ARIMA季節模型案例

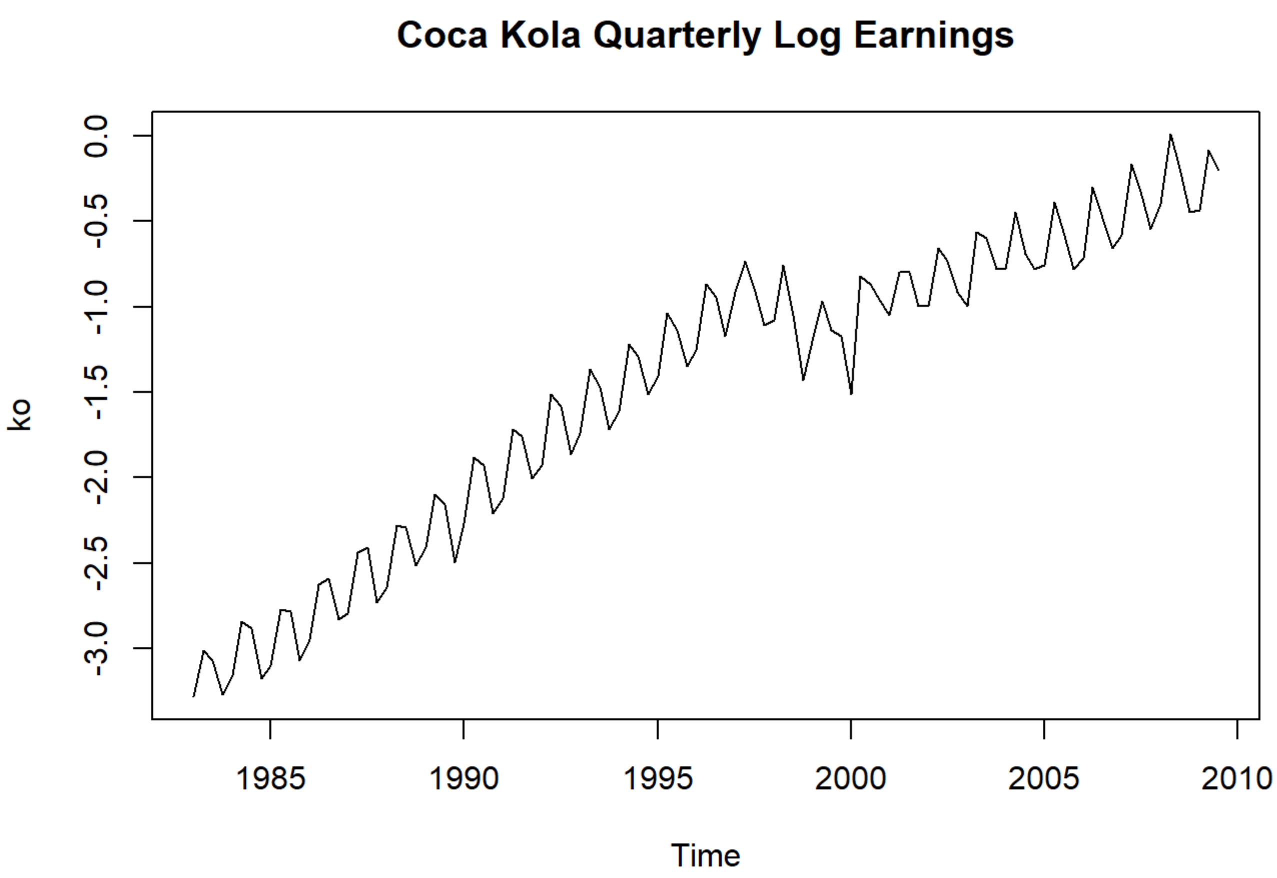
SJLin

2021/11/28

- 季節差分
- 乘性季節模型
- 季節虛擬變數

經濟和金融中的月資料、季資料一般有明顯的週期， 日資料也會有按照周、月、年週期的變化。 這樣的性質稱為季節性， 含有週期變化的時間序列稱為季節時間序列。如：可口可樂公司1983第1季到2009第3季公佈的季盈利資料。 每季的盈利資料在季結束後約一個月以後公佈。共107個觀測。 考慮季盈利的對數值，其時間序列圖：

```
load("ko.Rdata")
plot(ko, type="l", main="Coca Kola Quarterly Log Earnings")
```



該圖呈現出明顯的週期為4的波動。 如果是月資料，週期為12。 盈利做了對數變換， 其中一個理由是消除指數增長（倍數增長）現象， 對數變化可以將指數增長變成線性增長。

農產品等與天氣有關的衍生產品定價、能源期貨定階等與天氣有關的金融產品研究中， 季節性是重要的考慮因素。

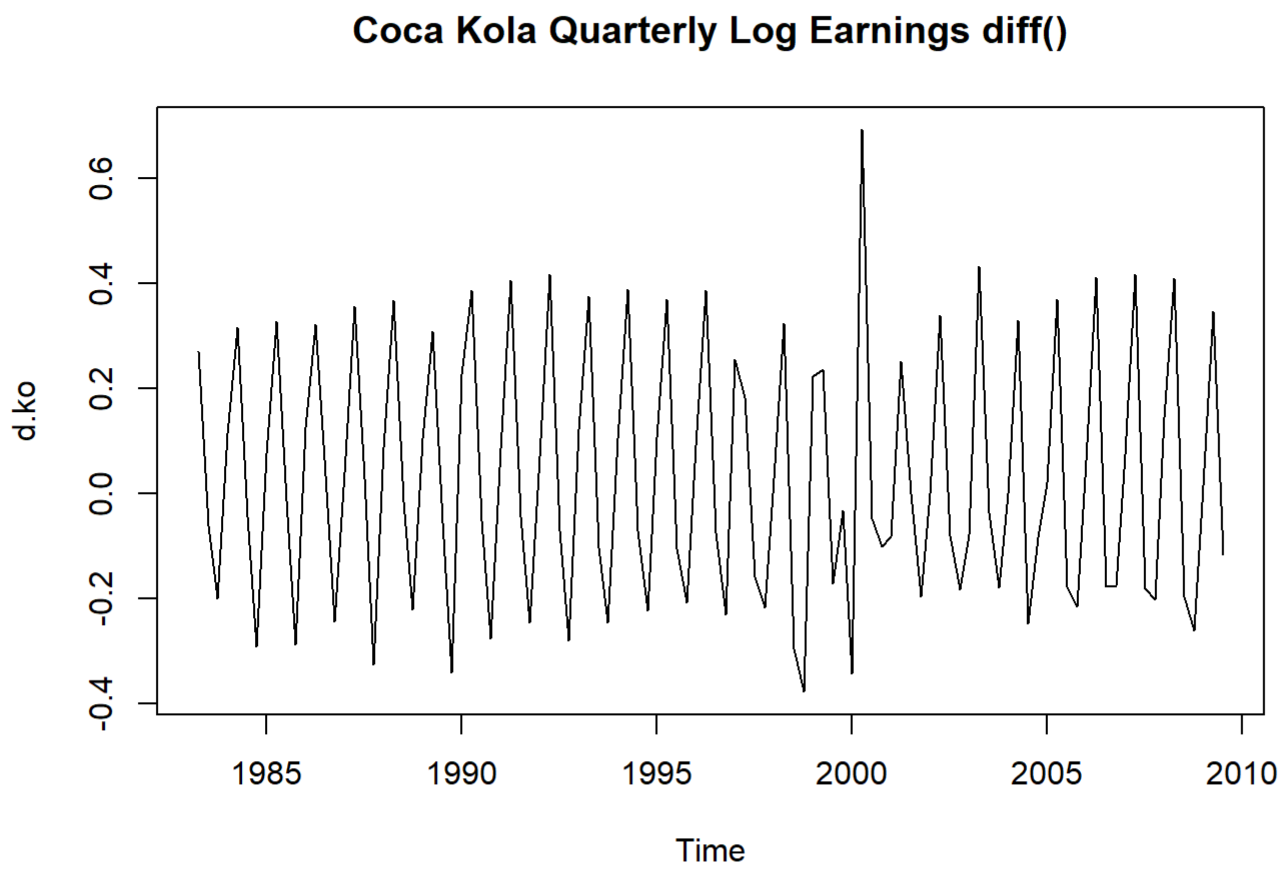
經濟研究中有時希望排除季節性影響， 用某種方法去掉季節性的序列稱為季節調整序列(seasonally adjusted series)， 如美國季節調整的GNP序列。 X12-ARIMA是一個常用的的季節調整方法。

季節因素的建模， 也包括動態模型（類似於隨機游動）和非隨機模型（類似於固定線性趨勢模型）。

季節差分

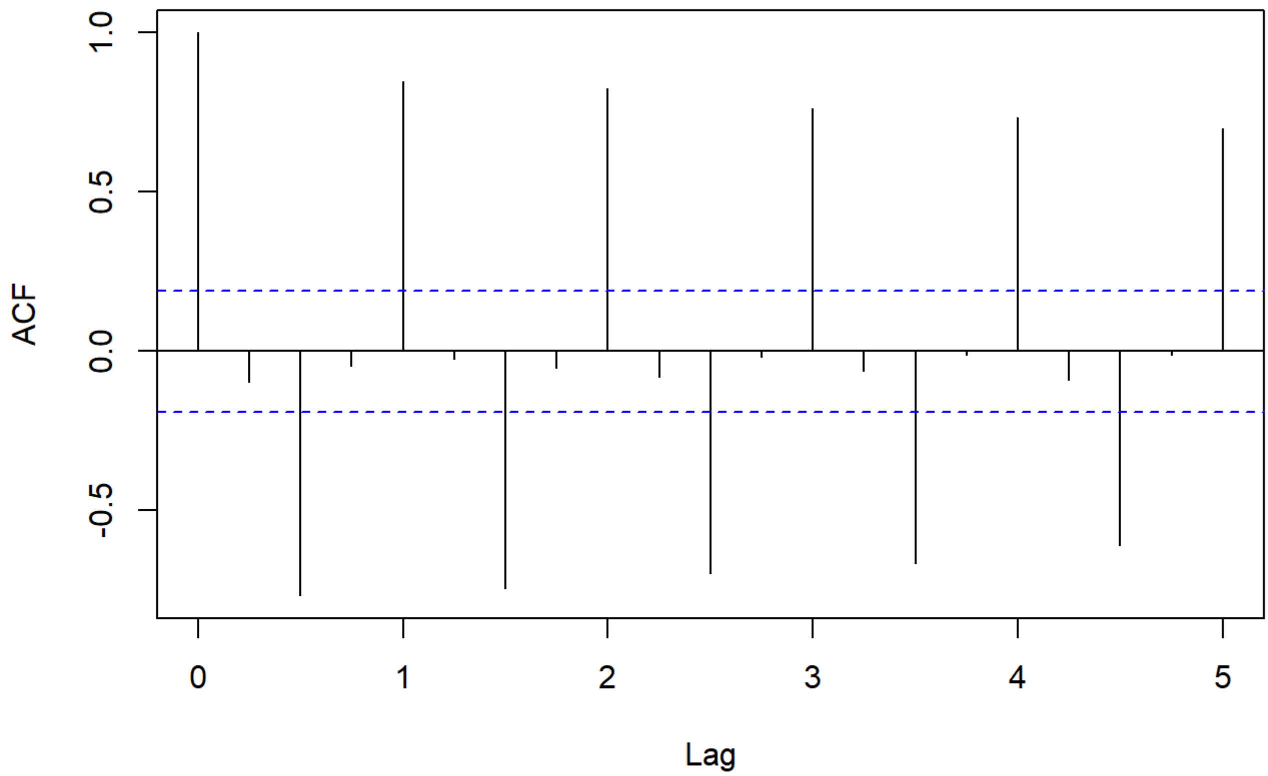
因為圖形呈現出線性增長，明顯是不平穩的，所以考慮其一階差分：

```
d.ko <- diff(ko)
plot(d.ko, type="l", main="Coca Kola Quarterly Log Earnings diff()")
```



一階差分後，消除了線性增長趨勢， 但是還有明顯的季節性， 這往往也是不平穩的表現， 看其ACF:

```
d.ko <- diff(ko)
acf(d.ko, main="")
```



消除季節性影響的方法是， 第二年第一季資料減去第一年第一季資料， 第二年第二季資料減去第一年第二季資料， 第二年第三季資料減去第一年第三季資料， 第二年第四季資料減去第一年第四季資料，這稱為季節差分，實際是計算$(1-L^4)x_t = x_t - x_{t-4}$。R中用 `diff(x, lag=4)` 計算這樣的季節差分。 例如：

```
tmp <- window(ko, start=c(1983,1), end=c(1985,4)); tmp
```

```
##           Qtr1      Qtr2      Qtr3      Qtr4
## 1983 -3.283414 -3.011862 -3.072613 -3.272804
## 1984 -3.158251 -2.842153 -2.885981 -3.177254
## 1985 -3.101093 -2.772589 -2.785471 -3.072613
```

```
diff(tmp, lag=4)
```

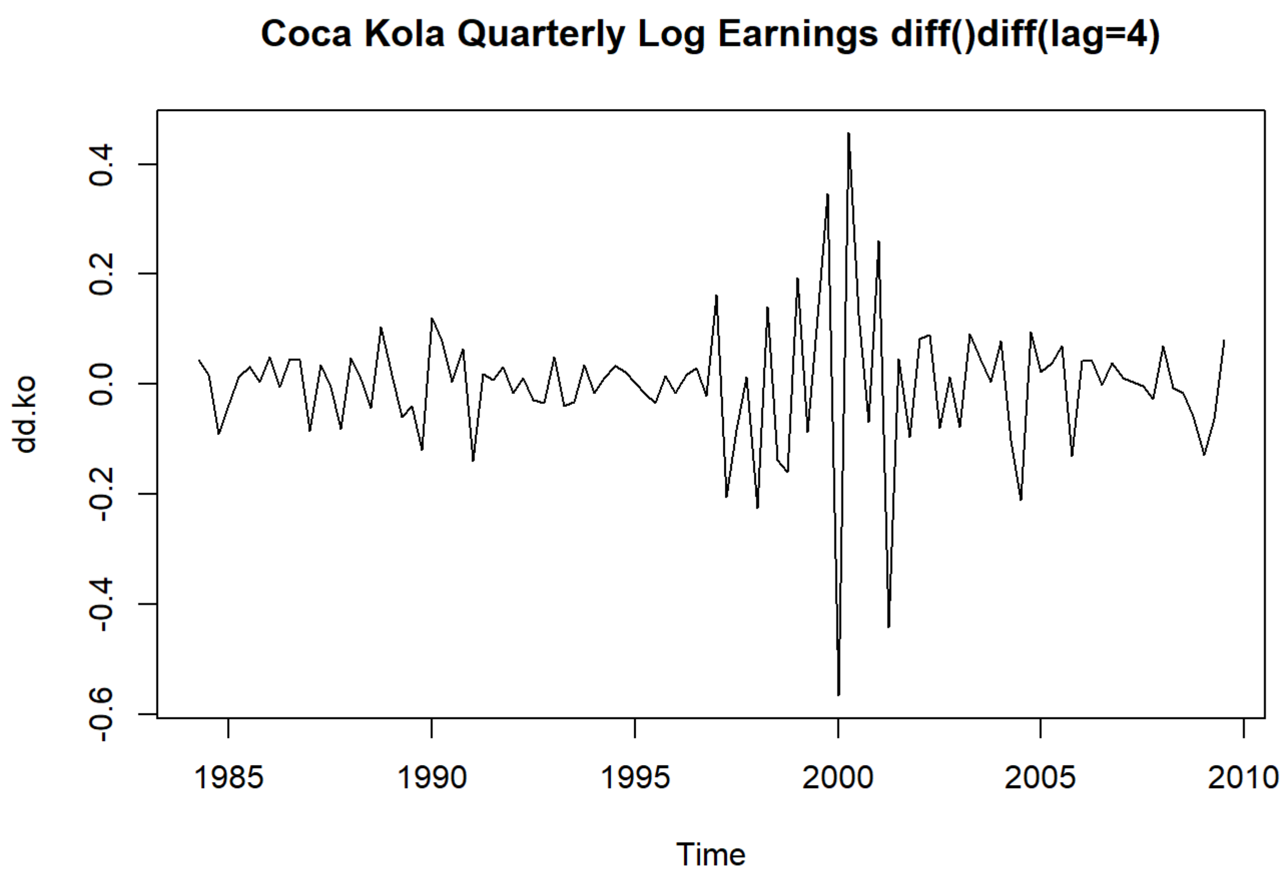
```
##            Qtr1       Qtr2       Qtr3       Qtr4
## 1984 0.12516314 0.16970847 0.18663191 0.09555002
## 1985 0.05715841 0.06956446 0.10051006 0.10464083
```

先做一階差分$(1-L)x_t$， 然後再做一季節差分$(1-L^4)(1-L)x_t = x_t - x_{t-1} - x_{t-4} + x_{t-5}$， 得到新的序列$\{y_t\}$：

```
dd.ko <- diff(diff(ko), lag=4)
str(dd.ko)
```

```
##  Time-Series [1:102] from 1984 to 2010: 0.0445 0.0169 -0.0911 -0.0384 0.0124 ...
```

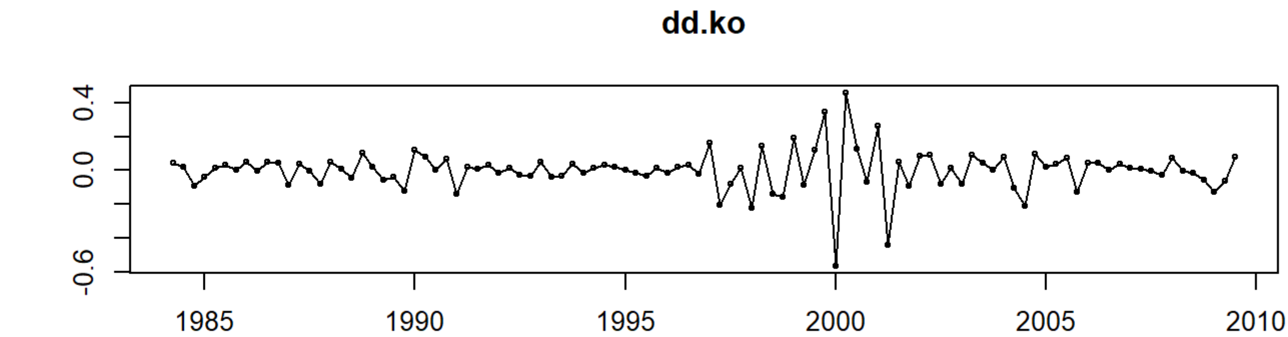
```
plot(dd.ko, main="Coca Kola Quarterly Log Earnings diff()diff(lag=4)")
```

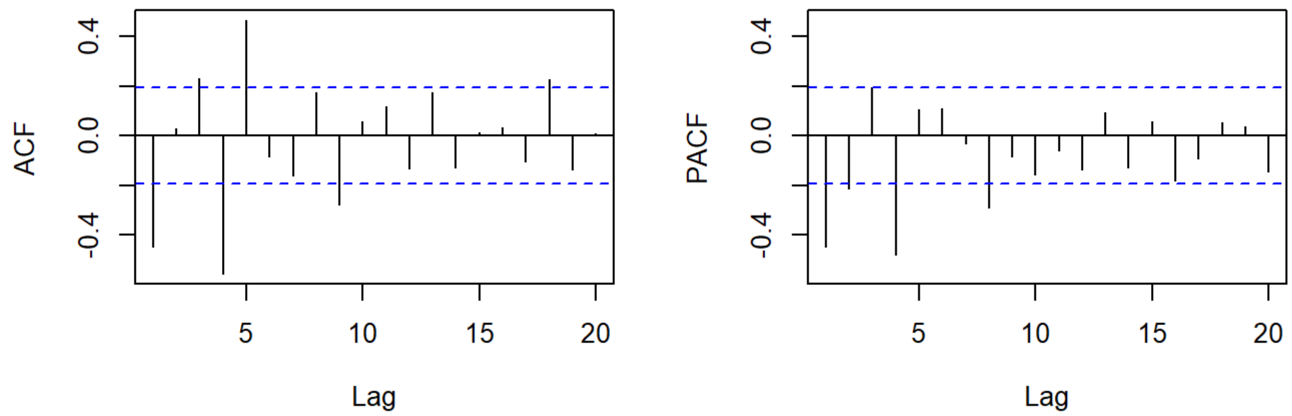


因為需要用到x_{t-5}，所以變換後的序列從原來的第6個觀測開始。從序列圖形看已經消除了線性趨勢與大部分季節波動。

查看一階差分與一階季差分後的序列的ACF:

```
tsdisplay(dd.ko)
```





此ACF的衰減已經比僅作一階差分的序列的ACF衰減快多了， 比較適合ARMA序列的ACF形狀。 注意ACF則落後1和落後4（落後4對應季節波動）都是顯著的負值， 這與一階差分和季差分有關。 ACF在落後5也是顯著的正值， 部分理由是$(1 - L^4)(1 - L)$中包含L^5項。 這些ACF表現是許多經過一階差分和季節差分的序列共同的特徵。

乘性季節模型

季節性的時間序列經過一階差分和季節差分後的序列常常仍在落後1、4、5這些位置呈現出自我相關 （如果是月資料，則為落後1、12、13）。 為此，對差分後序列建立MA序列，這樣，$\{x_t\}$的模型為

$$(1 - L)(1 - L^s)x_t = (1 - \theta L)(1 - \Theta L^s)\varepsilon_t$$

其中s是週期，對上面的季數據$s = 4$。 這個模型稱為航空模型， 最早用在分析航空乘客數月資料當中。

設Y_t為右邊的MA($s + 1$模型)：

$$Y_t = (1 - \theta L)(1 - \Theta L^s)\varepsilon_t = \varepsilon_t - \theta\varepsilon_{t-1} - \Theta\varepsilon_{t-s} + \theta\Theta\varepsilon_{t-s-1}$$

易見$EY_t = 0$,

$$\begin{aligned}
\gamma_0 =& \text{Var}(Y_t) = (1 + \theta^2)(1 + \Theta^2)\sigma^2 \\
\gamma_1 =& -\theta(1 + \Theta^2)\sigma^2 \\
\gamma_{s-1} =& \theta\Theta\sigma^2 \\
\gamma_s =& -\Theta(1 + \theta^2)\sigma^2 \\
\gamma_{s+1} =& \theta\Theta\sigma^2 \\
\gamma_k =& 0,\ k \neq 0, 1, s-1, s, s+1
\end{aligned}$$

所以$\{Y_t\}$的ACF為

$$\rho_1 = \frac{-\theta}{1+\theta^2},\ \rho_s = \frac{-\Theta}{1+\Theta^2}, \rho_{s-1} = \rho_{s+1} = \rho_1\rho_s = \frac{\theta\Theta}{(1+\theta^2)(1+\Theta^2)}$$

其它$\rho_k = 0$。

設

$$\xi_t = (1-\theta L)\varepsilon_t, \quad \eta_t = (1-\Theta L^s)\varepsilon_t,$$

則

$$\rho_1^{(\xi)} = \frac{-\theta}{1+\theta^2}, \rho_s^{(\eta)} = \frac{-\Theta}{1+\Theta^2},$$

且$\rho^{(\xi)}$和$\rho^{(\eta)}$在其它落後的值為零。 比較$\{Y_t\}$的ACFρ_k和$\rho^{(\xi)}$、$\rho^{(\eta)}$發現：

$$\rho_1 = \rho_1^{(\xi)},\ \rho_s = \rho_s^{(\eta)},\ \rho_{s-1} = \rho_{s+1} = \rho_1^{(\xi)}\rho_s^{(\eta)}$$

其中Y_t的ACF在落後$s \pm 1$的值可以認為是MA(1)和MA(4)的ACF的交互作用。 $\{Y_t\}$的模型稱為乘性季節MA模型(multiplicative MA model)。 可以認為序列的同一季節兩年之間的變化與相鄰兩個季節之間的變化時正交變化的。

航空模型與指數平滑也有關係。 將原來的模型改寫為

$$\frac{1-L}{1-\theta L}\left(\frac{1-L^s}{1-\Theta L^s}x_t\right) = \varepsilon_t$$

則可將其寫成兩步：

$$\frac{1-L}{1-\theta L}z_t = \varepsilon_t, \quad \frac{1-L^s}{1-\Theta L^s}x_t = z_t$$

第一個模型是一個ARIMA(0,1,1)模型， 對應於指數平滑； 第二個模型的輸入時第一個模型的序列， 對應於季節s的ARIMA$(0,1,1)_s$。 所以航空模型可以看成是兩個指數平滑的復合。

比航空模型略複雜一點兒的模型是增加一個落後2的MA項，即

$$(1-L)(1-L^s)x_t = (1-\theta_1 L - \theta_2 L^2)(1-\Theta L^s)\varepsilon_t$$

在$s > 4$較常出現。?resm 這時ACF不等於零的落後位置是$1, 2, s, s\pm 1, s\pm 2$。

另一個非乘積形式的帶有季節因素的MA模型是

$$X_t = (1-\theta L - \Theta L^s)\varepsilon_t$$

但是乘積形式的更常見。

在 `arima()` 函數中， 可以用 `seasonal=` 指定季節模型，包括季節AR階、季節差分階、季節MA階以及週期。 如

```
resm <- arima(
  ko, order=c(0,1,1), seasonal=list(order=c(0,1,1), period=4)
); resm
```

```
##
## Call:
## arima(x = ko, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 4))
##
## Coefficients:
##           ma1     sma1
##       -0.4096  -0.8203
## s.e.   0.0866   0.0743
##
## sigma^2 estimated as 0.00724:  log likelihood = 104.25,  aic = -202.5
```

結果模型為

$$(1 - L)(1 - L^4)x_t = (1 - 0.4096L)(1 - 0.8203L^4)\varepsilon_t$$

$\hat{\sigma}^2 = 0.00724$。注意 `arima()` 函數給出的MA多項式是$1 + \theta_1 L + \cdots + \theta_q L^q$形式的。

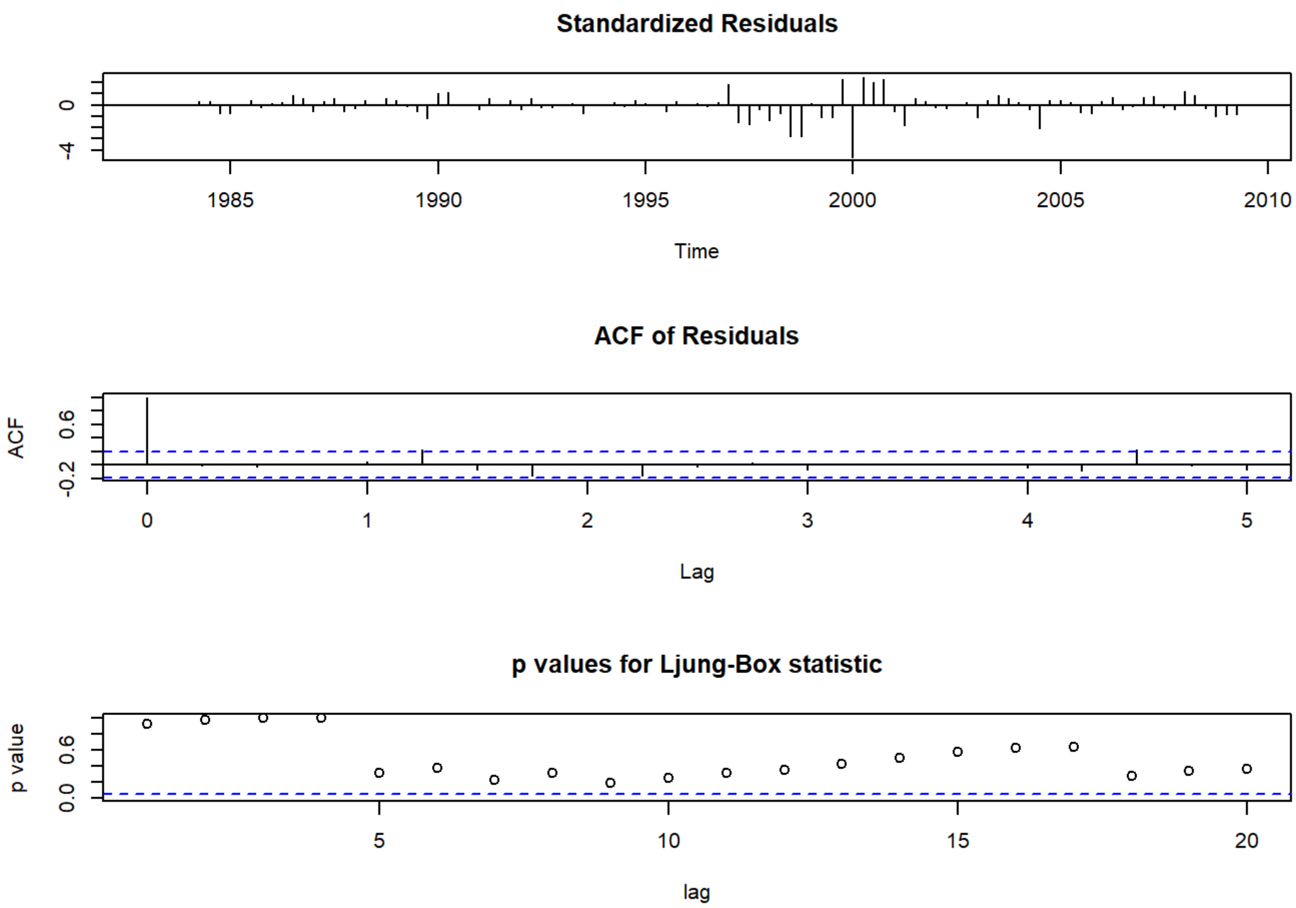
對模型殘差做 `LB` 白噪音 檢驗：

```
Box.test(resm$residuals, lag=12, fitdf=2)
```

```
##
##  Box-Pierce test
##
## data:  resm$residuals
## X-squared = 12.233, df = 10, p-value = 0.2698
```

結果不顯著，表示承認模型合適。 `arima()` 的結果還可以用 `tsdiag()` 函數檢驗殘差是否白噪音，如:

```
tsdiag(resm, gof=20)
```



其中 `gof=` 指定最大落後。診斷圖中的第一個圖是標準化殘差；第二個圖是殘差的ACF，除了落後0之外應該落入兩條水平線之間；第三個圖是不同落後值的LB白噪音檢驗的p值， 應該位於檢驗水平0.05代表的水平線上方。

下面用航空模型做超前多步預報。 用1983年到2007年這25年的100個觀測值建模， 對剩餘的7個值作超前多步預報：

```
tmp.y <- window(ko, start=start(ko), end=c(2007,4))
resm2 <- arima(tmp.y, order=c(0,1,1), seasonal=list(order=c(0,1,1), period=4))
resm2
```

```
##
## Call:
## arima(x = tmp.y, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 4))
##
## Coefficients:
##           ma1     sma1
##       -0.4209  -0.8099
## s.e.   0.0874   0.0767
##
## sigma^2 estimated as 0.007432:  log likelihood = 95.78,  aic = -185.57
```

```
pred1 <- predict(resm2, n.ahead=7)
cbind(Observed=ko[101:107], Predicted=pred1$pred, SE=pred1$se)
```

```
##              Observed  Predicted         SE
## 2008 Q1 -0.400477567 -0.5060620 0.08621248
## 2008 Q2  0.009950331 -0.1237792 0.09962409
## 2008 Q3 -0.186329578 -0.2669296 0.11143307
## 2008 Q4 -0.446287103 -0.4501580 0.12210527
## 2009 Q1 -0.430782916 -0.4219704 0.13894879
## 2009 Q2 -0.083381609 -0.0396876 0.15111786
## 2009 Q3 -0.198450939 -0.1828380 0.16237749
```

這是每股季盈利的對數值的預測。 但是，如何得到每股季盈利的預測？ 為了得到不偏估計， 需要利用對數常態分配的性質進行期望的校正。 如果隨機變量$\ln Y \sim \mathrm{N}(\mu, \sigma^2)$， 則$EY = \exp(\mu + \frac{1}{2}\sigma^2)$， $\mathrm{Var}(Y) = \exp(2\mu + \sigma^2)(e^{\sigma^2} - 1)$。

據此得到關於每股季盈利的預測：

```
pred2 <- exp(pred1$pred + 0.5*pred1$se^2)
se2 <- sqrt(exp(2*pred1$pred + pred1$se^2)*(exp(pred1$se^2)-1))
cbind(Observed=c(coredata(xts.koqtr))[101:107], Predicted=round(pred2, 2), SE=round(se2, 2))
```

```
##         Observed Predicted   SE
## 2008 Q1     0.67      0.61 0.05
## 2008 Q2     1.01      0.89 0.09
## 2008 Q3     0.83      0.77 0.09
## 2008 Q4     0.64      0.64 0.08
## 2009 Q1     0.65      0.66 0.09
## 2009 Q2     0.92      0.97 0.15
## 2009 Q3     0.82      0.84 0.14
```

或者，寫成R函數：

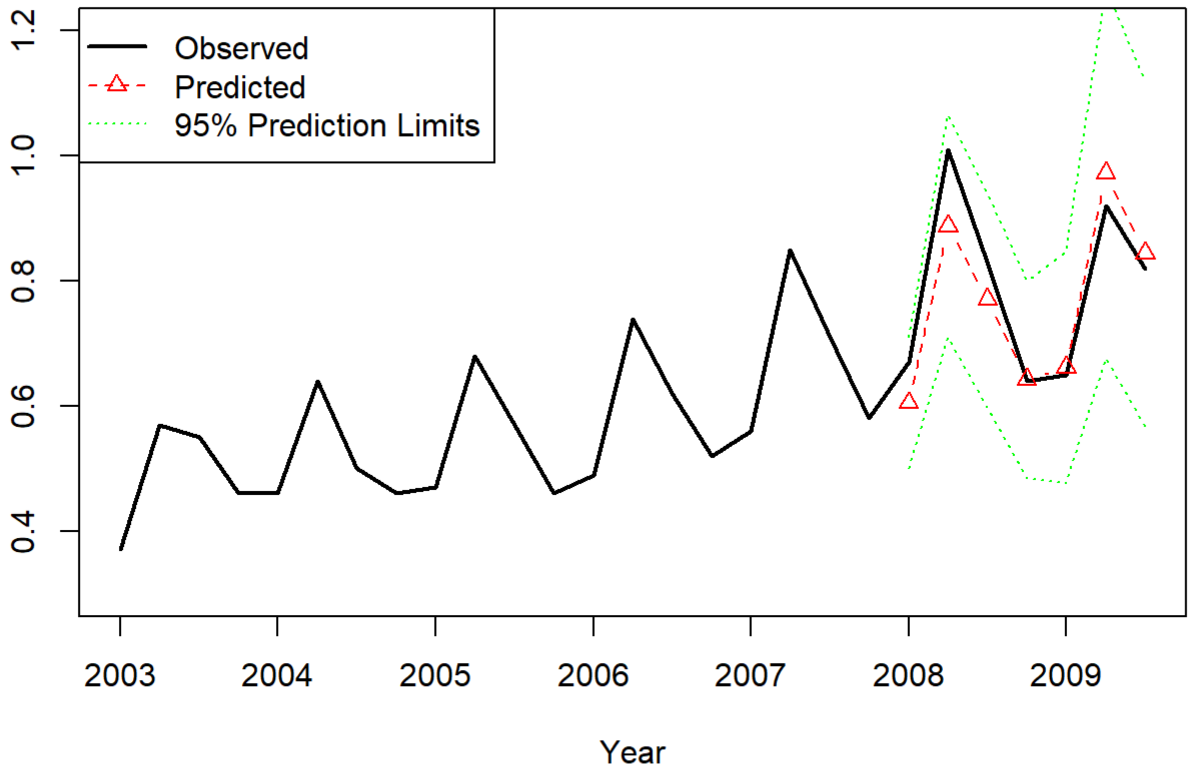
```
lognorm_adjust <- function(pred_list){
  pred <- pred_list$pred
  se <- pred_list$se
  xnew <- exp(pred + 0.5*se^2)
  senew <- sqrt(exp(2*pred + se^2)*(exp(se^2)-1))
  list(pred = xnew, se=senew)
}

pred1adj <- lognorm_adjust(pred1)
pred2 <- pred1adj$pred
se2 <- pred1adj$se
cbind(Observed=c(coredata(xts.koqtr))[101:107],
      Predicted=round(pred2, 2), SE=round(se2, 2))
```

```
##         Observed Predicted   SE
## 2008 Q1     0.67      0.61 0.05
## 2008 Q2     1.01      0.89 0.09
## 2008 Q3     0.83      0.77 0.09
## 2008 Q4     0.64      0.64 0.08
## 2009 Q1     0.65      0.66 0.09
## 2009 Q2     0.92      0.97 0.15
## 2009 Q3     0.82      0.84 0.14
```

從2003年開始，對預報效果作圖：

```
tmp.x <- ts(c(coredata(xts.koqtr["2003/"])), start=c(2003,1), frequency = 4)
tmp.p <- ts(c(pred2), start=c(2008,1), frequency = 4)
tmp.lb <- ts(c(pred2) - 2*c(se2), start=c(2008,1), frequency = 4)
tmp.ub <- ts(c(pred2) + 2*c(se2), start=c(2008,1), frequency = 4)
plot(tmp.x, ylim=c(0.3, 1.2), type="l", lwd=2,
     xlab="Year", ylab="")
lines(tmp.p, type="b", pch=2, lty=2, col="red")
lines(tmp.lb, lty=3, col="green")
lines(tmp.ub, lty=3, col="green")
legend("topleft", lty=c(1,2,3), lwd=c(2,1,1), pch=c(NA,2,NA),
       col=c("black", "red", "green"),
       legend=c("Observed", "Predicted", "95% Prediction Limits"))
```



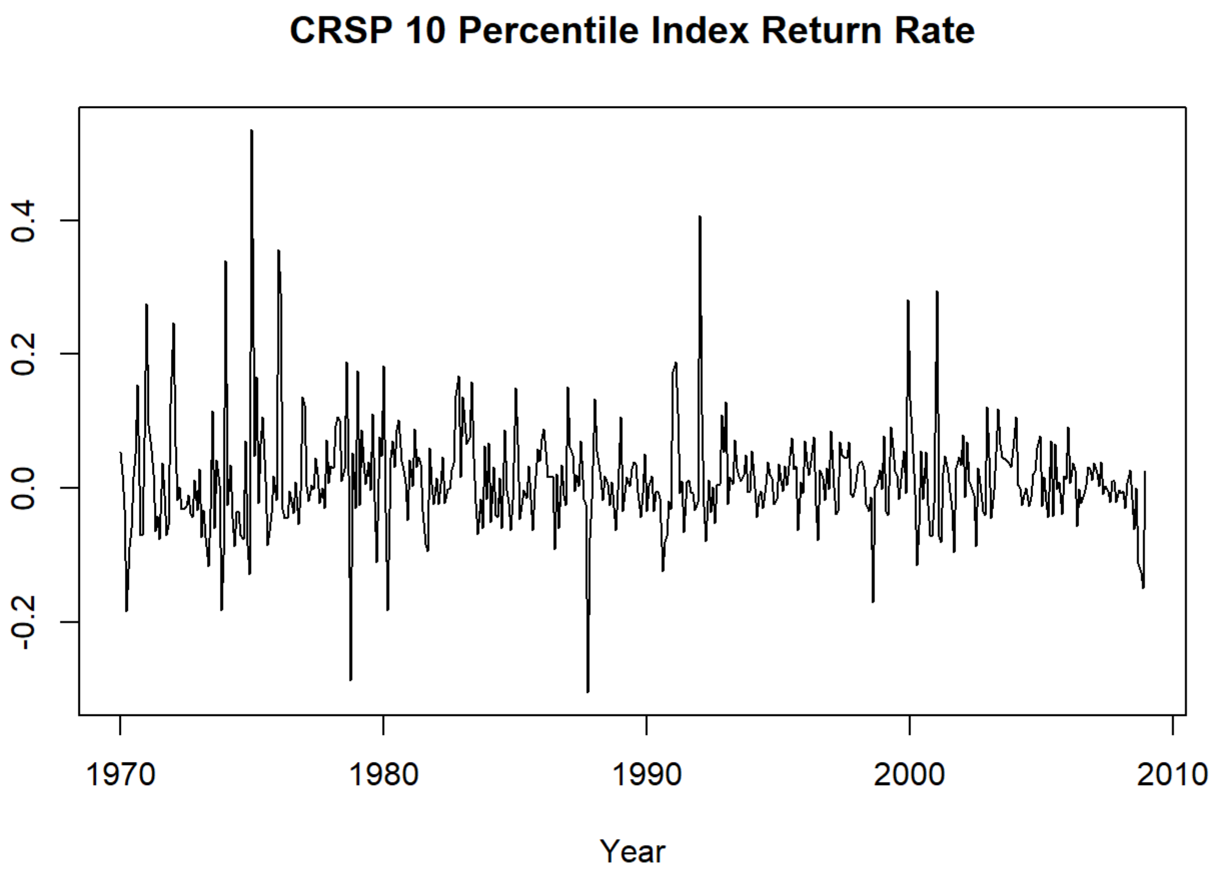
季節虛擬變數

另一種表示季節性的方法是用非隨機的回歸項表示固定的季節模式。 這樣的模式雖然也可以通過季節差分消除， 但固定的季節模型不應該用季節差分處理。

非隨機的季節因素用回歸虛擬變數表示。 $s = 4$時， 用3個虛擬變數就可以表示4個不同季節的固定水平。 為了判斷非隨機季節模型是否使用， 可以先配適動態的季節ARIMA$(1,0,1)(1,0,1)_s$模型， 當發現其中的季節因素可以忽略時， 就可考慮採用非隨機的季節模型。

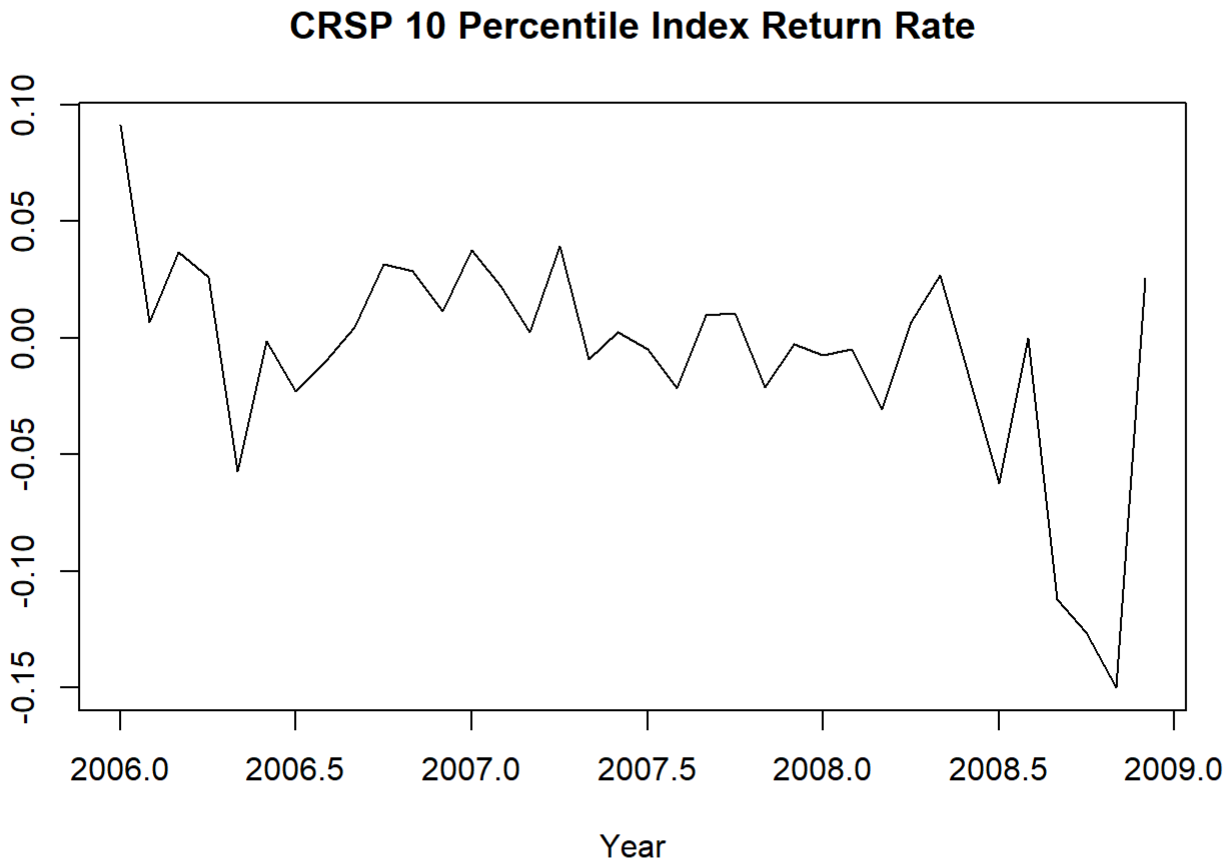
例:考慮從1970年1月到2008年12月共39年，共468個觀測值，CRSP最高10分位資產組合的月簡單報酬率，其序列圖：

```
plot(dec10, main="CRSP 10 Percentile Index Return Rate",
     xlab="Year", ylab="")
```



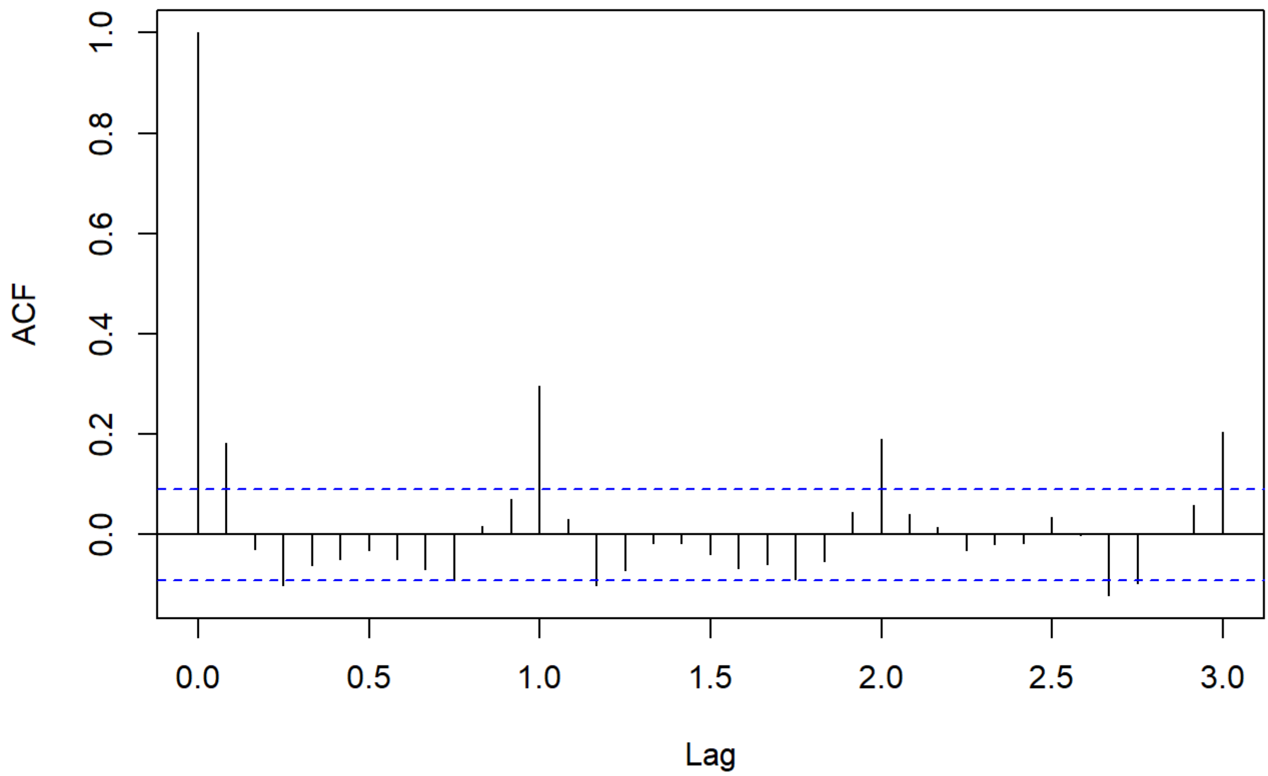
序列圖本身沒有表現出明顯的趨勢與週期性。取最近3年的數據來看，也看不出週期性：

```
plot(window(dec10, start=c(2006,1), end=c(2008,12)),
            main="CRSP 10 Percentile Index Return Rate",
     xlab="Year", ylab="")
```



作ACF圖：

```
acf(dec10, main="", lag.max=36)
```



ACF函數在12的倍數的落後上顯著不等於零， 這體現出了週期性。 來配適ARIMA$(1,0,1)(1,0,1)_{12}$:

```
resm1 <- arima(dec10, order=c(1,0,1),
               seasonal=list(order=c(1,0,1), period=12))
resm1
```

```
##
## Call:
## arima(x = dec10, order = c(1, 0, 1), seasonal = list(order = c(1, 0, 1), period = 12))
##
## Coefficients:
##           ar1     ma1    sar1     sma1  intercept
##       -0.0639  0.2508  0.9882  -0.9142     0.0117
## s.e.   0.2205  0.2130  0.0092   0.0332     0.0125
##
## sigma^2 estimated as 0.004704:  log likelihood = 584.69,  aic = -1157.39
```

得到的模型可以寫成

$$(1+0.0639L)(1-0.9882L^{12})(X_t-0.0117)=(1+0.2508L)(1-0.9142L^{12})\varepsilon_t$$

其中的$1-0.9882L^{12}$項與$1-0.9142L^{12}$近似可以消去， 所以這個結果提示可能不需要使用動態季節模型。

採用非隨機的虛擬變數模型來建模。 為簡單起見， 僅考慮一月份的效應， 因為實證發現一月份的報酬率傾向為正值。

```
jan <- as.numeric(c(cycle(dec10))==1)
lm1 <- lm(c(dec10) ~ jan); summary(lm1)
```

```
##
## Call:
## lm(formula = c(dec10) ~ jan)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.30861 -0.03475 -0.00176  0.03254  0.40671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.002864   0.003333   0.859    0.391
## jan         0.125251   0.011546  10.848   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06904 on 466 degrees of freedom
## Multiple R-squared:  0.2016, Adjusted R-squared:  0.1999
## F-statistic: 117.7 on 1 and 466 DF,  p-value: < 2.2e-16
```

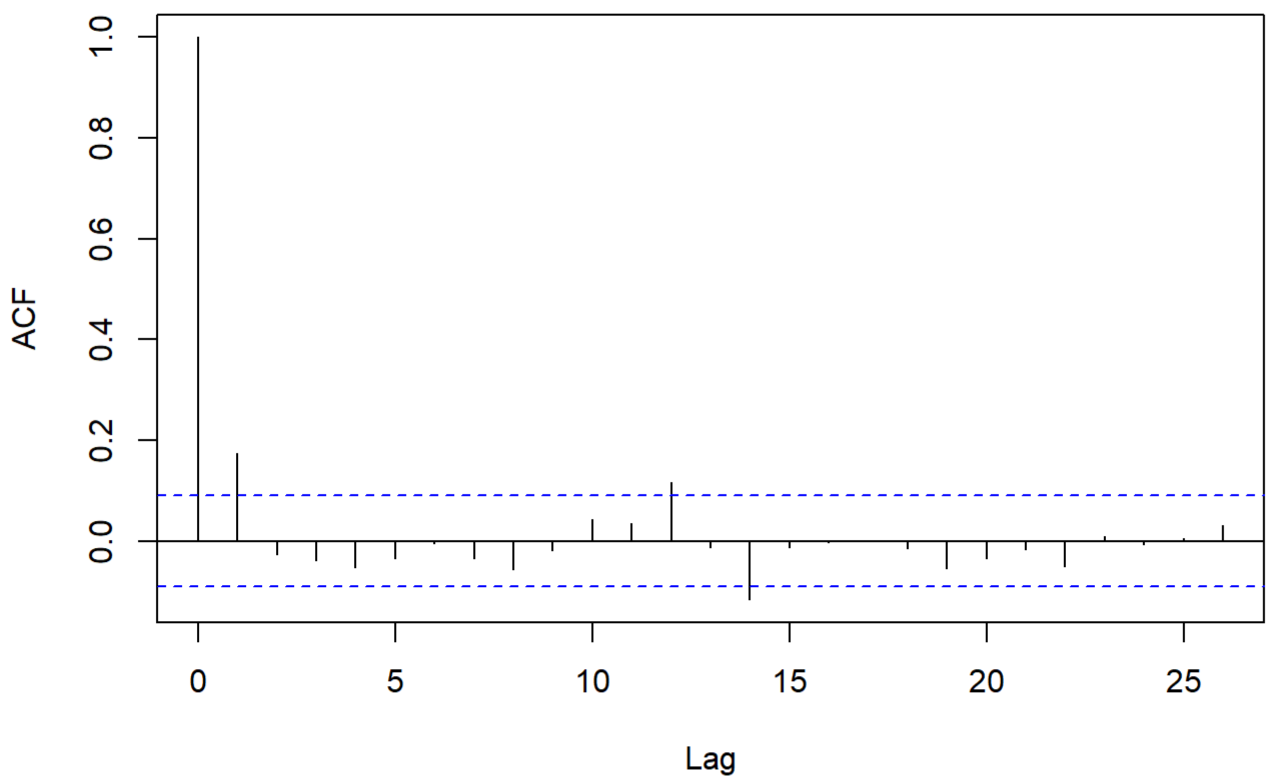
可以看出在以一月份為虛擬變數的線性回歸中， 自變量是顯著的。 配適的模型為

$$\hat{x}_t = 0.002864 + 0.1253 \text{Jan}_t$$

其中Jan_t當t的月份為一月份時等於1， 否則等於0。

但是，這個模型是有缺陷的， 因為線性回歸假定隨機誤差項獨立， 而這裡的隨機誤差項顯然是有自我相關的：

```
acf(residuals(lm1), main="")
```



其落後1還顯著不等於0。 在落後為12的倍數的位置已經不再顯著。