

# 時間序列分析期末報告

美國航空集團股價與美國新冠肺炎總感染人數

H24089519 吳育儒、F44054045 賴廷瑋、H34064058 莊芯瑜

110 年 1 月 7 日

# 目錄

一、研究動機 .....	2
二、資料介紹 .....	2
三、模型建立 .....	4
A. 美國航空集團股價 .....	4
(1) 時間序列的平穩 .....	4
(2) 模型配飾 .....	7
(3) 殘差檢定 .....	10
(4) McLeod-Li test .....	13
(5) 配飾 Garch 模型 .....	14
(6) 最終模型 .....	17
B. 美國每日新冠肺炎確診人數 .....	18
(1) 時間序列的平穩 .....	18
(2) SARIMA 模型配適 .....	22
(3) 殘差估計 .....	24
(4) McLeod-Li test .....	28
(5) 最終模型 .....	28
四、相關性檢定 .....	29
A. 建立回歸模型 .....	30
(1) 對 $Lag = -1$ 建立回歸模型 .....	30
(2) 對 $Lag = -7$ 建立回歸模型 .....	31
五、結論 .....	32
六、學習心得 .....	33
七、課堂建議 .....	34

八、期中考加分題 .....	35
九、額外建議 .....	37

## 一、研究動機

2020 年初爆發的新冠肺炎席捲全球，在三月造成全球股市崩盤，其中航空股更為慘烈。全球大部分航空公司因疫情限縮旅客流量，造成公司鉅額的虧損；然而我們好奇的是，面對疫情感染人數不斷地攀升與惡化，這些訊息是否有反應於航空股價上，或是航空股價仍由其他因素所主導？此一分析檢驗新冠肺炎感染人數是否與航空股價有連動關係，以提供疫情下航空股投資人更多的決策資訊。

## 二、資料介紹

因美國為疫情最為嚴重的國家之一，故本組針對美國新冠肺炎感染人數與美國航空集團股價進行分析。

### (一) 美國航空集團股價（股票代碼：AAL）

資料來源：Yahoo Finance

資料範圍：2020.05.01 至 2020.12.31

資料間隔：以日為單位

### (二) 美國新冠肺炎總感染人數

資料來源：CDC Covid Data Tracker

資料範圍：2020.05.01 至 2020.12.31

資料間隔：以日為單位

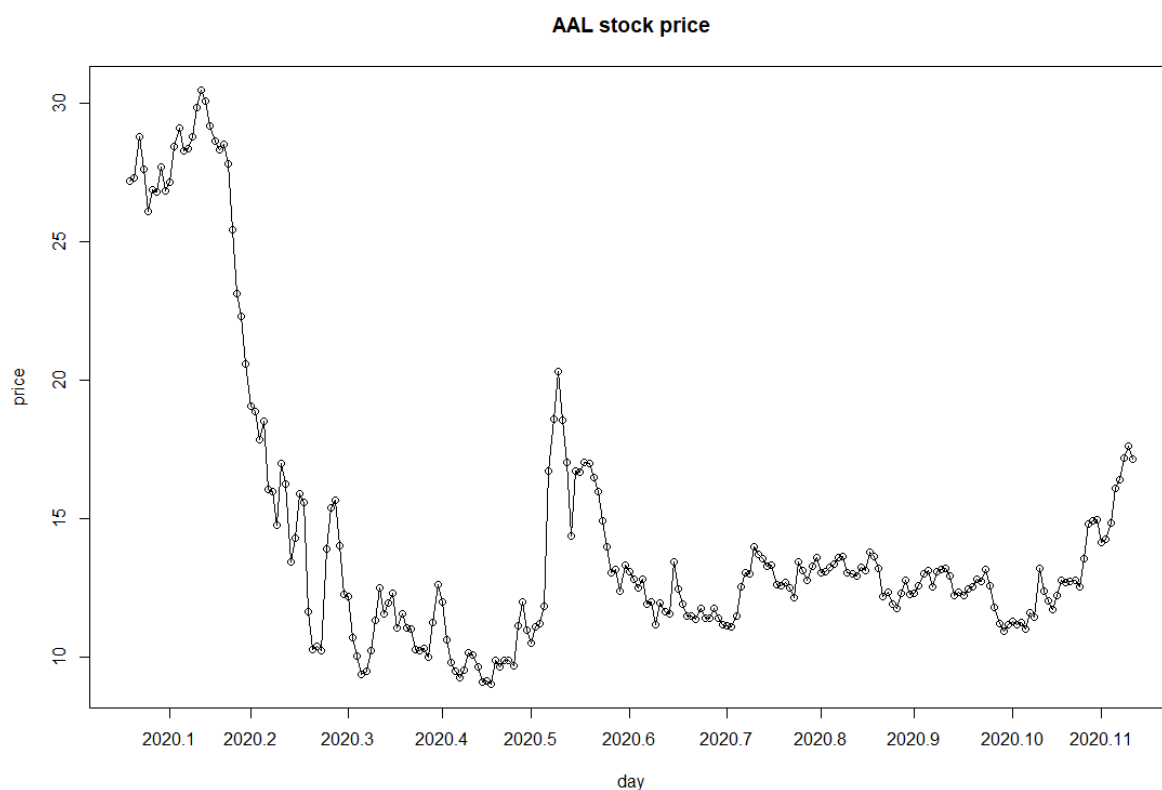


### 三、模型建立

#### A. 美國航空集團股價

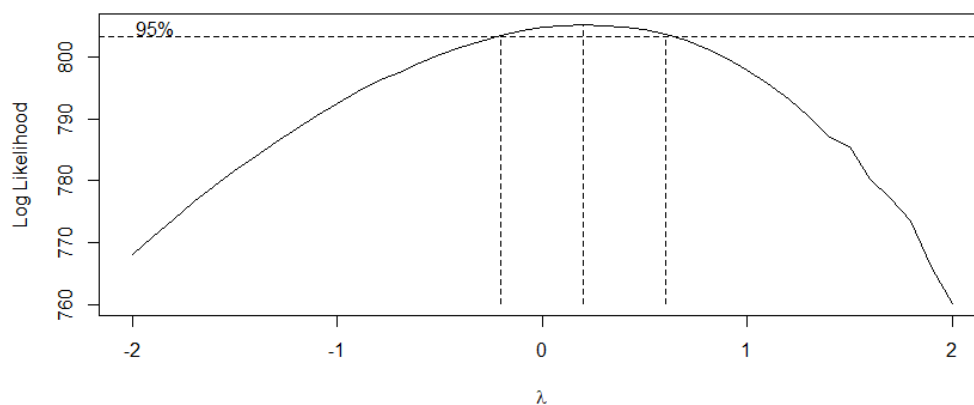
這次收集資料為美國航空集團股價，欄位名稱為 Close，時間從 2020/1/21~2020/12/31。

##### (1) 時間序列的平穩

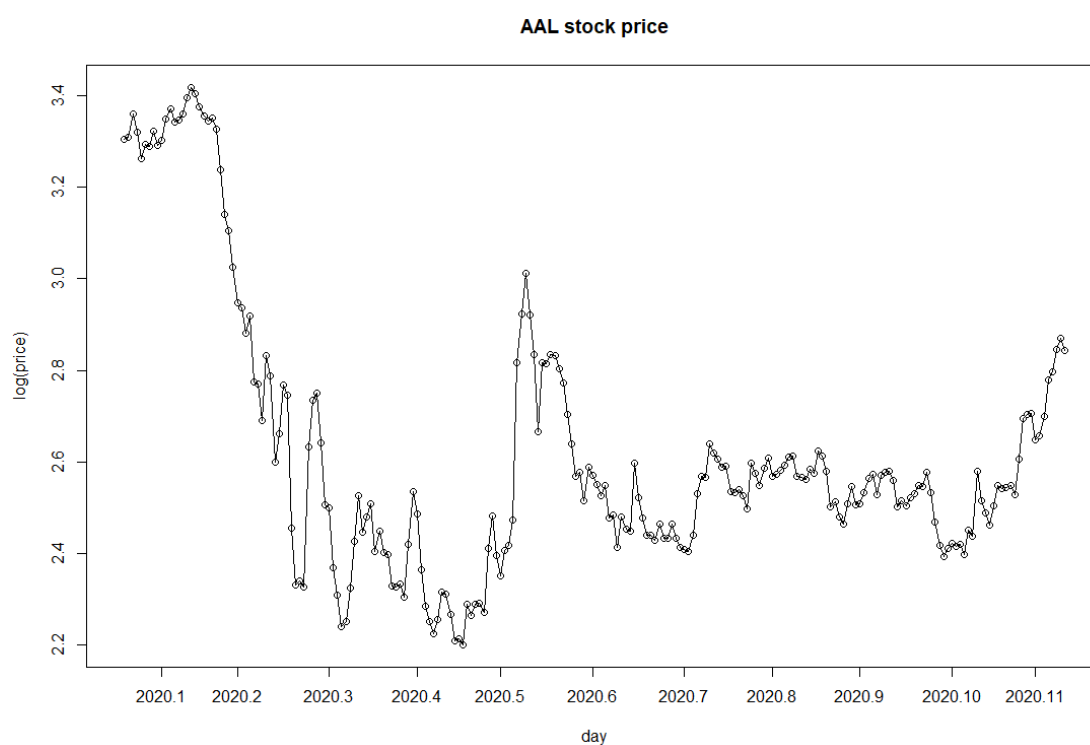


上圖為每天美國航空股票的時間序列圖，自 2020 年 1 月 21 日至 2020 年 12 月 9 日，以天為單位，共 226 筆資料。由此圖可以看出資料為不平穩之時間序列，有明顯逐年遞減的趨勢，

故下一步對美國航空股票做 Box-Cox 檢定。轉換方法如：
$$f(x) = \begin{cases} \log x & \text{if } \lambda = 0 \\ \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \end{cases}$$

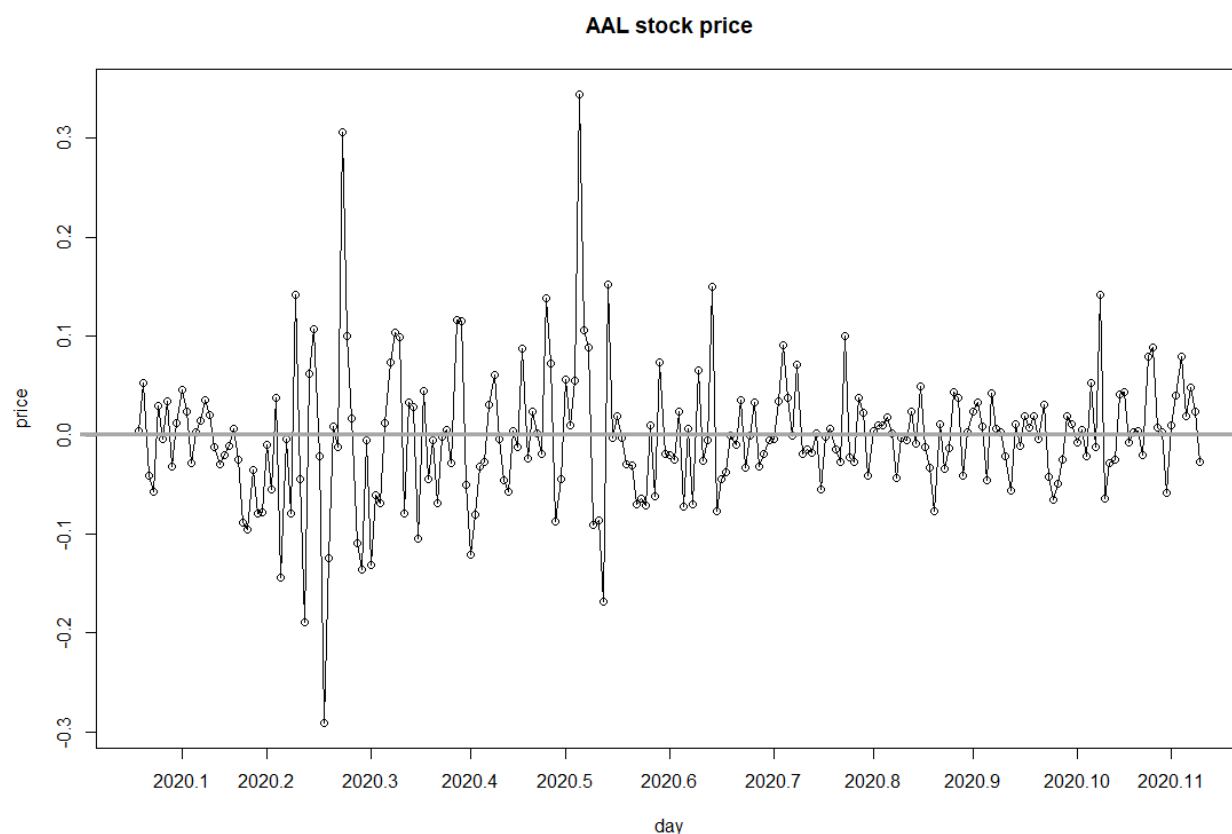


上圖為對美國航空股票資料做 BoxCox檢定所得到的圖，由圖可以看到 95%信賴區間包含 0，因此由 BoxCox檢定，我們對資料進行 log 轉換。



上圖為對美國航空股票資料 log 轉換所得的時間序列圖，但轉換後的資料仍有逐年遞減的趨勢，仍非平穩，於是進行 Augmented Dickey-Fuller Test(以下簡稱 ADF Test)與 KPSS Test 來確認是否需要進行差分。

	ADF Test		KPSS Test
$H_0$	Non-stationary	$H_0$	stationary
$H_a$	stationary	$H_a$	Non-stationary
顯著水準	0.05	顯著水準	0.05
Dickey-Fuller	-1.7777	KPSS Level	1.2067
p-value	0.6693	p-value	< 0.01
檢定結果	0.6693 > 0.05 ，不拒絕 $H_0$		
推論	此時間序列不平穩，仍需差分		

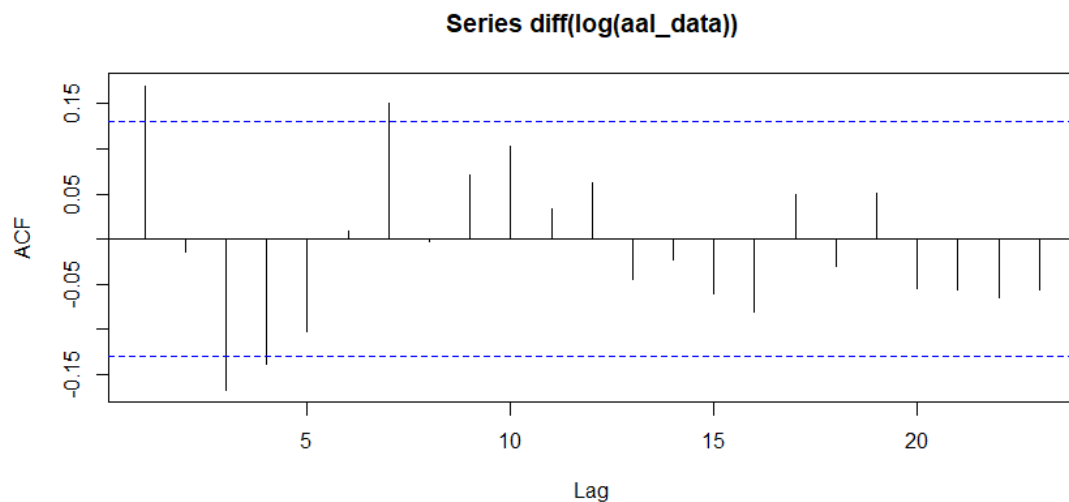


上圖為以取對數後的資料進行一次差分，沒有明顯的向上或向下的趨勢，大致平均分散在 Y 軸為 0 的上下，相較於只取對數的資料平穩許多。但仍有少數幾筆資料的數值離 Y 軸為 0 較遠，因此在進行 Augmented Dickey-Fuller Test(以下簡稱 ADF Test)與 KPSS Test 來確認是否需要進行二次差分。

	ADF Test		KPSS Test
$H_0$	Non-stationary	$H_0$	stationary
$H_a$	stationary	$H_a$	Non-stationary
顯著水準	0.05	顯著水準	0.05
Dickey-Fuller	-5.8709	KPSS Level	0.25288
p-value	< 0.01	p-value	> 0.1
檢定結果	< 0.05 ，拒絕 $H_0$	檢定結果	> 0.05 ，不拒絕 $H_0$
推論	此時間序列已經平穩，不需再差分		

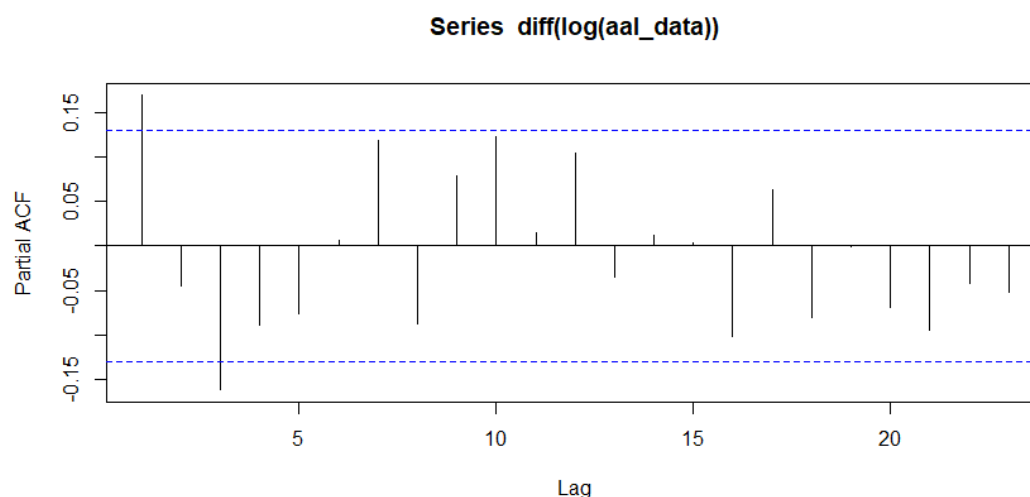
## (2) 模型配飾

由第一部分得到的結論可知，此筆資料取對數後仍需進行一次差分，下圖分別為對所有處理後資料的 ACF、PACF、EACF 圖。



這是資料的 ACF 圖，由圖中可以看到在第四步超過信賴區間，代表第四步顯著，雖然第七步亦超過信賴區間，但是除了這第七步以外，其他第四步後面的步數皆在信賴區間內，因此我們仍先以 MA(4)進行參數估計。





這是資料的 PACF 圖，由圖中可以看到第三步超過信賴區間，代表第三步顯著，且第三步以後的步數皆在信賴區間內，故有截斷的現象，因此我們取 AR(3)進行參數估計。

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

上圖為資料的 EACF 圖，ARIMA 的部分應配飾(1,1,1)。

由 ACF、PACF、EACF 圖得知，我們的候選模型為 MA(4)、AR(3)、ARIMA(1,1,1)，因此接下來對三個模型進行資料配飾。

### ①對 MA(3)模型進行參數估計

經過參數估計後，第四步的參數估計不顯著，因此再拿掉第四步後，故最後進行 MA(3)的參數估計，而因為第二步不顯著，故將第二步參數設為 0，因此得以下最後的參數估計。

下表為對 MA(3)模型進行參數估計。

$\nabla Y_t = Y_t - Y_{t-1} = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \theta_3 e_{t-3}, e_t \sim N(0, \sigma^2)$			
係數	$\theta_1$	$\theta_2$	$\theta_3$
估計值	0.1590	0	-0.1518
95%信賴區間	(0.0244, 0.2936)	(0, 0)	(-0.2864, -0.0172)
除了 $\theta_2$ 以外， $\theta_1$ 與 $\theta_3$ 的 95%信賴區間皆未包括 0，因此這兩個參數顯著，故得到模型為 $\nabla Y_t = Y_t - Y_{t-1} = e_t + 0.159e_{t-1} - 0.1518e_{t-3}, e_t \sim N(0, \sigma^2)$ ，列為候選模型。			

### ②對 AR(3)模型進行參數估計

下表為對 AR(3)模型進行參數估計。

$\nabla Y_t = Y_t - Y_{t-1} = \phi_1 \cdot \nabla Y_{t-1} + \phi_2 \cdot \nabla Y_{t-2} + \phi_3 \cdot \nabla Y_{t-3} + e_t, e_t \sim N(0, \sigma^2)$			
係數	$\phi_1$	$\phi_2$	$\phi_3$
估計值	0.1682	0	-0.1633
95%信賴區間	(0.0388, 0.2976)	(0, 0)	(-0.2925, -0.0341)
除了 $\phi_2$ 以外， $\phi_1$ 與 $\phi_3$ 的 95%信賴區間皆未包括 0，因此這兩個參數顯著，故得到模型為 $\nabla Y_t = Y_t - Y_{t-1} = 0.1682 \cdot \nabla Y_{t-1} - 0.1633 \cdot \nabla Y_{t-3} + e_t, e_t \sim N(0, \sigma^2)$ ，列為候選模型			

③對 ARIMA(1,1,1)模型進行參數估計

$\nabla Y_t = Y_t - Y_{t-1} = \phi_1 \cdot \nabla Y_{t-1} + e_t + \theta_1 e_{t-1}, e_t \sim N(0, \sigma^2)$		
係數	$\phi_1$	$\theta_1$
估計值	0.0811	0.0933
95%信賴區間	(-0.4039, 0.5661)	(-0.3801, 0.5567)
$\phi_1$ 與 $\theta_1$ 的 95%信賴區間皆包括 0，因此這兩個參數不顯著，故不將此模型列入候選模型。		

由取對數且差分後的时间序列圖可以發現在 2 月和 5 月似乎有異常值，所以檢測以上各個模型是否有 IO 或 AO 的存在。由下表可以發現兩個候選模型，MA(3)與 AR(3)都有 IO 和 AO 的存在，但為了避免拿掉太多離群值，我們僅移除 IO，並進行下一部分的殘差檢定。

候選模型	AO	IO
ARIMA(0,1,3)	41, 45, 95	41, 45, 95
ARIMA(3,1,0)	37, 41, 45, 95	41, 45, 95

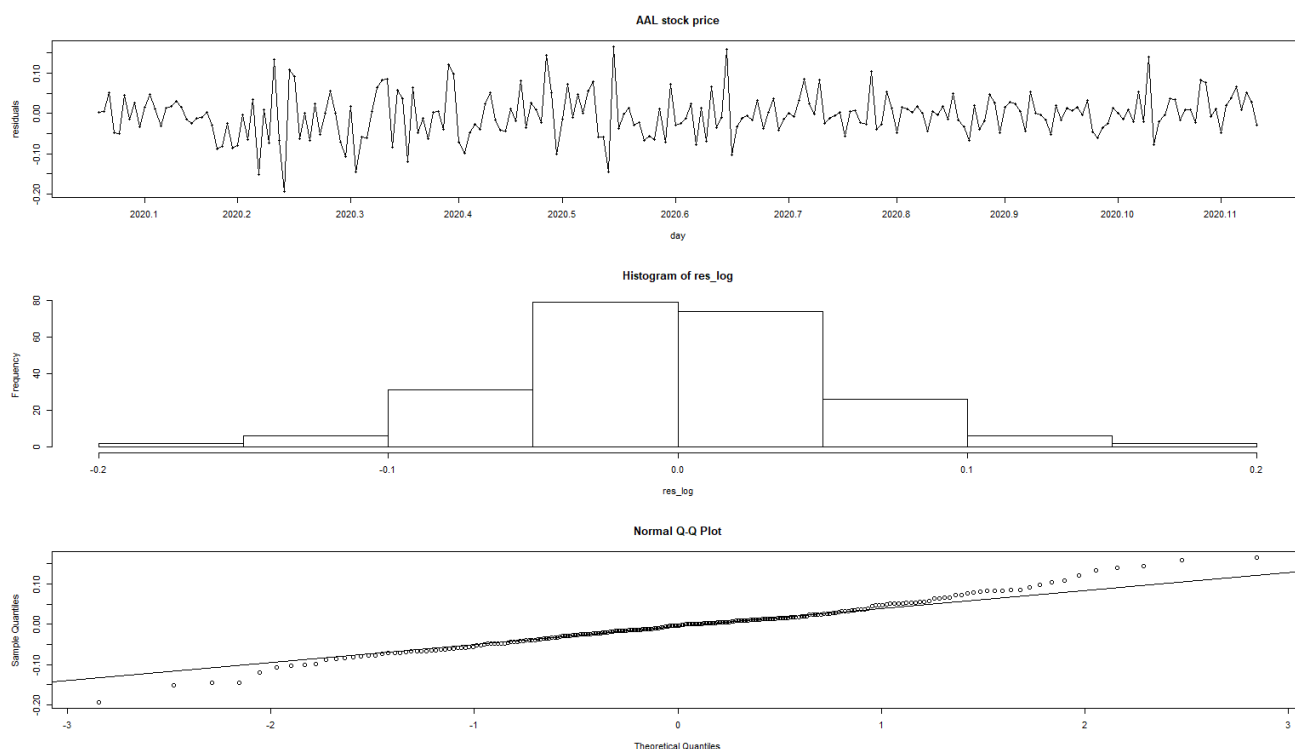
(3) 殘差檢定

對候選模型進行比較：

候選模型	AIC
①ARIMA(0,1,3) + io(41,45,95)	-659.62
②ARIMA(3,1,0) + io(41, 45, 95)	-671.85

由於最後兩個候選皆通過殘差檢定，僅常態沒過，故由上表兩個模型 AIC 的比較可知，①模型配飾比②好，因此接下來僅列出對 ARIMA(3,1,0) + io(41, 45, 95)模型的殘差檢定。

模型為： $\nabla Y_t = Y_t - Y_{t-1} = 0.1478 \cdot \nabla Y_{t-1} - 0.1302 \cdot \nabla Y_{t-3} + e_t - 0.2794IO.41 + 0.2917 \cdot IO.45 + 0.3436 \cdot IO.95, e_t \sim N(0, \sigma^2)$

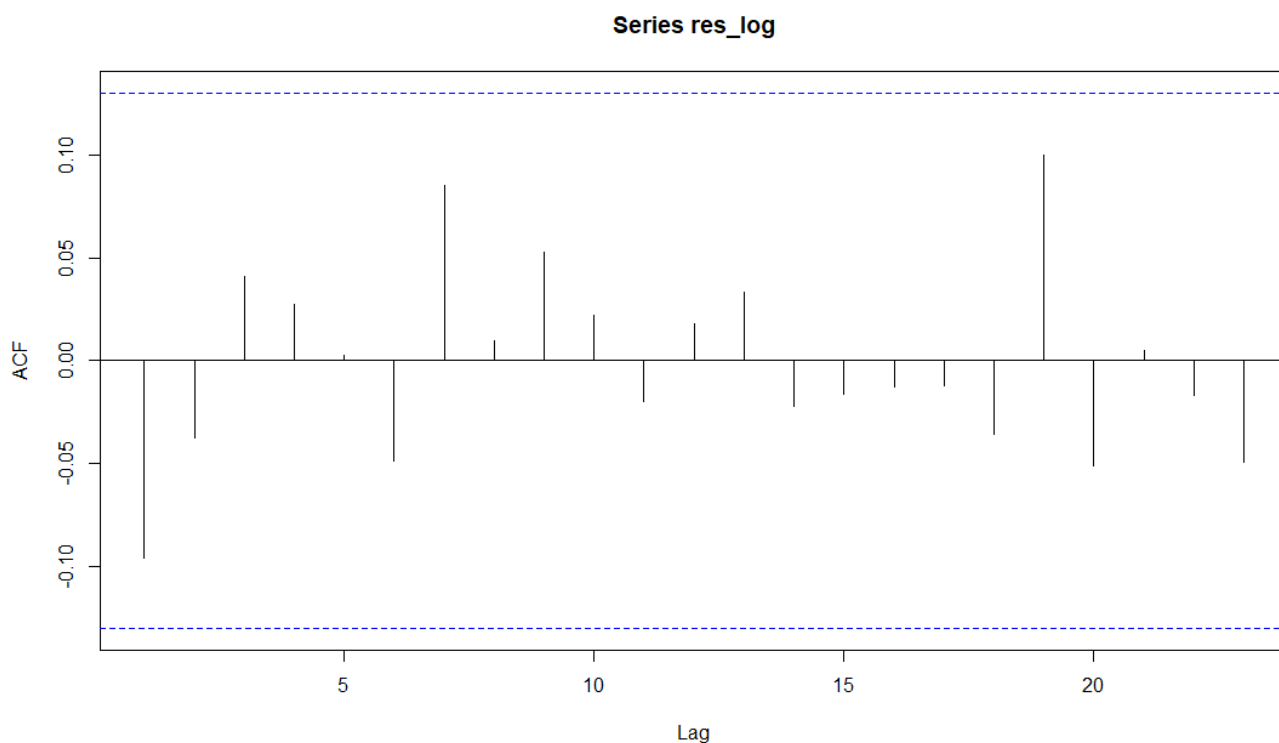


上面三張圖分別為模型的殘差時間序列圖、殘差的直方圖、QQ-plot。殘差的時間序列圖無明顯的趨勢與季節效應，但變異數在某些時間段有變大的趨勢；由直方圖可推測殘差的平均值約為 0；在 QQ-plot 的左側與右側明顯不符合常態，有厚尾的現象，故推測此時間序列有 ARCH 效應。

對模型進行 T-test 與 Shapiro-Wilk normality test。

	T-test	Shapiro-Wilk normality test
$H_0$	$\mu=0$	殘差分布符合常態
$H_a$	$\mu \neq 0$	殘差分布不符合常態
檢定值	-0.99724	0.98368
p-value	0.3197	0.01054

對殘差做 t-test，因為  $p\text{-value} > 0.05$ ，故接受  $H_0$ 。搭配上方的直方圖，我們可得殘差的平均值為 0 的結論。對殘差做 Shapiro-Wilk normality test，因為  $p\text{-value} < 0.05$ ，故不接受  $H_0$ ，搭配上方的 QQ-plot，我們得殘差不符合常態的結論。



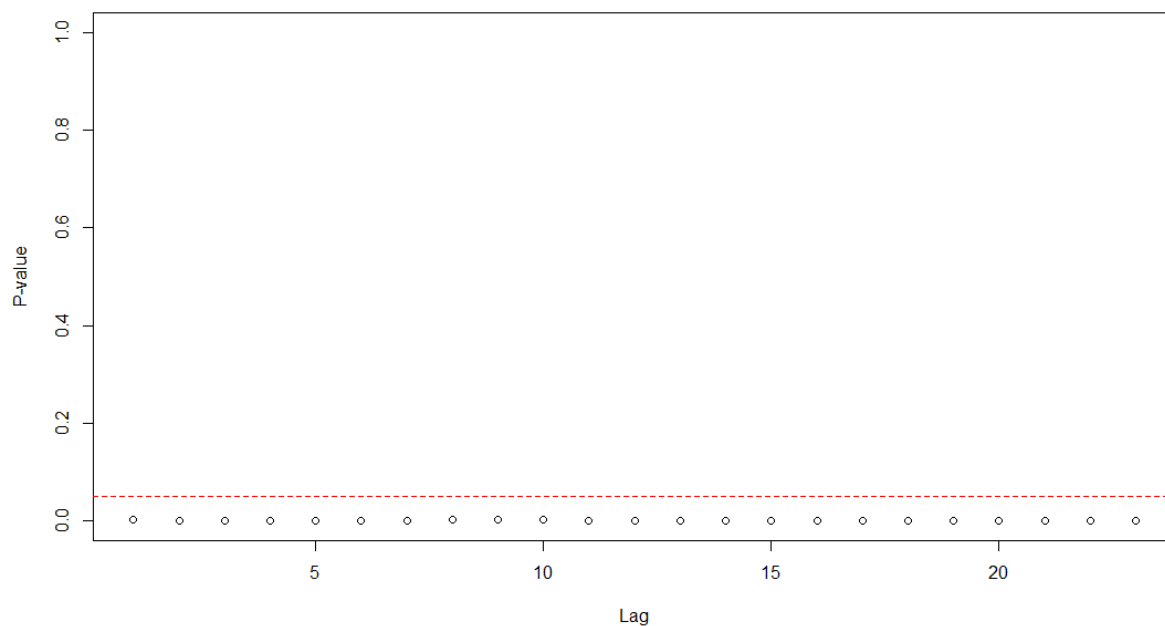
上圖為殘差的 ACF 圖，由圖可看到所有步數皆在信賴區間內，因此殘差無時間序列相關。

最後將 ARIMA(0,1,3) + io(41,45,95)與 ARIMA(3,1,0) + io(41, 45, 95)模型進行比較。

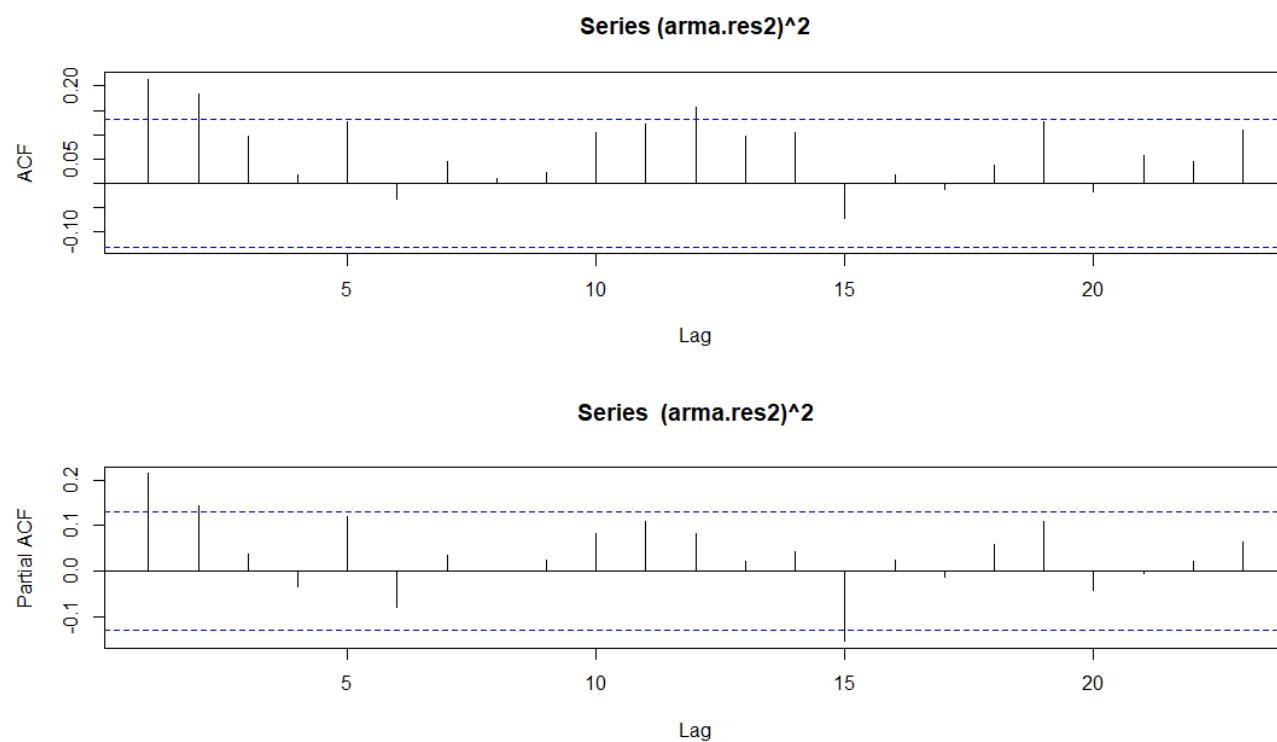
	t-test	Shapiro-Wilk normality test	ACF 圖	AIC
ARIMA(0,1,3) + io(41,45,95)	殘差平均為 0	不符合常態	殘差無時間序列相關	-659.62
ARIMA(3,1,0) + io(41, 45, 95)	殘差平均為 0	不符合常態	殘差無時間序列相關	-660.75

由上表以 ARIMA(3,1,0) + io(41, 45, 95)， $\nabla Y_t = Y_t - Y_{t-1} = 0.1478 \cdot \nabla Y_{t-1} - 0.1302 \cdot \nabla Y_{t-3} + e_t - 0.2794I0.41 + 0.2917 \cdot I0.45 + 0.3436 \cdot I0.95$ ， $e_t \sim N(0, \sigma^2)$ 模型最為我們現階段最佳的模型。由於殘差的 Q-Q-plot 與殘差的時間序列圖推知，資料可能有 ARCH 效應，因此我們對資料進行 McLeod.Li test。

#### (4) McLeod.Li test



上圖為對資料進行 McLeod.Li test 所取得的圖，由圖可看到 p-value 都在信賴區間內，因此此資料有 ARCH 效應。接著看殘差平方的 ACF、PACF 圖。



由殘差平方的 ACF 圖可看到第 1, 2, 12 步都顯著；PACF 圖則為第 1, 2, 15 步顯著。由於無法由圖看出明顯的 ARCH 配飾，因此以試誤法分別配飾 GARCH(1,1), (1,0), (2,0)。

### (5) 配飾 Garch 模型

模型	t-test	Shapiro-Wilk normality test	ACF 圖	AIC
GARCH(1,1)	殘差平均為 0	不符合常態	殘差無時間序列相關	-3.047901
GARCH (1,0)	殘差平均為 0	不符合常態	殘差無時間序列相關	-3.054722
GARCH (2,0)	殘差平均為 0	不符合常態	殘差無時間序列相關	-3.059874

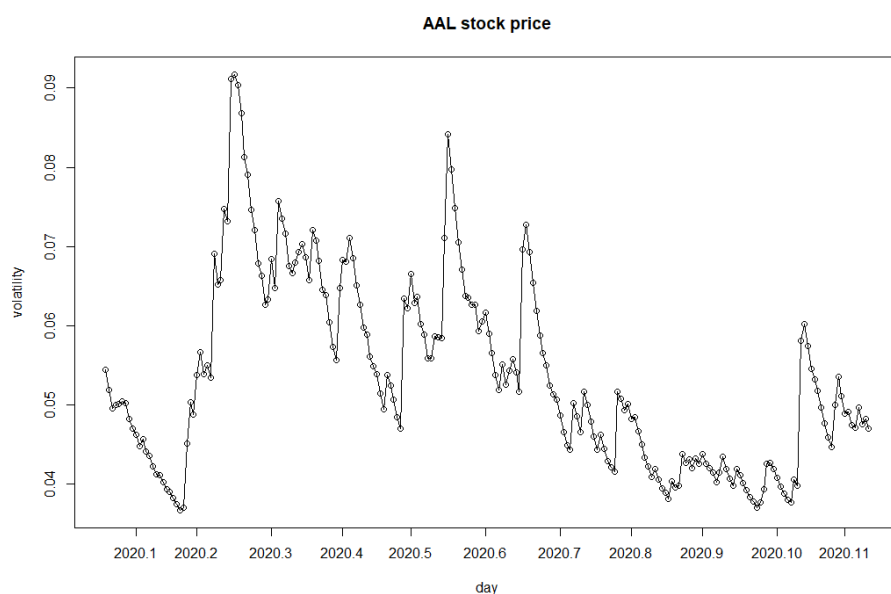
由上表可知三組模型的常態都未通過，因此我們看 AIC 得出配飾 GARCH(2, 0)最合適。下一段會詳細說明配飾 GARCH(2, 0)模型。

下表為配飾 GARCH(2, 0)模型的參數估計

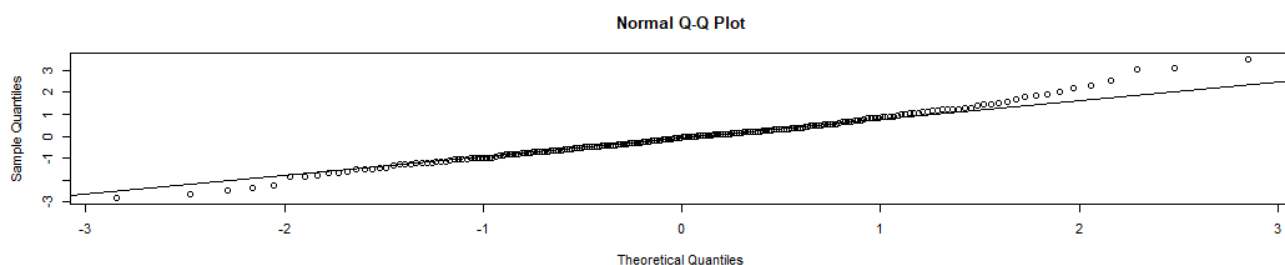
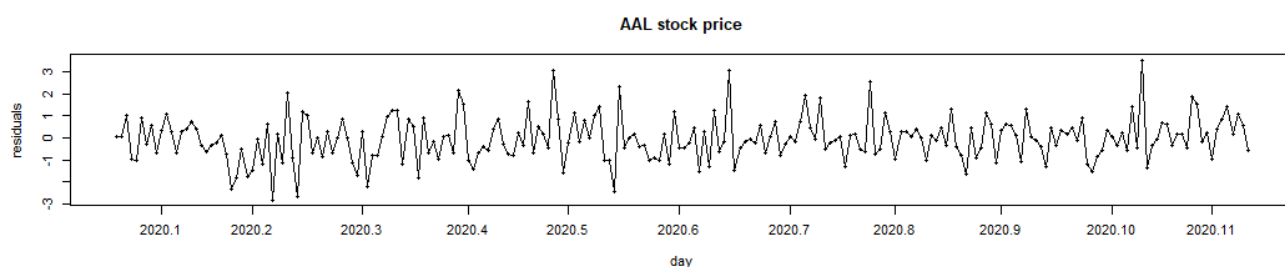
$\begin{cases} \nabla Y_t = Y_t - Y_{t-1} = 0.1478 \cdot \nabla Y_{t-1} - 0.1302 \cdot \nabla Y_{t-3} + e_t - 0.2794IO.41 + 0.2917 \cdot IO.45 + 0.3436 \cdot IO.95 \\ e_t = \sigma_t \cdot \tau_t, \quad \tau_t \sim N(0, \sigma^2) \\ \sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \alpha_2 r_{t-2}^2 \end{cases}$			
係數	$\alpha_0$	$\alpha_1$	$\alpha_2$
估計值	0.0017824	0.1959085	0.2118338
t-value	6.124	2.101	2.178
p-value	9.15e-10	0.0356	0.0294
$\alpha_0$ 、 $\alpha_1$ 、 $\alpha_2$ 的 p-value 皆小於 0.05，因此這三個參數顯著不等於 0。			

接著對 GARCH(2, 0)模型做殘差檢定，模型為

$$\begin{cases} \nabla Y_t = Y_t - Y_{t-1} = 0.1478 \cdot \nabla Y_{t-1} - 0.1302 \cdot \nabla Y_{t-3} + e_t - 0.2794IO.41 + 0.2917 \cdot IO.45 + 0.3436 \cdot IO.95 \\ e_t = \sigma_t \cdot \tau_t, \quad \tau_t \sim N(0, \sigma^2) \\ \sigma_t^2 = 0.00017284 + 0.1959085 r_{t-1}^2 + 0.2118338 r_{t-2}^2 \end{cases}$$



這是 GARCH(2, 0)模型中 volatility 的時間序列圖，由圖中可以看到前半年的變異較大，後半年的變異變小。



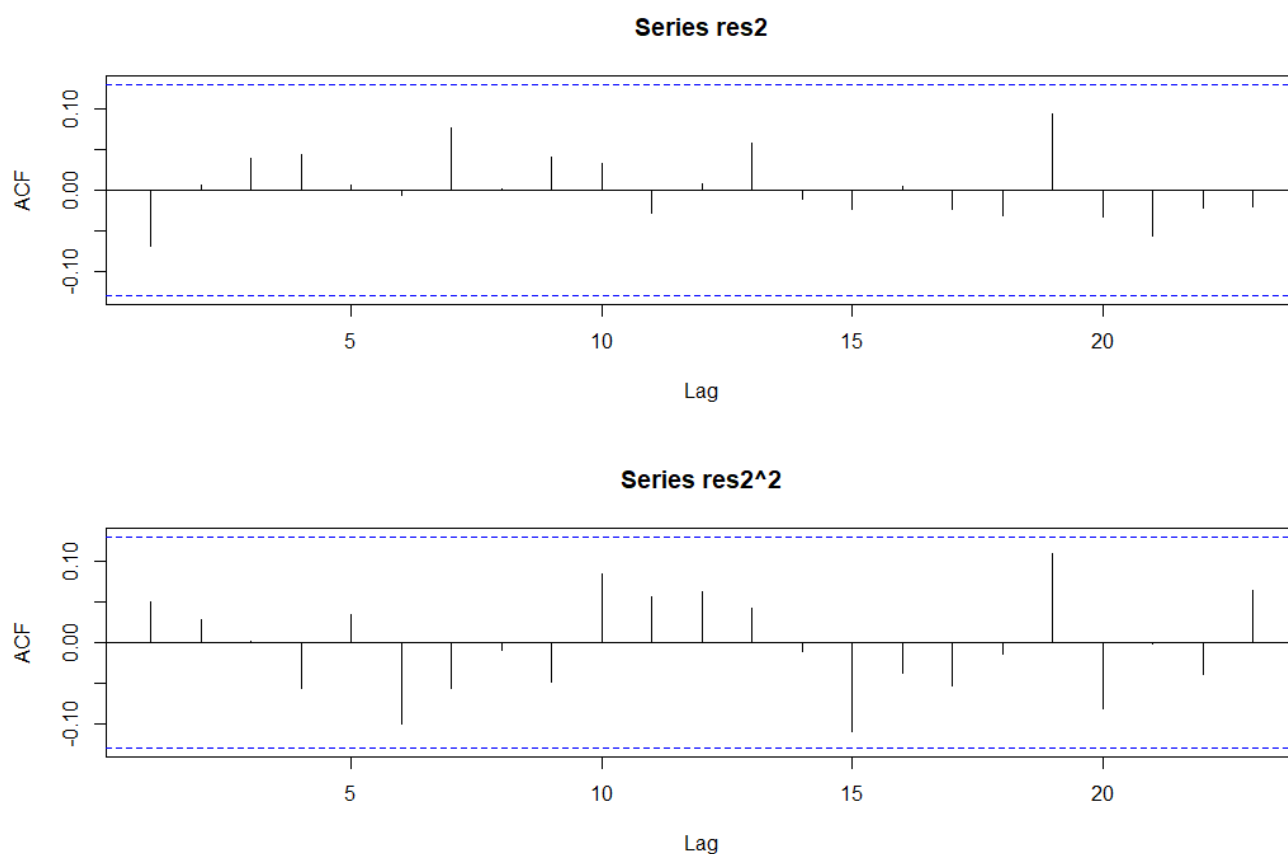
上面三張圖分別為模型的殘差時間序列圖、殘差的直方圖、QQ-plot。殘差的時間序列圖無明顯的趨勢與季節效應，但變異數在某些時間段有變大的趨勢；由直方圖可推測殘差的平均值約為 0；在 QQ-plot 的左側與右側有些值不符合常態。



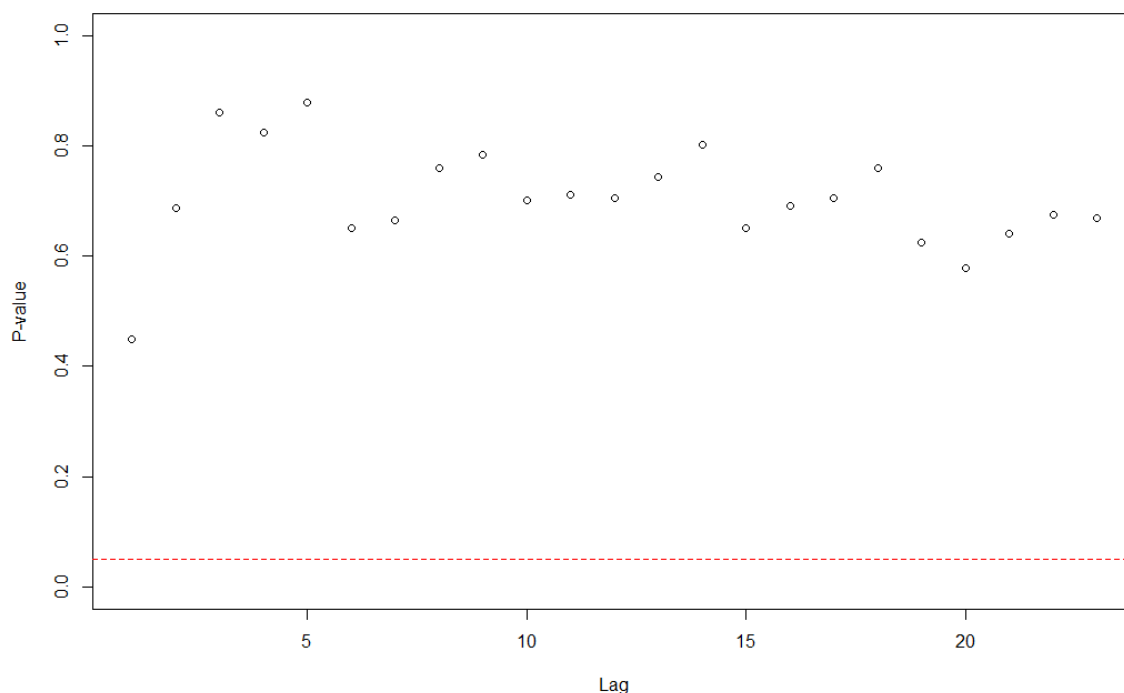
對模型進行 T-test 與 Shapiro-Wilk normality test。

	T-test	Shapiro-Wilk normality test
$H_0$	$\mu=0$	殘差分布符合常態
$H_a$	$\mu\neq 0$	殘差分布不符合常態
檢定值	-0.74892	0.98394
p-value	0.4547	0.01162

對殘差做 t-test，因為  $p\text{-value}>0.05$ ，故接受  $H_0$ 。搭配上方的直方圖，可得殘差的平均值為 0 的結論。對殘差做 Shapiro-Wilk normality test，因為  $p\text{-value}<0.05$ ，故不接受  $H_0$ ，搭配上方的 QQ-plot，殘差不符合常態的結論。



上圖為殘差與殘差平方的 ACF 圖，由圖可看到所有步數皆在信賴區間內，因此殘差無時間序列相關。



最後我們再做一次 McLeod.Li test，此時所有步數的 p value 都落在 0.05 之上，因此得到已經無 ARCH 效應的結論。

## (6) 最終模型

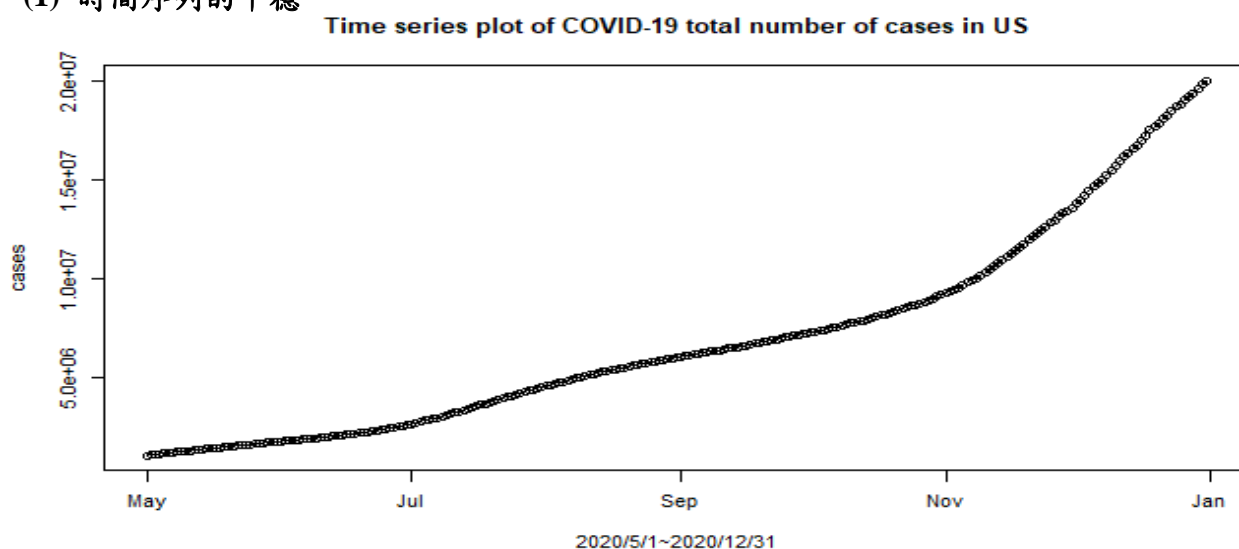
$$\left\{ \begin{array}{l} \nabla Y_t = Y_t - Y_{t-1} = 0.1478 \cdot \nabla Y_{t-1} - 0.1302 \cdot \nabla Y_{t-3} + e_t - 0.2794 IO.41 + \\ \quad 0.2917 \cdot IO.45 + 0.3436 \cdot IO.95 \\ e_t = \sigma_t \cdot \tau_t, \quad \tau_t \sim N(0, \sigma^2) \\ \sigma_t^2 = 0.00017284 + 0.1959085 r_{t-1}^2 + 0.2118338 r_{t-2}^2 \end{array} \right.$$

我們先對資料做 log return，選擇 AR(3)做為我們的最佳模型，接著拿掉三個離群值，並對資料做 Garch 配飾，而以 Garch(2, 0)配適的最好。然而，我們最後配飾的 ARCH 模型仍然無法通過檢定，因此我們得到以上最終的模型。

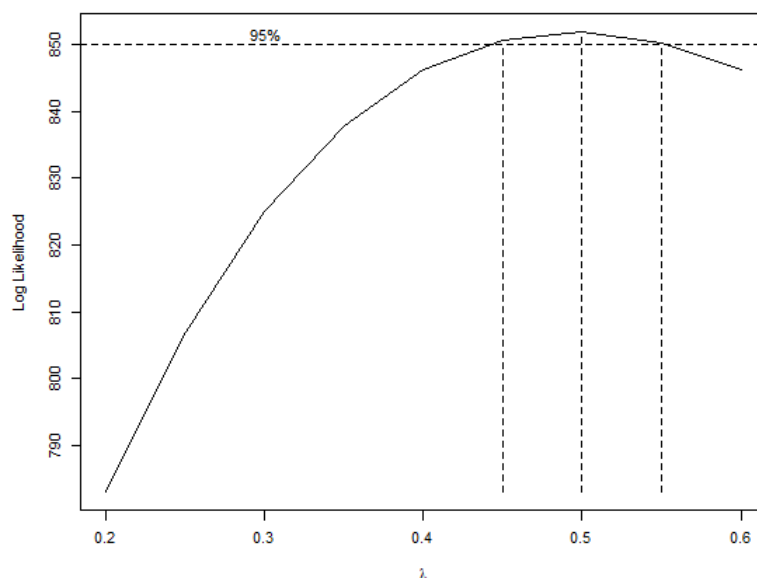
## B. 美國每日新冠肺炎確診人數

這次收集資料為美國每日新冠肺炎確診人數，時間從 2020/1/21~2020/12/31，但因為 1/21~4/30 的新冠肺炎確診人數偏少，且後來產生無法建模的情形，因此，觀察範圍改為 2020/5/1~2020/12/31 的新冠肺炎確診人數。

### (1) 時間序列的平穩



由此時間序列圖可以發現每日的新冠肺炎的確診人數隨著時間的是呈持續攀升的趨勢，因此，推測為此筆資料是不平穩的。因此，為了移除趨勢，故對此筆資料進行 Box cox 檢定。



由此圖可以得知  $\lambda = 0.5$ ，因此由 Box cox 轉換可以得知不需要做資料做轉換。

而時間序列的波動仍隨著時間的變化而改變，所以仍然無法解決趨勢現象。因此，為了解決此問題，我們將判斷差分的必要與否。

為了判斷此筆資料是否需要做差分，採用 KPSS 檢定及 ADF 檢定來檢驗。

	ADF Test		KPSS Test
$H_0$	Non-stationary	$H_0$	stationary
$H_a$	stationary	$H_a$	Non-stationary
顯著水準	0.05	顯著水準	0.05
Dickey-Fuller	-0.30688	KPSS Level	3.7544
p-value	0.99	p-value	0.01
檢定結果	$0.99 > 0.05$ ，不拒絕 $H_0$	檢定結果	$0.01 < 0.05$ ，拒絕 $H_0$
推論	此時間序列不平穩，仍需差分		

綜合以上兩種檢定得知資料是不平穩的，因此需要對這筆資料進行差分，再觀察其時間序列圖是否平穩。因此，對資料再觀察是否需要進行二次差分。

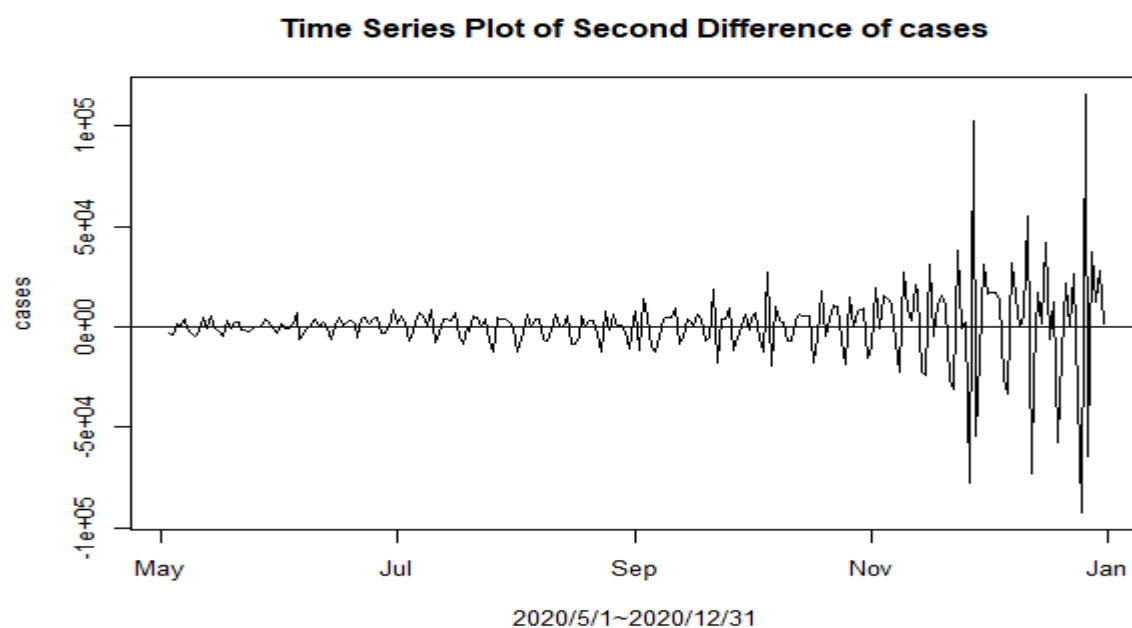
	ADF Test		KPSS Test
$H_0$	Non-stationary	$H_0$	stationary
$H_a$	stationary	$H_a$	Non-stationary
顯著水準	0.05	顯著水準	0.05
Dickey-Fuller	-0.88972	KPSS Level	3.5029
p-value	0.9256	p-value	0.01
檢定結果	$0.9256 > 0.05$ ，不拒絕 $H_0$	檢定結果	$0.01 < 0.05$ ，拒絕 $H_0$
推論	此時間序列不平穩，仍需差分		

因此由以上結果得知，需要對這筆資料做二次差分。

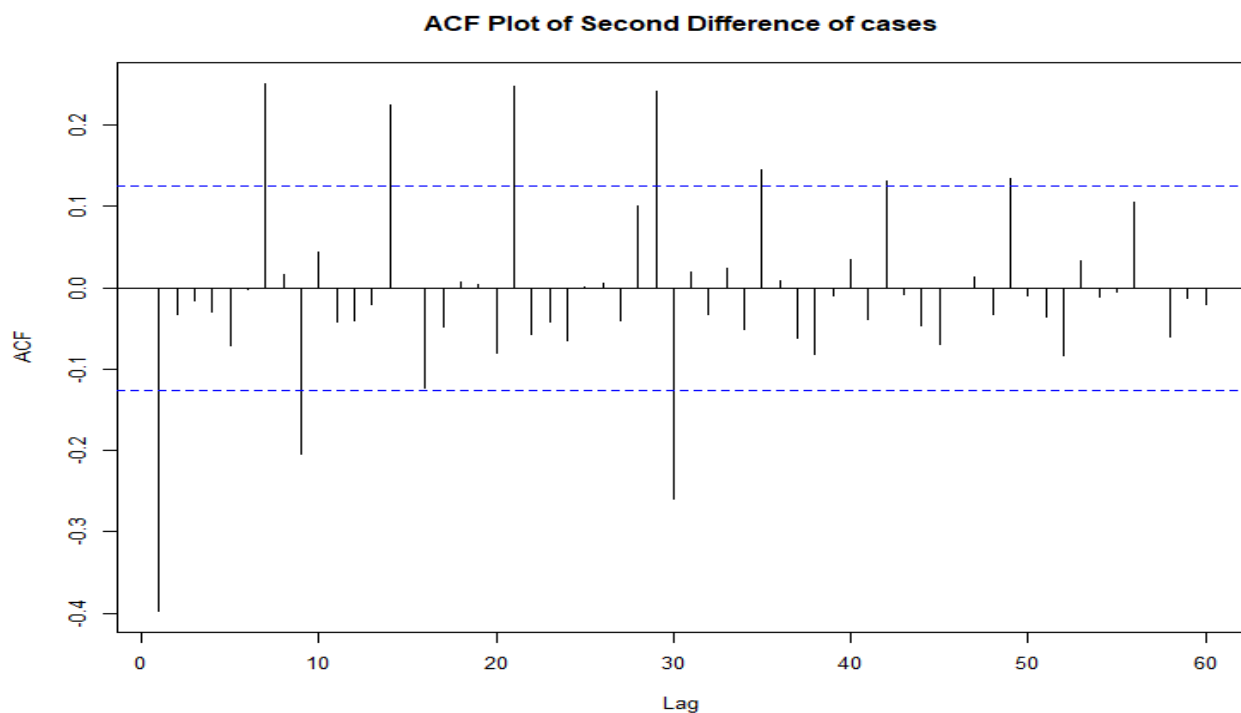
因此由以下的檢定發現經過二次差分後，可確定只需做到二次差分即可。

	ADF Test		KPSS Test
$H_0$	Non-stationary	$H_0$	stationary
$H_a$	stationary	$H_a$	Non-stationary
顯著水準	0.05	顯著水準	0.05
Dickey-Fuller	-7.5623	KPSS Level	0.16257,
p-value	0.01	p-value	0.1
檢定結果	$0.01 > 0.05$ ，拒絕 $H_0$	檢定結果	$0.1 > 0.05$ ，不拒絕 $H_0$
推論	此時間序列平穩，不需在差分		

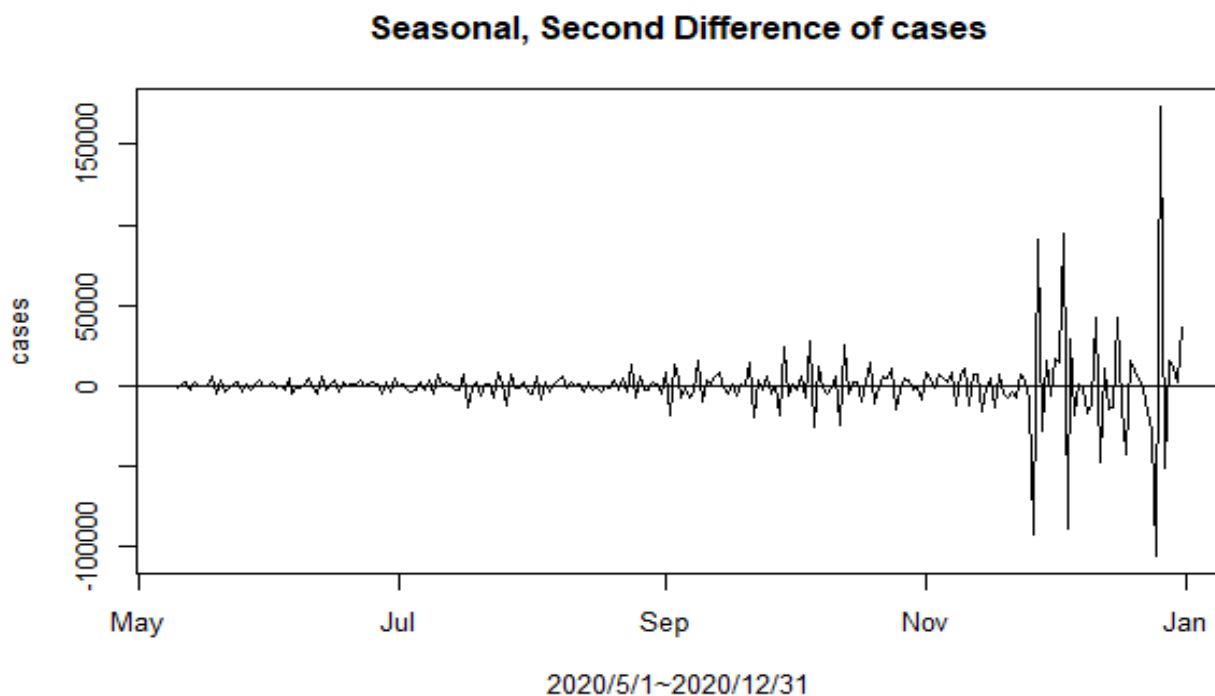
為了看是否真的為平穩，以二次差分後的時間序列圖來觀察。



從這張時間序圖可以發現沒有明顯的向上或向下趨勢，且大致平均分散在 Y 軸為 0 的上下，但可以發現它在高低點走勢幾乎相同，因此懷疑還有季節效應存在，且每七步就有相似圖形，所以再做七步差分。



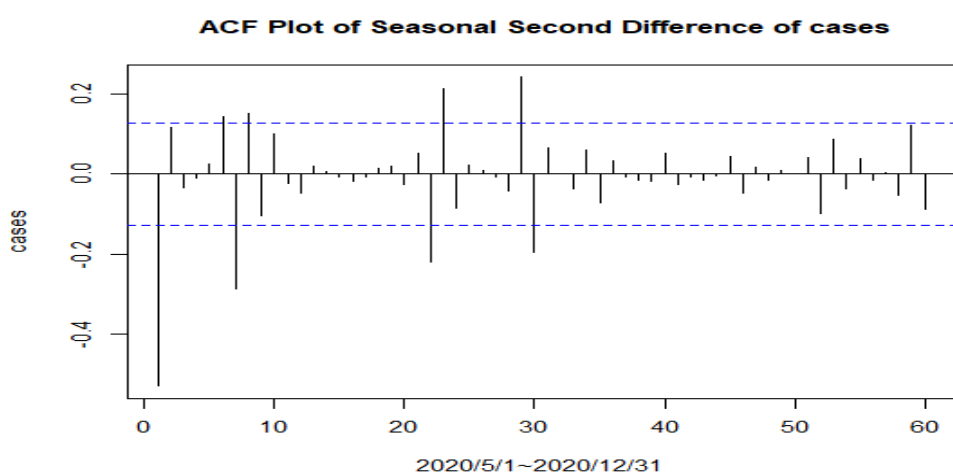
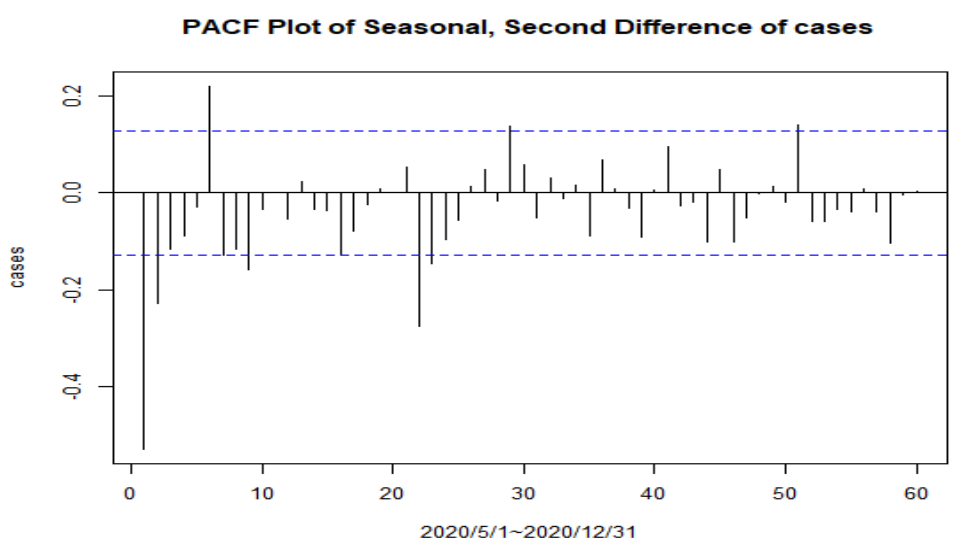
上圖為進行二次差分後的 ACF 圖，可以明顯看出它在第 7、14、21、29、35、42 步都特別顯著，且幾乎每一步有值，因此確定需要加上季節效應。



上圖為以加以 7 步差分後的时间序列圖，每一年相似的走勢已經被去除，看起來更為平穩。

## (2) SARIMA 模型配適

由第一部分得到的結論可知，仍需要進行二次差分及一次季節差分。以下為進行上述所有處理後資料的 ACF、PACF、EACF 圖。

ACF	<div><p>ACF Plot of Seasonal Second Difference of cases</p></div>	從這張圖可以得知第 7、22、29 特別顯著，且上述步數的左右一步也幾乎是顯著，但也很像每一步都有值。																																																																																	
PACF	<div><p>PACF Plot of Seasonal, Second Difference of cases</p></div>	從這張圖可以得知在第 21 步以前的數顯著，而後面逐漸收斂為 0。																																																																																	
EACF	<table><tr><th>AR\MA</th><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr><tr><td>0</td><td>x</td><td>0</td><td>0</td><td>0</td><td>0</td><td>x</td><td>x</td><td>x</td></tr><tr><td>1</td><td>x</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>x</td><td>0</td></tr><tr><td>2</td><td>x</td><td>0</td><td>x</td><td>0</td><td>0</td><td>0</td><td>x</td><td>0</td></tr><tr><td>3</td><td>x</td><td>x</td><td>0</td><td>0</td><td>0</td><td>0</td><td>x</td><td>0</td></tr><tr><td>4</td><td>x</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>x</td><td>0</td></tr><tr><td>5</td><td>0</td><td>x</td><td>0</td><td>0</td><td>0</td><td>0</td><td>x</td><td>0</td></tr><tr><td>6</td><td>x</td><td>x</td><td>0</td><td>x</td><td>x</td><td>0</td><td>x</td><td>0</td></tr><tr><td>7</td><td>x</td><td>x</td><td>x</td><td>0</td><td>x</td><td>x</td><td>x</td><td>0</td></tr></table>	AR\MA	0	1	2	3	4	5	6	7	0	x	0	0	0	0	x	x	x	1	x	0	0	0	0	0	x	0	2	x	0	x	0	0	0	x	0	3	x	x	0	0	0	0	x	0	4	x	0	0	0	0	0	x	0	5	0	x	0	0	0	0	x	0	6	x	x	0	x	x	0	x	0	7	x	x	x	0	x	x	x	0	由此可知在三角形內包括最多圈圈的是 ARMA(1,2,1)
AR\MA	0	1	2	3	4	5	6	7																																																																											
0	x	0	0	0	0	x	x	x																																																																											
1	x	0	0	0	0	0	x	0																																																																											
2	x	0	x	0	0	0	x	0																																																																											
3	x	x	0	0	0	0	x	0																																																																											
4	x	0	0	0	0	0	x	0																																																																											
5	0	x	0	0	0	0	x	0																																																																											
6	x	x	0	x	x	0	x	0																																																																											
7	x	x	x	0	x	x	x	0																																																																											

綜合以上的判斷，選出了以下的模型：

候選模型	AIC
SARIMA(1,2,0) × (0,1,1) <sub>7</sub>	5240.29
SARIMA(1,2,0) × (0,1,3) <sub>7</sub>	5243.6
SARIMA(1,2,1) × (0,1,1) <sub>7</sub>	5218.45

因為最後我們選擇的模型是 SARIMA(1,2,1) × (0,1,1)<sub>7</sub>，因此僅列出此模型的參數估計。

對 SARIMA(1,2,1) × (0,1,1)<sub>7</sub> 模型進行參數估計

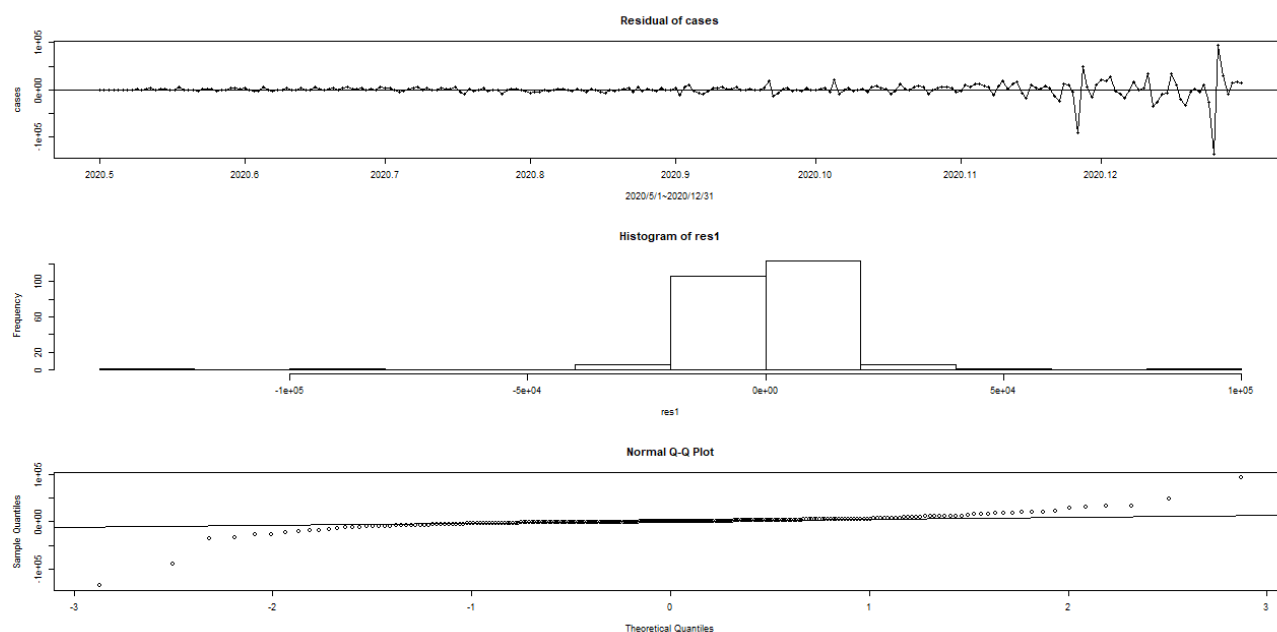
$Y_t = (2 + \phi_1)Y_{t-1} - (1 - 2\phi_1)Y_{t-2} + \phi_1 Y_{t-3} + Y_{t-7} - (\phi_1 - 2)Y_{t-8} + (1 + 2\phi_1)Y_{t-9} - \phi_1 Y_{t-10} + e_t - \Theta_1 e_{t-7} - \theta_1 e_{t-1} - \Theta_1 \theta_1 e_{t-8}, \quad e_t \sim N(0, \sigma^2)$			
係數	$\phi_1$	$\theta_1$	$\Theta_1$
估計值	0	-0.6258	-0.8078
標準誤	0	0.0475	0.0532
$\theta_1$ 與 $\Theta_1$ 的 95% 信賴區間皆未包括 0，因此這兩個參數顯著，故得到模型為 $Y_t = 2Y_{t-1} - Y_{t-2} + Y_{t-7} + 2Y_{t-8} + Y_{t-9} + e_t + 0.8078e_{t-7} + 0.6258e_{t-1} - 0.5e_{t-8},$ log likelihood = -2606.23, AIC = 5218.45			



### (3) 殘差估計

檢定殘差

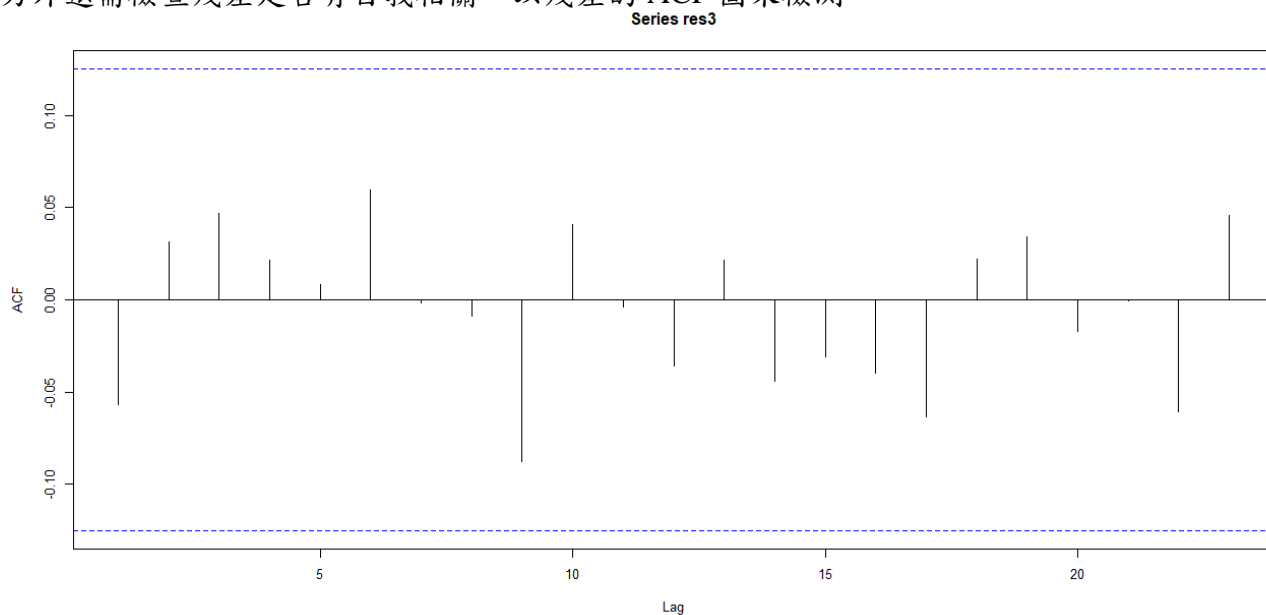
以 SARIMA (1,2,1)×(0,1,1)<sub>7</sub> 為例



取上面三張圖分別為模型的殘差時間序列圖、殘差的直方圖、QQ-plot。從殘差的時間序列圖發現除了 11 月底及 12 月份可以發現有較明顯的離群值以外，其他月份都平均分布在 0 附近。

殘差的常態性檢定由直方圖與 QQ-plot 來檢測。由直方圖可推測殘差的平均值約為 0；由 QQ-plot 的左側與右側有許多點皆沒有落在直線上，有厚尾的現象，故推測此時間序列有 GARCH 效應。

另外還需檢查殘差是否有自我相關，以殘差的 ACF 圖來檢測



由圖可知沒有超出信賴界，故符合殘差無序列相關之假設。

對候選模型進行 T-test 與 Shapiro-Wilk normality test。

	T-test	Shapiro-Wilk normality test
$H_0$	$\mu=0$	殘差分布符合常態
$H_a$	$\mu \neq 0$	殘差分布不符合常態

候選模型	AIC	t-test	結果	SW TEST	結果	ACF 序列相關
SARIMA(1,2,0)×(0,1,1) <sub>7</sub>	5240.29	0.43113	接受 $H_0$	0.59579	拒絕 $H_0$	有序列相關
SARIMA(1,2,0)×(0,1,3) <sub>7</sub>	5243.6	0.43113	接受 $H_0$	0.59579	拒絕 $H_0$	有序列相關
SARIMA(1,2,1)×(0,1,1) <sub>7</sub>	5218.45	0.64758	接受 $H_0$	0.61115	拒絕 $H_0$	無序列相關

下一部份為配飾加入離群值的模型比較

由第七步差分後的時間序列圖可以發現，在 11 及 12 月似乎有異常值，所以檢測以上各個模型是否有 IO 或 AO 的存在。

候選模型	AO	IO
SARIMA(1,2,0)×(0,1,1) <sub>7</sub>	210,215,216,218,225, 230~233,238~240	210,211,225,226,230,233 239,240
SARIMA(1,2,0)×(0,1,3) <sub>7</sub>	194,210,215,216,218, 225,230~233,238~240	210,211,217,225,226,230 239,240
SARIMA(1,2,1)×(0,1,1) <sub>7</sub>	187,210,218,225,231,232 238~240,243,244	210,217,225,233 239,240

因為由 detectAO 與 detectIO 偵測到的臨界值太多，因此選擇利用時間序列圖來觀察有哪些離群值。

候選模型	T test	SW test	ACF	AIC
①SARIMA(1,2,1)×(0,1,1) <sub>7</sub> + IO(210,211,239,240)	接受 $H_0$	拒絕 $H_0$	有序列相關	4929.25
②SARIMA(1,2,1)×(0,1,1) <sub>7</sub> + IO(210)	接受 $H_0$	拒絕 $H_0$	無序列相關	5149.25
③SARIMA(1,2,1)×(0,1,1) <sub>7</sub> + AO(210,211,239,240)	接受 $H_0$	拒絕 $H_0$	有序列相關	4988.25
④SARIMA(1,2,1)×(0,1,1) <sub>7</sub> + AO(210)	接受 $H_0$	拒絕 $H_0$	無序列相關	5180
⑤SARIMA(1,2,1)×(0,1,1) <sub>7</sub> + IO(240)	接受 $H_0$	拒絕 $H_0$	無序列相關	5185.23

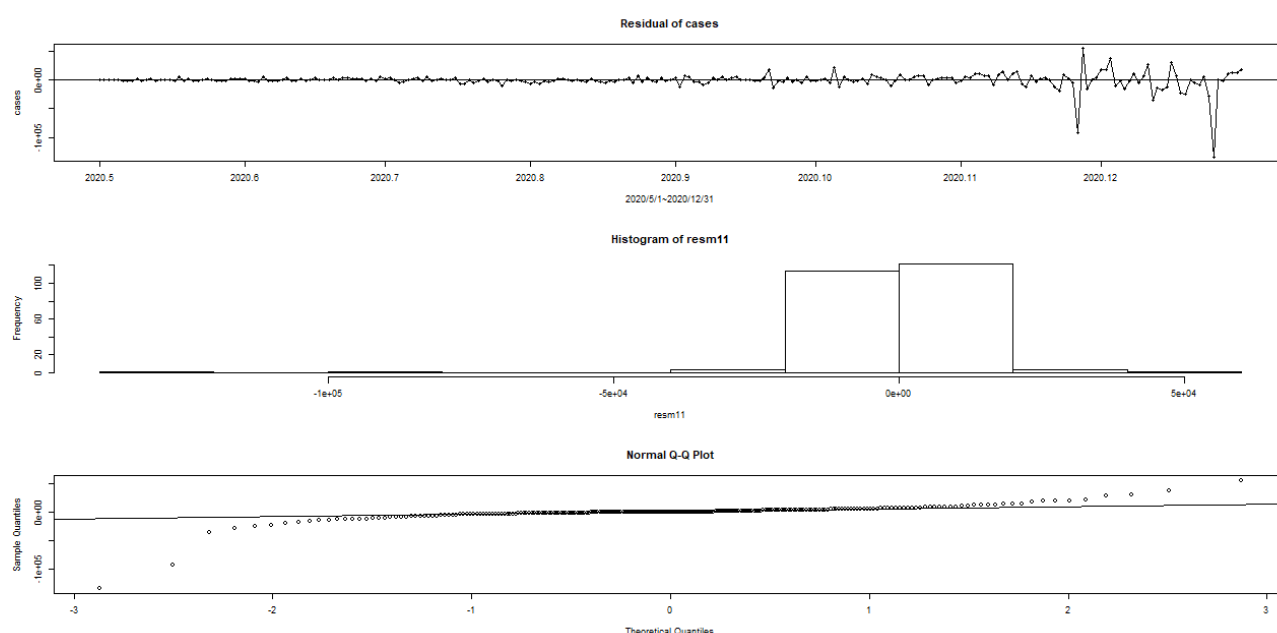
由於②③⑤的 AIC 值相近，我們分別對各模型做參數估計並做 McLeod.Li test，以 SARIMA(1,2,1)×(0,1,1)<sub>7</sub> + IO(210)和 SARIMA(1,2,1)×(0,1,1)<sub>7</sub> + AO(210)來說，我們發現這兩個模型雖然要配 Garch 模型，但是最終配飾的 Garch 模型之殘差都有序列相關，因此我們不考慮。最後以 SARIMA(1,2,1)×(0,1,1)<sub>7</sub> + IO(240)做為我們最後的模型，因此下一段僅對此模型詳細說明參數估計、殘差檢定。

### SARIMA(1,2,1)×(0,1,1)<sub>7</sub>+ IO(240)模型的參數估計

因為 $\phi_1$ 為不顯著，因此先把估計值配為 0 再進行參數估計。

$Y_t = (2 + \phi_1)Y_{t-1} - (1 - 2\phi_1)Y_{t-2} + \phi_1 Y_{t-3} + Y_{t-7} - (\phi_1 - 2)Y_{t-8} + (1 + 2\phi_1)Y_{t-9} - \phi_1 Y_{t-10} + e_t - \Theta_1 e_{t-7} - \theta_1 e_{t-1} - \Theta_1 \theta_1 e_{t-8} + \text{IO.240}, \quad e_t \sim N(0, \sigma^2)$				
係數	$\phi_1$	$\theta_1$	$\Theta_1$	IO.240
估計值	0	-0.3845	-0.7066	109897.07
標準誤	0	0.0741	0.0729	18699.98
$\theta_1$ 與 $\Theta_1$ 與 IO.210 的 95%信賴區間皆未包括 0，因此這兩個參數顯著，故得到模型為 $Y_t = 2Y_{t-1} - Y_{t-2} + Y_{t-7} + 2Y_{t-8} + Y_{t-9} + e_t + 0.7066e_{t-7} + 0.3845e_{t-1} + 0.7066e_{t-8} + 109897.07 \times \text{IO.240},$ $\log \text{likelihood} = -2589.61, \quad \text{AIC} = 5185.23$				

### SARIMA(1,2,1)×(0,1,1)<sub>7</sub>+ IO(240)模型的殘差檢定



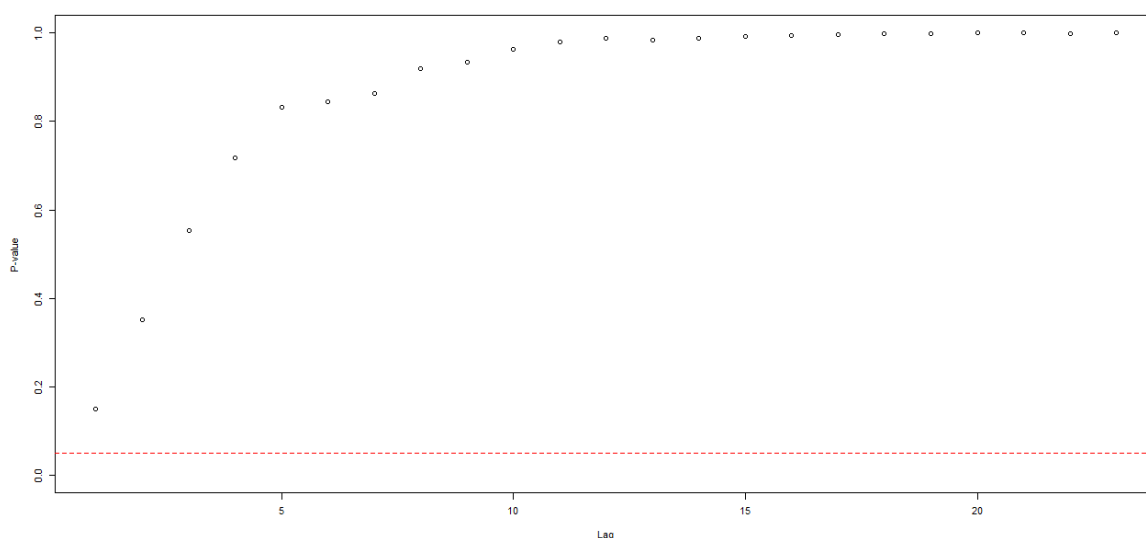
上面三張圖分別為模型的殘差時間序列圖、殘差的直方圖、QQ-plot。殘差的時間序列圖無明顯的趨勢與季節效應，但變異數在某些時間段有變大的趨勢；由直方圖可推測殘差的平均值約為 0；在 QQ-plot 的左側與右側明顯不符合常態，有厚尾的現象，故推測此時間序列可能有 ARCH 效應。

對模型進行 T-test 與 Shapiro-Wilk normality test。

	T-test	Shapiro-Wilk normality test
$H_0$	$\mu=0$	殘差分布符合常態
$H_a$	$\mu\neq 0$	殘差分布不符合常態
檢定值	-0.20158	0.5663,
p-value	0.8404	$< 2.2e-16$

對殘差做 t-test，因為  $p\text{-value} > 0.05$ ，故不拒絕  $H_0$ 。搭配上方的直方圖，我們可得殘差的平均值為 0 的結論。對殘差做 Shapiro-Wilk normality test，因為  $p\text{-value} < 0.05$ ，故不接受  $H_0$ ，搭配上方的 QQ-plot，殘差不符合常態的結論。

#### (4) McLeod.Li test



此時所有步數的 p value 都落在 0.05 之上，因此得到已經無 ARCH 效應的結論。

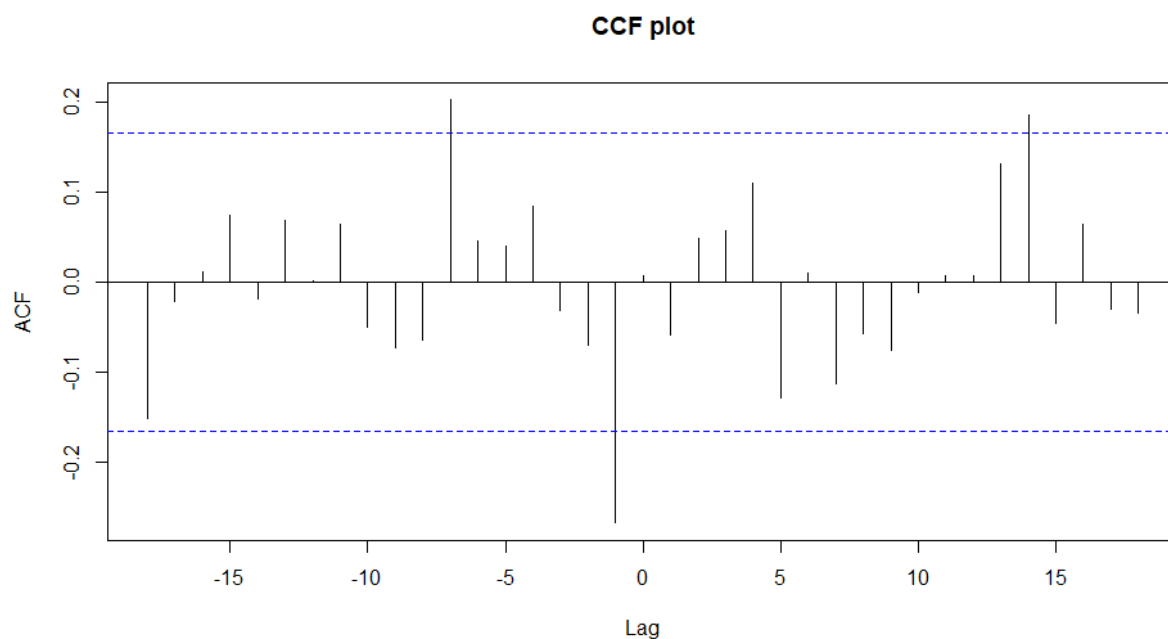
#### (5) 最終模型

$$Y_t = 2Y_{t-1} - Y_{t-2} + Y_{t-7} + 2Y_{t-8} + Y_{t-9} + e_t + 0.7066e_{t-7} + 0.3845e_{t-1} + 0.7066e_{t-8} + 109897.07 \times \text{IO.240}, \quad e_t \sim N(0, \sigma^2)$$

我們先對資料做二次差分及一次 7 步季節差分後，選擇  $\text{SARIMA}(1,2,1) \times (0,1,1)_7$  做為我們的最佳模型，接著拿掉一個離群值(IO.240)，因此我們得到以上最終的模型。

#### 四、相關性檢定

為了檢定美國新冠肺炎每日累積感染人數與 AAL 股價是否存在序列相關性，將兩原始序列資料各自配適模型後的殘差作 CCF 圖 (predictor (X) 設定為每日累積感染人數配適模型後之殘差，response (Y) 為 AAL 股價配適模型後的殘差)。



觀察 CCF 圖後發現，在  $\text{lag} = -1, -7, 14$  時的 sample ccf 超出信賴界，表示可能分別在時間  $t-1$ 、 $t-7$  的感染人數模型之殘差會顯著影響股價模型在時間  $t$  的殘差。另外，照常理來說，新冠肺炎感染人數應為股價的 predictor，若股價為感染人數的 predictor 較不符合常理( $\text{lag} = 14$  時超出信賴界)，故後續回歸模型建立不考慮  $\text{lag} = 14$  的情況。

## A. 建立回歸模型

### (1) 對 Lag = -1 建立回歸模型

$$Y_t = \beta_0 + \beta_1 X_{t-1} + e_t$$

( $X_{t-1}$  為在 t-1 時間點之感染人數模型殘差， $Y_t$  為在 t 時間點之股價模型殘差)

Coefficients	Estimated Value	Standard Error	P-value
$\beta_0$	-0.0007546	0.0033227	0.82067
$\beta_1$	-1.6707576	0.51164	0.00138

由於發現模型截距項參數估計不顯著(p 值遠大於 0.05)，故拿掉截距項後再進行一次配適

$$Y_t = \beta_1 X_{t-1} + e_t$$

Coefficients	Estimated Value	Standard Error	P-value
$\beta_1$	-1.6656	0.5094	0.00136
Multiple R-squared = 0.07191, Adjusted R-squared= 0.06518			

此時參數估計顯著不為 0 (p 值小於 0.05)，故由參數估計為負值以及 CCF 圖可知，在**前一天**時間點的感染總人數模型殘差與當天股票模型殘差有顯著的**負向關係**。

(2) 對 Lag = -7 建立回歸模型

$$Y_t = \beta_0 + \beta_1 X_{t-7} + e_t$$

( $X_{t-7}$  為在 t-7 時間點之感染人數模型殘差， $Y_t$  為在 t 時間點之股價模型殘差)

Coefficients	Estimated Value	Standard Error	P-value
$\beta_0$	0.001008	0.003467	0.7718
$\beta_1$	1.382762	0.546051	0.0125

由於發現模型截距項參數估計不顯著(p 值遠大於 0.05)，故拿掉截距項後再進行一次配適

$$Y_t = \beta_1 X_{t-7} + e_t$$

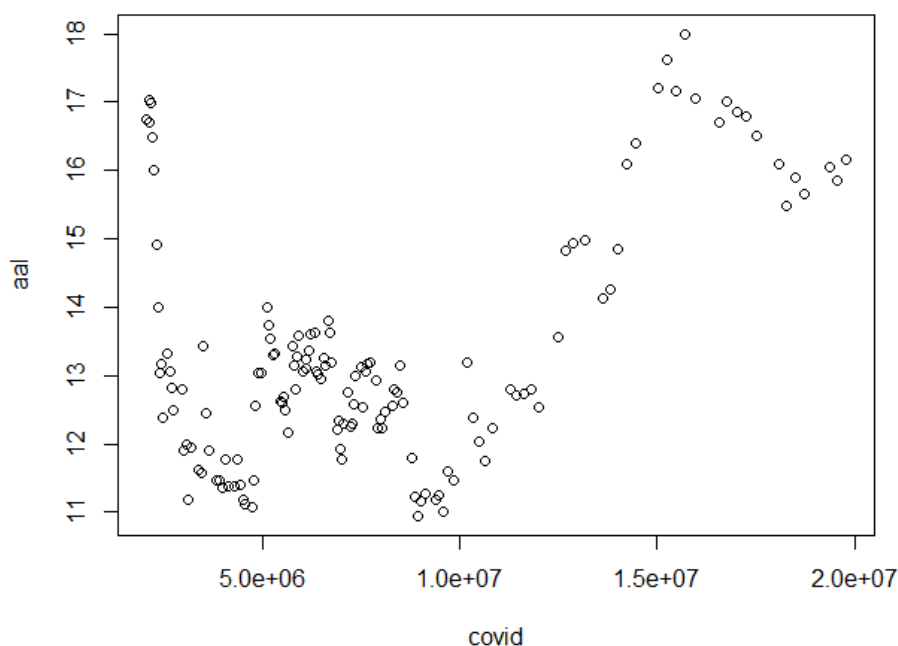
Coefficients	Estimated Value	Standard Error	P-value
$\beta_1$	1.3750	0.5435	0.0126
Multiple R-squared = 0.04625, Adjusted R-squared= 0.03902			

此時參數估計也顯著不為 0 (p 值小於 0.05)，故由參數估計為正值以及 CCF 圖可知，感染總人數模型殘差在**前七天**時間點與股票模型殘差有顯著的**正向關係**。



## 五、結論

新冠肺炎總感染人數與美國航空股價散布圖(皆為原始資料)



為驗證結論，將新冠肺炎感染人數與美國航空股價原始資料作散布圖後可以觀察到，前期感染人數與股價具有負向的關係，而後期兩者之間較多有正向的關係。原對 CCF 圖的結果感到困惑(感染人數前一步和前七步居然和股價分別有負向和正向的關係)，但比對原始資料散步圖的結果，解決了我們原本的困惑。我們原本假設感染人數應與航空股價呈現高度負相關，但相關性檢定中 Lag= -1, -7 的兩個回歸模型之 R-squared 皆非常低(約為 7% 左右)，我們結論是美國航空集團股價(AAL)確實與前一天美國新冠肺炎總感染人數呈現負相關，且與前七天總感染人數呈現正相關，但是解釋力並不高，美國航空集團股價(AAL)的變動由其他更多的因素所主導。

## 六、學習心得

莊芯瑜：

我修完這堂課後，我的整體評價是，可以學到很多新知識，又有實作分析資料，老師上課的節奏輕快，我覺得最棒的是，當我們真正在分析真實資料時，老師願意花時間跟我們解說，我覺得這反而是這堂課可以讓我整個融會貫通的價值。所以如果是認真且想學習新知的同學，我一定會建議他修！

此外，我比較例外一點(?)，我真的很喜歡老師每個禮拜派作業，雖然我也是花很多時間在寫作，但因為每個禮拜的作業讓我更有計畫且規律性地在複習時間序列，尤其期中考前那九周，平均分配掉我念時序的時間，加上助教非常盡責地為我解惑時間序列的功課，所以期中考時也就有不錯的表現。

最後，建議如果是外系同學要修這堂課，還是要有一點統計背景再來修，會比較清楚知道老師的意思，另外，這個課期末的大 project 一定要慎選組員，而我也要非常感謝我的組員，在沒有思考方向時，總會透過討論得到最佳解，所以我真的非常喜歡這堂課，我也非常認真修習，最後謝謝老師和助教的辛苦與指導！

賴廷瑋：

這堂課是目前收穫最多的課之一，雖然 loading 重，但真的學得很扎實。我覺得收穫最多的並不是了解時間序列的模型有哪些，反而是對時間序列資料概念的掌握，這部分老師解釋的也很清楚。例如我這學期在另一門深度學習的課程也針對股票建立模型，因為修了時序這門課，就了解要針對平穩的資料建模、時間序列之間的相關性並非是直接看原始資料的相關性…等基本的概念。未來碰到時間序列資料時，該做哪些處理、要注意哪些事有更深的了解，另外我很喜歡這門課的教材，每個單元都會有實際的 R code，而且概念解釋簡單有力。

最後謝謝老師，也謝謝超 carry 的組員們！

## 七、課堂建議

老師您好，經過我們組別討論，我們認為課堂建議第一，希望助教時間可以安排在上課後幾天，因為這學期面對的狀況是，老師禮拜三才把課上完，每次可能可以開始寫作業的時間為禮拜三晚上或四，這樣都沒辦法當天禮拜三去問老師問題。建議第二，我覺得老師在講每個章節的時候可能可以先講一些這個章節的大方向，這樣可能會比較有概念知道老師接下來要講的內容之連貫性。

# 八、期中考加分題

$$\epsilon_t \sim N(0, \sigma^2).$$

$$\text{ARMA}(1,2) \text{ Model} = Y_t = 0.8 Y_{t-1} + \epsilon_t + 0.7 \epsilon_{t-1} + 0.6 \epsilon_{t-2}.$$

$$\Leftrightarrow Y_t = \phi Y_{t-1} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} \Rightarrow \begin{cases} \phi = 0.8 \\ \theta_1 = -0.7 \\ \theta_2 = -0.6. \end{cases}$$

by back shifting,

$$(1 - \phi B) Y_t = (1 - \theta_1 B - \theta_2 B^2) \epsilon_t.$$

$$\Rightarrow Y_t = \frac{(1 - \theta_1 B - \theta_2 B^2)}{1 - \phi B} \epsilon_t = \left[ \sum_{\lambda=0}^{\infty} (\phi B)^{\lambda} - \theta_1 \sum_{\lambda=0}^{\infty} (\phi B)^{\lambda+1} - \theta_2 \sum_{\lambda=0}^{\infty} (\phi B)^{\lambda+2} \right] \epsilon_t \quad \text{et } \uparrow$$

$$\text{其中 } \sum_{\lambda=0}^{\infty} (\phi B)^{\lambda} = 1 + \phi B + \phi^2 B^2 + \phi^3 B^3 + \dots \quad - ①$$

$$\left\{ \begin{aligned} -\theta_1 \sum_{\lambda=0}^{\infty} (\phi^{\lambda+1} B^{\lambda+1}) &= -\theta_1 B - \theta_1 \phi B^2 - \theta_1 \phi^2 B^3 - \theta_1 \phi^3 B^4 - \dots \quad - ② \\ -\theta_2 \sum_{\lambda=0}^{\infty} (\phi^{\lambda+2} B^{\lambda+2}) &= -\theta_2 B^2 - \theta_2 \phi B^3 - \theta_2 \phi^2 B^4 - \theta_2 \phi^3 B^5 - \dots \quad - ③ \end{aligned} \right.$$

將 ①+②+③, 得

$$Y_t = \left[ 1 + (\phi - \theta_1) B + \sum_{\lambda=2}^{\infty} (\phi^{\lambda} - \theta_1 \phi^{\lambda-1} - \theta_2 \phi^{\lambda-2}) B^{\lambda} \right] \epsilon_t. \quad - ④$$

$$\psi_{\lambda} = \begin{cases} 1, & \lambda=0. \\ \phi - \theta_1, & \lambda=1 \\ \phi^{\lambda} - \theta_1 \phi^{\lambda-1} - \theta_2 \phi^{\lambda-2}, & \lambda \geq 2. \end{cases}$$

見下頁.

令  $l$  為 forecast horizon, 則利用  $\oplus$  得:

$$Y_{t+l} = \psi_0 e_{t+l} + \psi_1 e_{t+l-1} + \psi_2 e_{t+l-2} + \psi_3 e_{t+l-3} = \sum_{\lambda=0}^{\infty} \psi_{\lambda} e_{t+l-\lambda}.$$

令  $\hat{Y}_t(l)$  為  $l$  step ahead forecast, 則

$$\hat{Y}_t(l) = E(Y_{t+l} | Y_t, Y_{t-1}, Y_{t-2}, \dots, Y_1).$$

$$\times E(e_{t+j} | Y_t, Y_{t-1}, Y_{t-2}, \dots, Y_1) = \begin{cases} 0, & j > 0 \\ e_{t+j}, & j \leq 0. \end{cases}$$

$$\text{則 } \hat{Y}_t(l) = \sum_{\lambda=0}^{\infty} \psi_{\lambda} e_{t+l-\lambda}.$$

$$\text{則 forecast error } e_t(l) = Y_{t+l} - \hat{Y}_t(l) = \sum_{\lambda=0}^{l-1} \psi_{\lambda} e_{t+l-\lambda}.$$

故針對  $Y_t = 0.8Y_{t-1} + e_t - 0.7e_{t-1} - 0.6e_{t-2}$ , 第  $l$  步的 forecast error 之 variance 為

$$\text{Var}(e_t(l)) = \sum_{\lambda=0}^{l-1} \psi_{\lambda}^2 \cdot \text{var}(e_{t+l-\lambda}) = \sigma^2 \sum_{\lambda=0}^{l-1} \psi_{\lambda}^2 \quad \#.$$

$$\text{其中 } \psi_{\lambda}^2 = \begin{cases} 1, & \lambda=0 \\ (0.8+0.7)^2, & \lambda=1 \\ (0.8^{\lambda} + 0.7 \times 0.8^{\lambda-1} + 0.6 \times 0.8^{\lambda-2})^2, & \lambda \geq 2 \end{cases}$$

## 九、額外建議

莊芯瑜：

我覺得前面幾章或許可以縮短一些證明的部分，這樣也許最後就可以有五周講最後三章，那這樣就可以仿照前面先介紹模型，在找跑程式時就可以更加瞭解背後的整個理論。此外，如果可以，也希望老師可以除了介紹模擬的例子，也可以多介紹真實的例子運用在 Intervention, Garch 等等。

吳育儒：

我想或許可以建議系辦把時序的課程排成一天三小時的課，而非兩天的課程，或許這樣能夠讓講解更有效率一點。如此就能解決後半部無法上到的部份。