method 的最終殘差 ACF 及 PACF(圖十八)自相關係數較小，但仍然接近兩個標準誤差邊緣，因此考慮再次修正模型，將 lag 6 考慮進 AR 參數中。



圖二十五、SARMA(0, 1)(0, 1)[12]模型的最終殘差 ACF 及 PACF

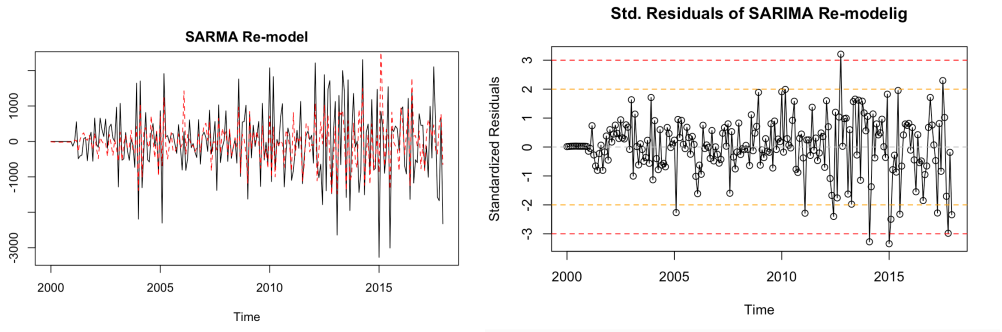
5. 模型修正(Re-modeling)

從 SARMA(0, 1)(0, 1)[12]的最中殘差分析中，考慮再將 lag 6 考慮進 AR 的參數中，但由於模型複雜，因此修正模型為 SARMA(6, 1)(0, 1)[12] with fixed AR parameter，在 R 語言的參數設定上 fixed = c(0, 0, 0, 0, 0, NA)，模型真正形式如下：

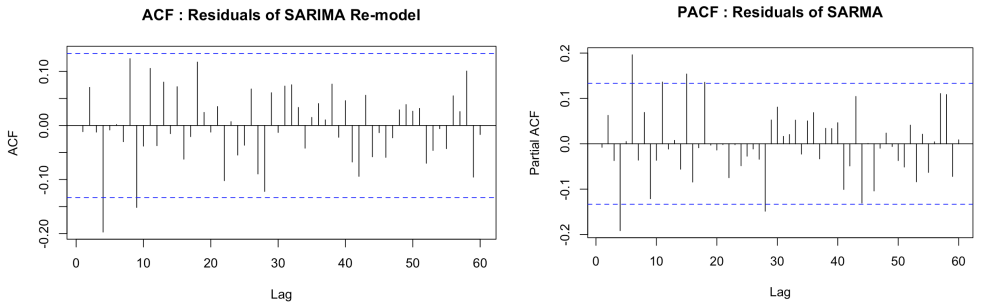
$$(1-\phi_6 B^6)Y_t = (1-\theta_1 B^1)(1-\Theta_1 B^{12})e_t$$

$$\Longrightarrow Y_t = \phi_6 Y_{t-6} + e_t - \theta_1 e_{t-1} - \Theta_1 e_{t-12} + \theta_1\Theta_1 e_{t-13}$$

修正後模型配適圖(圖二十六)與修正前(圖二十四)差不多，但在 ACF 與 PACF 的表現上(圖二十七)，lag 1、lag6、lag12 的自相關終於消除，除了 lag 4 稍微凸出一點以外，殘差幾乎呈現類似白噪音(Whit Noise)的情形，也終於得到理想的最終模型！



圖二十六、SARMA(6, 1)(0, 1)[12]模型配適結果與殘差圖



圖二十七、SARMA(6, 1)(0, 1)[12]模型殘差的 ACF 與 PACF

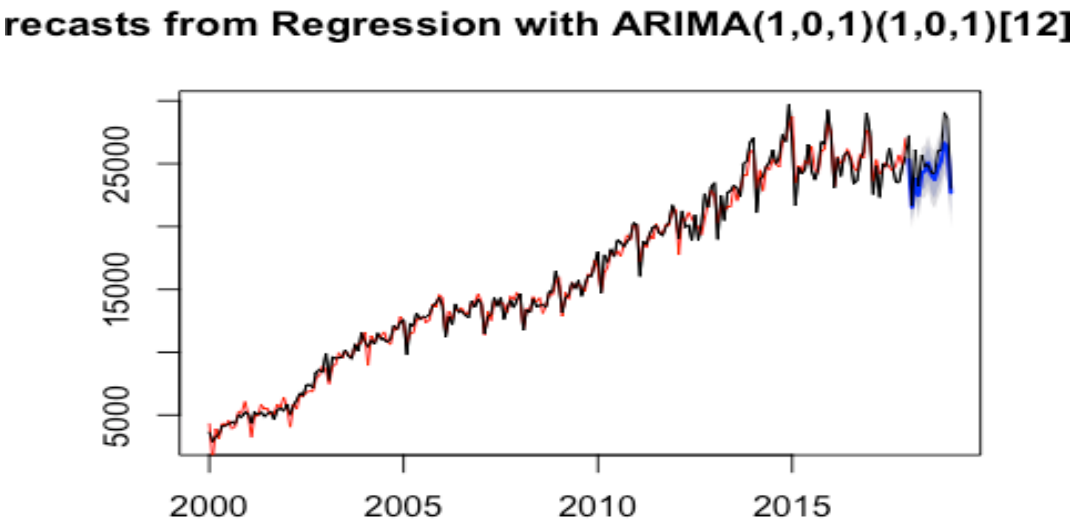
四、模型預測

模型的優劣也可以透過預測能力的強弱來區分，本章節將比較 Determinist Method、Stochastic Method，並另外加入折衷兩者的 Mix Method 一同比較，透過訓練 2000 年 1 月到 2017 年 12 月的 A2 原始資料，來預測 2018 年 1 月到 2019 年 2 月共 14 個月的結果。

1. Determinist Method

$(1-\Sigma_i^p \phi_i B_i)Y_t = (1-\Sigma_j^q \theta_j B^j)e_t + (\Sigma \gamma_k m_k) + (\beta_1 t + \beta_0)$

此方法的在模型配飾上非常貼近 A2 原始資料，預測線也非常貼近實際值，在預測初期的震盪情形仍有加強的空間，但實際值都落在預測值的 95%區間之中。

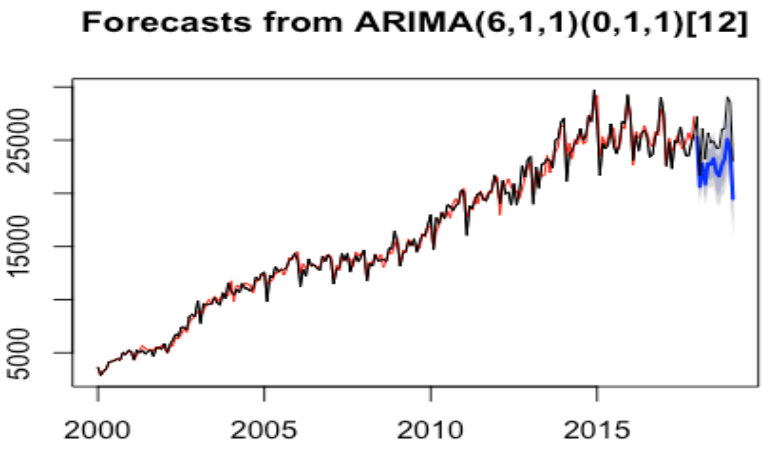


圖二十八、Determinist Method 預測結果

2. Stochastic Method

$(1-B)^d(1-B^s)^D(1-\Sigma_i^p \phi_i B^i)(1-\Sigma_j^P \Phi_j B^{js})Y_t$
$= (1-\Sigma_k^q \theta_k B^k)(1-\Sigma_l^Q \theta_l B^{ls})e_t$

此方法的在模型配飾上同樣非常貼近 A2 原始資料，預測線則有向下平移的情形，但震盪情形幾乎完全相同，實際值都勉強落在預測值的 95%區間之中。



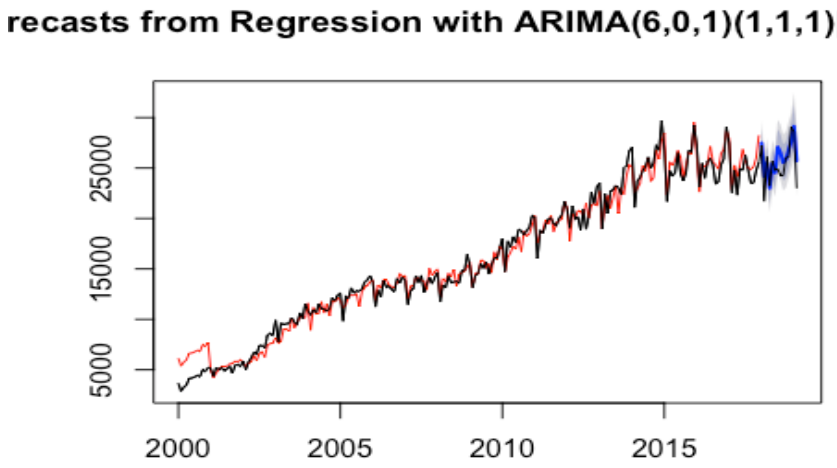
圖二十九、Stochastic Method 預測結果

3. Mix Method

考量到 Determinist Method 的預測趨勢較貼近實際值，但 Stochastic Method 的鋸齒震盪更貼近實際情況，因此綜合兩者的優點，混合了 Determinist Trend 跟 Stochastic Seasonality，採用了以下模型：

$$(1-B^s)^D(1-\Sigma_i^p\phi_iB^i)(1-\Sigma_j^P\Phi_jB^{js})Y_t$$
$$=(1-\Sigma_k^q\theta_kB^k)(1-\Sigma_l^Q\theta_lB^{ls})e_t+(\beta_1t+\beta_0)$$

模型結果(圖三十)相較於前兩者，除了配飾線線在初期有些微偏差以外，預測線不僅抓到了趨勢，並且也比較咬合實際值得震盪情形。



圖三十、Mix Method 預測結果

五、結論

本篇試圖以時間數列模型配飾並預測機動車輛及道路交通事故資料中的 A2 類案件數量，以 Determinist Method 跟 Stochastic Method 兩種方法，逐步地排除趨勢效應(Trend)及季節性效應(Seasonality)，並以解釋性較高的方式配飾模型並透過最終殘差是否符合白噪音(White Noise)來診斷模型優劣，最後在預測上也發現了了 Determinist Method 較能符合趨勢、 Stochastic Method 更能咬合震盪情形，最後結合兩者的優點發展了 Mix Method 同時考量 Determinist Trend 跟 Stochastic Seasonality，也在模型預測中得到不錯的預測結果。

不論 Determinist Method 或 Stochastic Method，都顯示了當期的 A2 案件數量可能跟前 1 期、前 6 期、前 12 期、前 13 期的 A2 案件數量有關，同時跟前 13 期的殘差值有關，前 1 期、前 12 期、前 13 期的「殘差值」有關，並且 A2 案件數量的成長趨勢確實存在，並且月份導致的季節性效應也不能忽略，以上是本次研究的結論。