Statistics 435/535/711 28 March 2022
P. Shaman

ARIMA Models: Order Determination and Seasonal Models

A. Order determination. Several methods are used to judge the goodness and adequacy of an ARIMA model one has fit. In all cases one is trying to determine if the residuals are "estimating" white noise.

1. Examine the sample acf and sample pacf of the residuals. These should show statistical insignificance (but allowing for some false positives, hopefully at lags which are not of importance) if one has reduced to white noise.

2. Still better, estimate the spectral density of the residuals. If this is essentially flat, one has reduced to white noise. This is preferable to the first method cited above.

3. Employ the portmanteau test. This is the method used by Tsay in his book. Let r_k denote the kth residual autocorrelation, and m be moderately large, relative to the sample size. The Ljung-Box statistic (Q-statistic) is

$$Q_m = n^*(n^*+2)\sum_{k=1}^{m}\frac{1}{n^*-k}r_k^2,$$

where $n^* = n - d$ is the number of observations after differencing. One rejects the residual white noise hypothesis if Q_m is greater than $\chi^2(m-p-q-1;\alpha)$, the upper α percentile of the chi-square distribution with $m - p - q - 1$ degrees of freedom ($p+q+1$ is the number of parameters estimated in the fitted ARIMA model).

4. Use an order determination criterion, such as AIC, formulated by Akaike, or SBC, formulated by Schwarz. These criteria are obtained as follows:

$$AIC = -2\log LH + 2(p+q+1), \qquad SBC = -2\log LH + \log n^*(p+q+1),$$

where LH is the value of the likelihood function for the fitted model. One chooses the model for which AIC (or SBC) is smallest. The two criteria are based upon different theoretical bases and customarily do not lead to the same choice of model. There are numerous model selection criteria which have been developed and discussed in the literature, and AIC and SBC are among the most widely used.

From the definitions we see that AIC imposes a penalty of 2 units per parameter employed, and that SBC imposes a penalty which can be much larger, $\ln n^*$ units per parameter. Thus SBC tends to select models with fewer parameters than AIC. My strong preference is to use AIC for ARIMA model fitting.

5. Use overfitting. That is, once you think you have a good model, refit with more AR parameters or more MA parameters. *Avoid adding more AR and MA parameters at the same time.*

B. Seasonal models. Let s denote the seasonal period. One differences seasonally via

$$(1-B^s)y_t = y_t - y_{t-s}.$$

Box-Jenkins seasonal models employ an AR seasonal operator of order P,

$$1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps},$$

an MA seasonal operator of order Q,

$$1+\Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs},$$

and a differencing operator of order D,

$$(1-B^s)^D.$$

The ordinary and seasonal operators are applied multiplicatively. This leads to parsimony. A multiplicative seasonal

$$\text{ARIMA}(p,d,q)\ (P,D,Q)_s$$

model is specified by

$$(1-\phi_1 B - \cdots - \phi_p B^p)(1-\Phi_1 B^s - \cdots - \Phi_P B^{Ps})(1-B)^d (1-B^s)^D y_t$$

$$= (1+\theta_1 B + \cdots + \theta_q B^q)(1+\Theta_1 B^s + \cdots + \Theta_Q B^{Qs})\varepsilon_t.$$

In this formulation, p must be less than s.

There can be more than one seasonal component, for example,

$$\text{ARIMA}(p,d,q)(P_1,D_1,Q_1)_{s_1}(P_2,D_2,Q_2)_{s_2}.$$

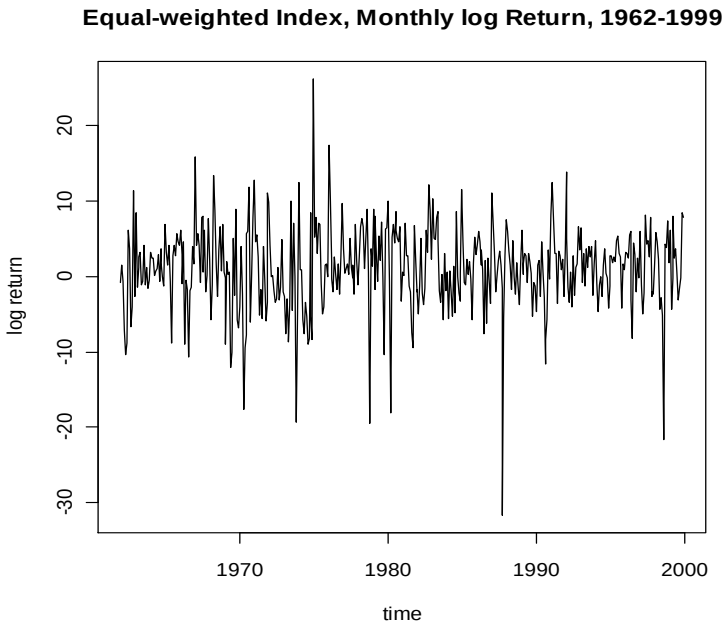
ARMA Model Fitting—CRSP Equal-Weighted Index

We examine the monthly log returns of the CRSP (Center for Research in Security Prices, at the Booth School of Business, University of Chicago) equal-weighted index from January 1962 to December 1999, 456 observations altogether. The data are in the file m-ew6299.txt. This data set is used in Problem 13 of Chapter 2 of Tsay's book.

```
> ew<-read.csv("F:Stat71122Spring/m-ew6299.txt")
> attach(ew)
> head(ew)
  crspewreturn year month
1       -0.792 1962     1
2        1.532 1962     2
3       -0.596 1962     3
4       -7.049 1962     4
5      -10.319 1962     5
6       -8.880 1962     6
```

Here is a plot of the data.

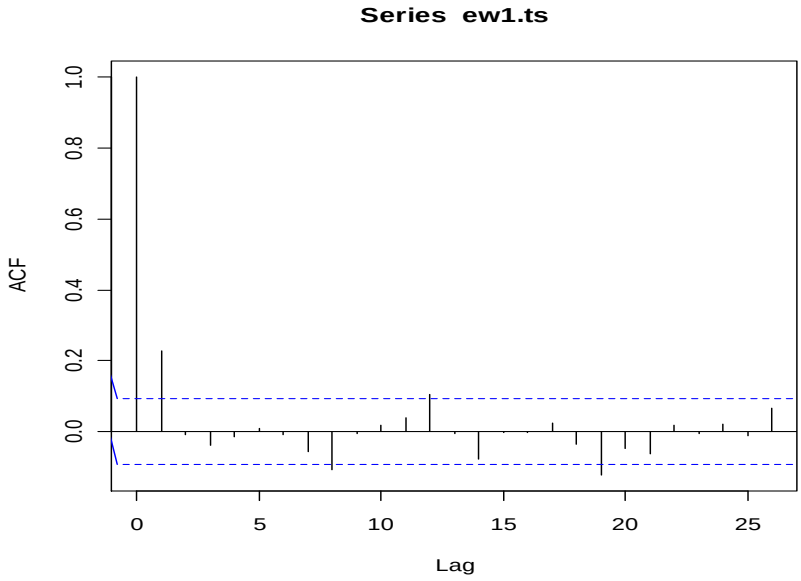
```
> ew.ts<-ts(crspewreturn,start=c(1962,1),freq=12)
> plot(ew.ts,xlab="time",ylab="log return",main="Equal-weighted Index,
Monthly log Return, 1962-1999")
```

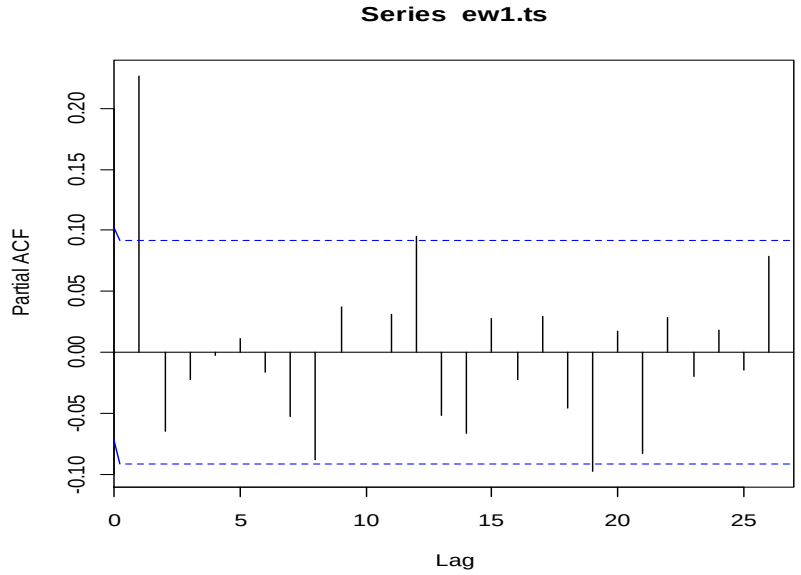


There are some outlier values, and some indication of changing volatility. For now we will postpone addressing the outliers and focus on the log returns.

Let's fit an ARMA model. As noted, to begin we won't attempt to modify any aberrant values.

Here are the autocorrelation function and partial autocorrelation function estimates:



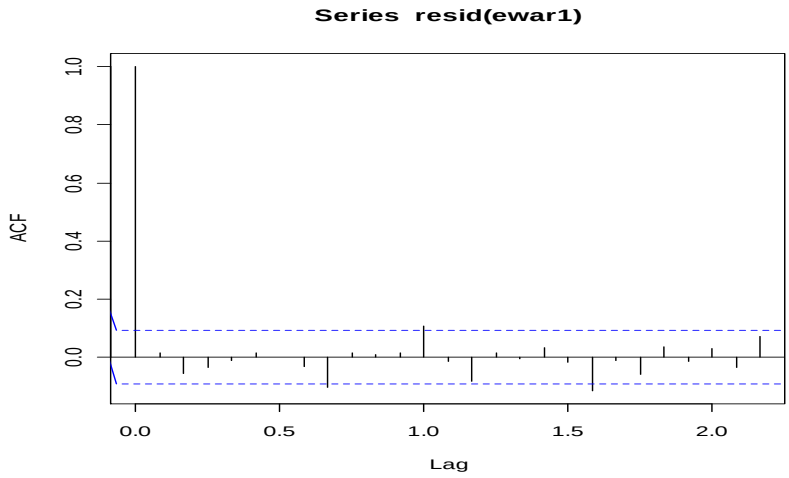


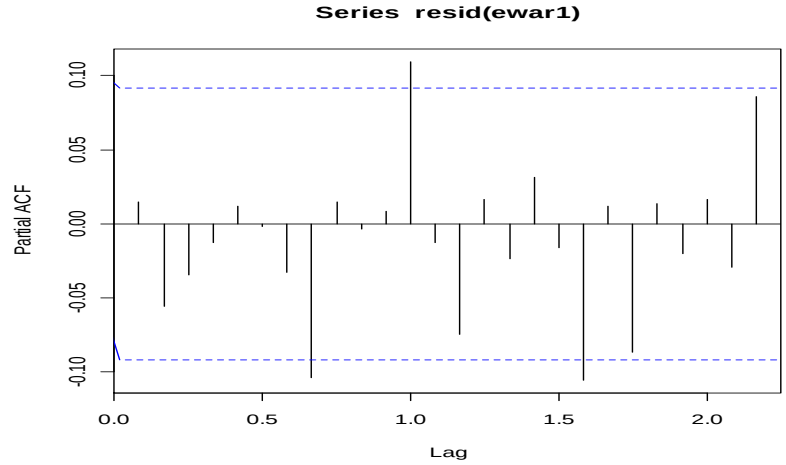
The acf and pacf estimates suggest fitting either an AR(1) or an MA(1) model. The barely significant lag 8 correlation is perhaps a false positive. The significant lag 12 acf and pacf values may warrant some attention, because lag 12 is related to seasonal structure. In addition, there is significance in both plots at lag 19, but barely so.

Here are the AR(1) and MA(1) fits.

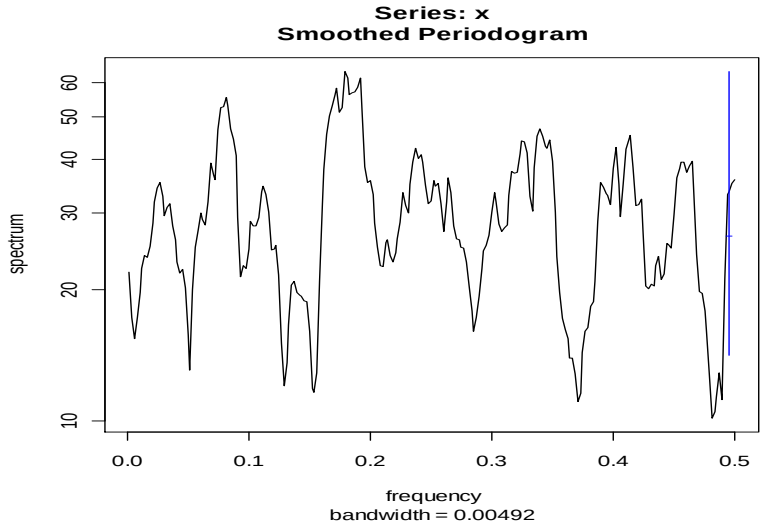
```
> ewar1<-arima(ew.ts,order=c(1,0,0))
> ewar1

Call:
arima(x = ew.ts, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.2267     1.0626
s.e.  0.0456     0.3297

sigma^2 estimated as 29.68:  log likelihood = -1420.11,  aic = 2846.22

> library("lmtest")
> coeftest(ewar1)

z test of coefficients:

          Estimate Std. Error z value  Pr(>|z|)
ar1       0.226653   0.045623  4.9679 6.767e-07 ***
intercept 1.062620   0.329698  3.2230  0.001269 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ewma1<-arima(ew.ts,order=c(0,0,1))
> ewma1

Call:
arima(x = ew.ts, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      0.2385     1.0605
s.e.  0.0449     0.3153

sigma^2 estimated as 29.59:  log likelihood = -1419.37,  aic = 2844.73

> coeftest(ewma1)

z test of coefficients:

          Estimate Std. Error z value  Pr(>|z|)
ma1       0.238495   0.044925  5.3088 1.104e-07 ***
intercept 1.060512   0.315331  3.3632 0.0007705 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual diagnostics for the fitted models follow. The first plots are for the AR(1) fit.

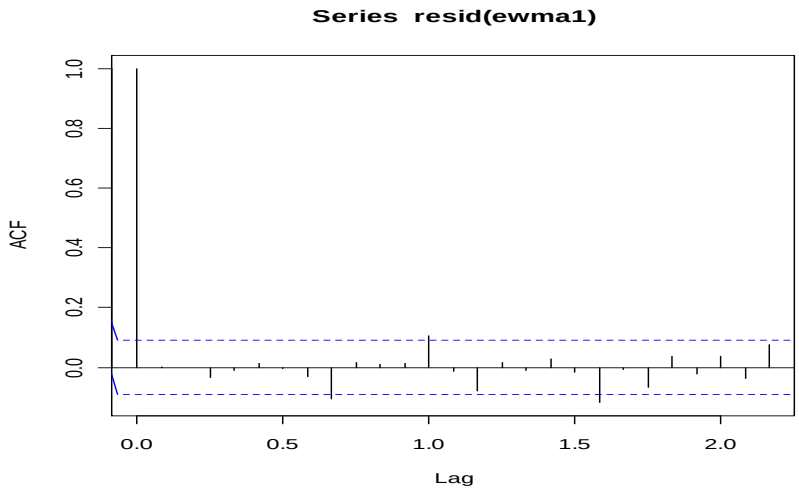


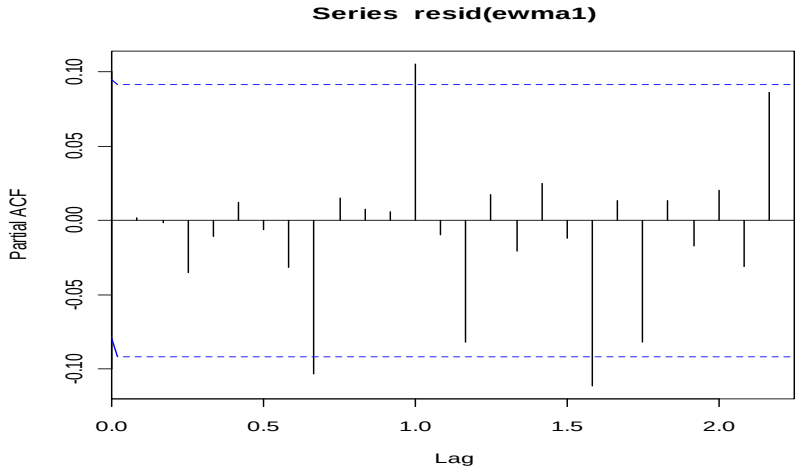


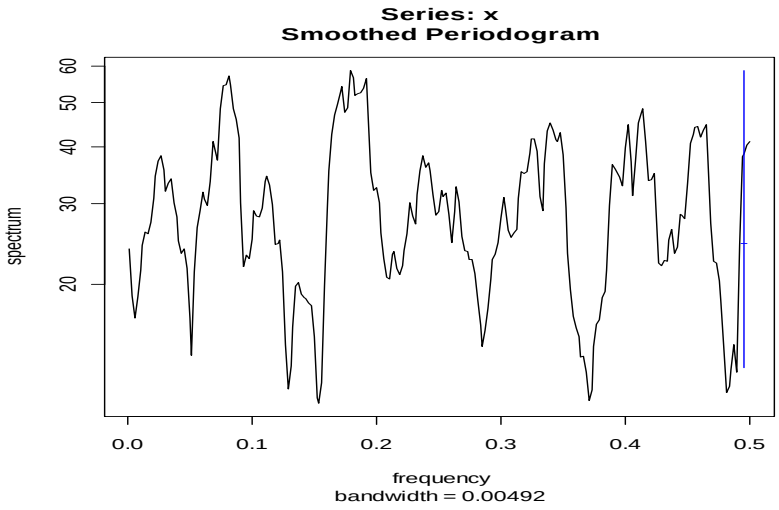

```
> spectrum(ts(resid(ewar1)),span=8)
```



Residual plots for the MA(1) fit follow.







For both fits the residual acf and pacf calculations show significance at level 0.05 at lags 8, 12, and 19, but barely so. The AIC diagnostic gives preference to the MA(1) fit. The residual spectral plots for the two models look the same. In fact, the two are virtually the same. The AR(1) model fit is

$$(1 - 0.2266B)(y_t - 1.0626) = \varepsilon_t,$$

and the MA(1) model fit is

$$y_t = 1.0605 + (1 + 0.2385B)\varepsilon_t.$$

The latter may be written in autoregressive form as

$$(1 - 0.2385B + 0.0569B^2 - 0.0136B^3 + \cdots)(y_t - 1.0605) = \varepsilon_t,$$

approximately an AR(2) model. Let's consider overfitting with AR(2), MA(2), and ARMA(1,1) models.

```
Call:
arima(x = ew.ts, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      0.2410  -0.0640     1.0615
s.e.  0.0467   0.0468     0.3093

sigma^2 estimated as 29.56:  log likelihood = -1419.18,  aic = 2846.35

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1        0.241033   0.046727  5.1583 2.492e-07 ***
ar2       -0.063996   0.046775 -1.3682 0.1712544
intercept  1.061536   0.309294  3.4321 0.0005989 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
arima(x = ew.ts, order = c(0, 0, 2))

Coefficients:
         ma1     ma2  intercept
      0.2403  0.0064     1.0605
s.e.  0.0473  0.0484     0.3174

sigma^2 estimated as 29.58:  log likelihood = -1419.36,  aic = 2846.71

z test of coefficients:

          Estimate Std. Error z value  Pr(>|z|)
ma1       0.240294   0.047266  5.0839 3.697e-07 ***
ma2       0.006358   0.048407  0.1313  0.895503
intercept 1.060500   0.317397  3.3412  0.000834 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
arima(x = ew.ts, order = c(1, 0, 1))

Coefficients:
         ar1     ma1  intercept
      0.0203  0.2195     1.0612
s.e.  0.1784  0.1729     0.3169

sigma^2 estimated as 29.58:  log likelihood = -1419.36,  aic = 2846.72

z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1       0.020348   0.178440  0.1140 0.9092104
ma1       0.219528   0.172913  1.2696 0.2042324
intercept 1.061169   0.316943  3.3481 0.0008136 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The three overfits are not needed, as one can see from the *t*-tests for the coefficients. The AR(2) model is interesting, because it mimics the MA(1). The AR(2) fit is

$$(1 - 0.2410B + 0.0640B^2)(y_t - 1.0615) = \varepsilon_t.$$

See the AR representation of the MA(1) model, above.

AIC chooses the MA(1) model from among all the fits.

The MA(2) model fit is

$$y_t = 1.0605 + (1 + 0.2403B + 0.0064B^2)\varepsilon_t.$$

In AR representation this is

$$(1 - 0.2403B + 0.0514B^2 + \cdots)(y_t - 1.0605) = \varepsilon_t.$$

And the ARMA(1,1) model fit is

$$(1 - 0.0203B)(y_t - 1.0612) = (1 + 0.2195B)\varepsilon_t.$$

The autoregressive representation of this model is

$$(1 - 0.2399B + 0.0527B^2 + \cdots)(y_t - 1.0612) = \varepsilon_t.$$

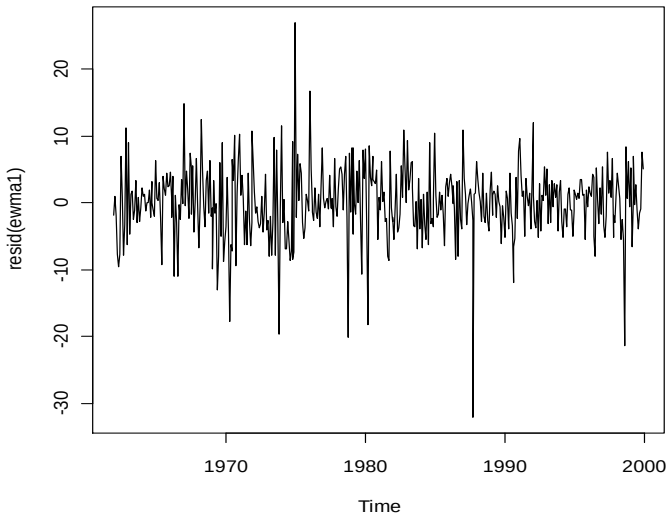
Thus, the MA(1), AR(2), MA(2), and ARMA(1,1) model fits are all essentially the same.

Let's interpret the MA(1) fit using its AR representation. First, the MA(1) model estimates annual growth of approximately 12(1.0605) = 12.7, which translates to 13.6 per cent. The deviation of the current monthly return from the estimated monthly average

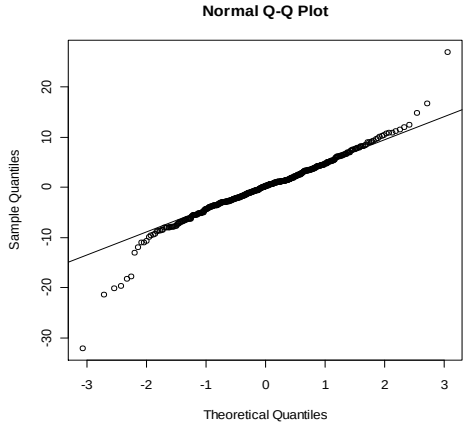
return, 1.0605 per cent, is modelled as 0.24 times the deviation in the previous month, minus 0.06 times the deviation two months ago, and so on, plus a new current random shock with standard deviation $(29.59)^{1/2} = 5.44$ per cent. The signal is weak.

Next, we address outliers and seasonality. Let's look at residual diagnostics for the MA(1) fit.

```
> plot(resid(ewma1))
```




```
> qqnorm(resid(ewma1))
> qqline(resid(ewma1))
```




```
> sort(resid(ewma1))[1:10]
 [1] -32.02327 -21.38313 -20.07529 -19.63106 -18.23474 -17.69099 -13.07781
 [8] -11.86764 -11.01002 -10.96967

> sort(resid(ewma1))[447:456]
 [1] 10.73535 10.83243 10.93867 11.13956 11.52465 12.04404 12.44446 14.86299
 [9] 16.61820 26.88965
```

The large positive residual, 26.89, occurred in January 1975. This residual value is 4.9 times the standard deviation. During April–December 1974 there were decreases in the

index in all months except October, which saw an increase. The changes were, beginning with April, −5.711, −7.533, −3.406, −5.332, −8.973, −8.230, 8.521 (October), −5.275, and −8.416. Then January 1975 experienced a substantial upward movement, a return of 26.175, or 29.92 per cent.

The large negative residual, −32.02, is October 1987. This residual value is in magnitude 5.9 times the standard deviation. There was massive selling on 19 October 1987, so-called Black Monday—the Dow-Jones Industrial Average lost 20 per cent of its value. For the month the S&P500 return was down 21.76. The return for the equal-weighted index for the month was –31.588.

In addition, there is a clearly isolated large negative residual, −21.38, for August 1998. On 7 August 1998, the U. S. embassies in Nairobi, Kenya and Dar es Salaam, Tanzania were bombed. Altogether about 225 people were killed. There are four other large negative residuals close in value to this data point.

To deal with seasonality and some of the large outliers, let's start with a regression model. We'll take the regression residuals, hence deseasonalizing the time series and neutralizing several large outliers, and then fit ARMA models to the residuals. In order to conveniently view the significance of the seasonal structure month-by-month, we define and use the set of monthly dummies on page 28 of the 12 January notes, rather than allow R to define monthly dummies.

Here are the construction of the monthly dummies and the outlier dummies:

```
> SJ1<-c(rep(c(1,rep(0,10),-1),38))
> SJ2<-c(rep(c(0,1,rep(0,9),-1),38))
> SJ3<-c(rep(c(rep(0,2),1,rep(0,8),-1),38))
> SJ4<-c(rep(c(rep(0,3),1,rep(0,7),-1),38))
> SJ5<-c(rep(c(rep(0,4),1,rep(0,6),-1),38))
> SJ6<-c(rep(c(rep(0,5),1,rep(0,5),-1),38))
> SJ7<-c(rep(c(rep(0,6),1,rep(0,4),-1),38))
> SJ8<-c(rep(c(rep(0,7),1,rep(0,3),-1),38))
> SJ9<-c(rep(c(rep(0,8),1,rep(0,2),-1),38))
> SJ10<-c(rep(c(rep(0,9),1,0,-1),38))
> SJ11<-c(rep(c(rep(0,10),1,-1),38))

> obs157<-c(rep(0,156),1,rep(0,299))
> obs310<-c(rep(0,309),1,rep(0,146))
> obs440<-c(rep(0,439),1,rep(0,16))

> seasmodel<-
lm(crspewreturn~SJ1+SJ2+SJ3+SJ4+SJ5+SJ6+SJ7+SJ8+SJ9+SJ10+SJ11+obs157+ob
s310+obs440);summary(seasmodel)

Call:
lm(formula = crspewreturn ~ SJ1 + SJ2 + SJ3 + SJ4 + SJ5 + SJ6 +
    SJ7 + SJ8 + SJ9 + SJ10 + SJ11 + obs157 + obs310 + obs440)
```

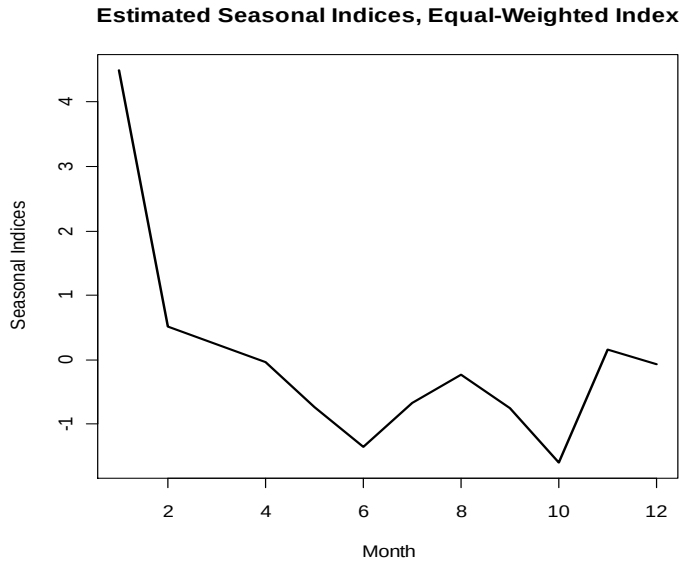
```
Residuals:
     Min       1Q   Median       3Q      Max
-20.5579  -2.7274   0.4102   3.0088  12.5559

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.13216    0.23543   4.809 2.09e-06 ***
SJ1           4.48881    0.78793   5.697 2.23e-08 ***
SJ2           0.52279    0.77843   0.672   0.5022
SJ3           0.24674    0.77843   0.317   0.7514
SJ4          -0.02848    0.77843  -0.037   0.9708
SJ5          -0.74095    0.77843  -0.952   0.3417
SJ6          -1.35766    0.77843  -1.744   0.0818 .
SJ7          -0.66508    0.77843  -0.854   0.3934
SJ8          -0.23243    0.78793  -0.295   0.7681
SJ9          -0.74616    0.77843  -0.959   0.3383
SJ10         -1.58805    0.78793  -2.015   0.0445 *
SJ11          0.16179    0.77843   0.208   0.8355
obs157       20.55403    5.07767   4.048 6.10e-05 ***
obs310      -31.13211    5.07767  -6.131 1.94e-09 ***
obs440      -22.54973    5.07767  -4.441 1.13e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.01 on 441 degrees of freedom
Multiple R-squared: 0.2241,     Adjusted R-squared: 0.1995
F-statistic: 9.099 on 14 and 441 DF,  p-value: < 2.2e-16
```

And the estimate for the December seasonal index:

```
> Dec<--sum(coef(seasmodel)[2:12])
> Dec
[1] -0.06131757
```

The January index is highly statistically significant, indicating an estimated return 4.5 per cent greater than the overall average monthly return. The indices for June and October are marginally significant, with estimated returns 1.4 per cent and 1.6 per cent, respectively, less than the overall average monthly return.

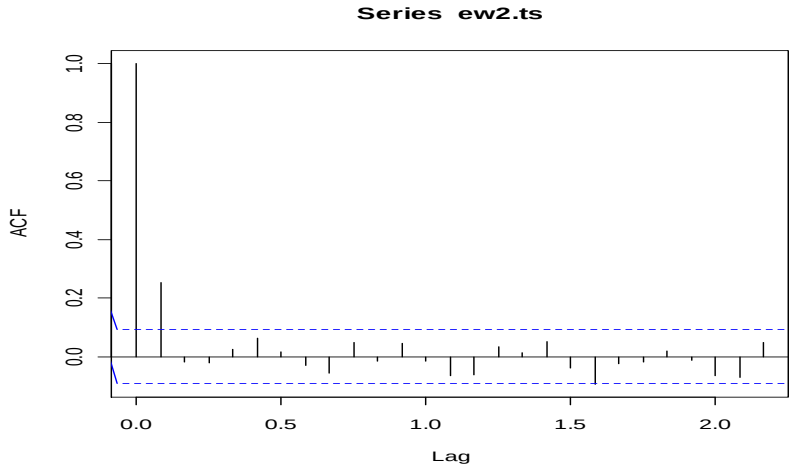
```
> plot(ts(c(coef(seasmodel)[2:12],Dec)),xlab="Month",ylab="Seasonal
Indices",main="Estimated Seasonal Indices, Equal-Weighted
Index",lty=1,lwd=2)
```



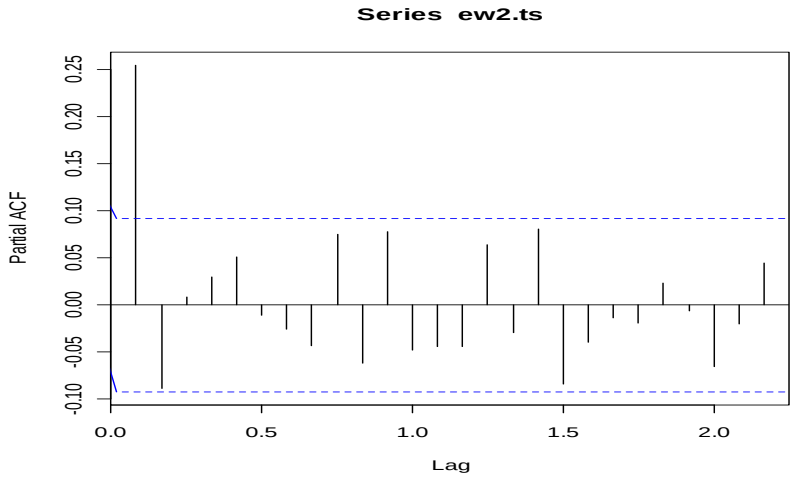
As the plot shows, for most of the months the return is estimated to be at the level of the overall mean monthly return.

Let's fit an ARMA model to the residual series from the regression.

```
> ew2.ts<-ts(resid(seasmodel),start=c(1962,1),freq=12)
> acf(ew2.ts)
```



```
> pacf(ew2.ts)
```



The acf and pacf plots suggest AR(1) and MA(1) fits. Note that there is no significance in the plots at lags 8, 12, and 19. And, based on the previous analyses, an AR(2) fit should also be tried.

```
> ew2ma1<-arima(ew2.ts,order=c(0,0,1))
> ew2ma1

Call:
arima(x = ew2.ts, order = c(0, 0, 1))

Coefficients:
         ma1  intercept
      0.2800    -0.0011
s.e.  0.0451     0.2844

sigma^2 estimated as 22.53:  log likelihood = -1357.3,  aic = 2720.6
> coeftest(ew2ma1)

z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ma1        0.2800488  0.0450514  6.2162 5.093e-10 ***
intercept -0.0010953  0.2844145 -0.0039    0.9969
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ew2ar1<-arima(ew2.ts,order=c(1,0,0))
> ew2ar1

Call:
arima(x = ew2.ts, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.2556     0.0002
s.e.  0.0454     0.2995
```

```
sigma^2 estimated as 22.7:  log likelihood = -1358.95,  aic = 2723.89
> coeftest(ew2ar1)

z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1       0.25560494 0.04538980  5.6313 1.788e-08 ***
intercept 0.00021762 0.29948789  0.0007    0.9994
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ew2ar2<-arima(ew2.ts,order=c(2,0,0))
> ew2ar2

Call:
arima(x = ew2.ts, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      0.2779  -0.0884    -0.0015
s.e.  0.0467   0.0468     0.2741

sigma^2 estimated as 22.52:  log likelihood = -1357.17,  aic = 2722.34
> coeftest(ew2ar2)

z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
ar1        0.2778978  0.0467364  5.9461 2.747e-09 ***
ar2       -0.0884149  0.0468021 -1.8891   0.05888 .
intercept -0.0015475  0.2741134 -0.0056   0.99550
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The AIC values are 2720.60, 2722.34, and 2723.89 for the MA(1), AR(2), and AR(1) models, respectively. The MA(1) fit, applied to the deseasonalized series, is

$$y_t = -0.0011 + (1 + 0.2800B)\varepsilon_t.$$

Rewriting this in autoregressive form, we have

$$(1-0.2800B+0.0784B^2-0.0220B^3+\cdots)(y_t+0.0011)=\varepsilon_t.$$

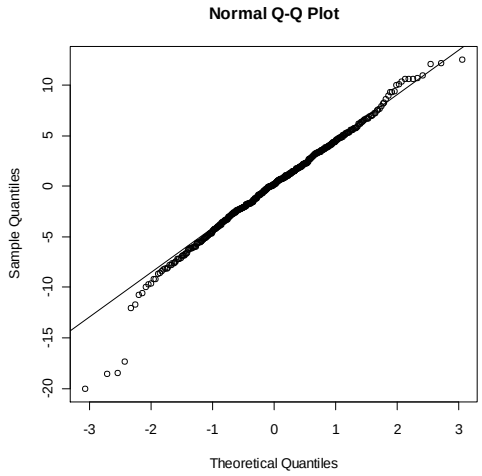
The AR(2) fit, applied to the deseasonalized series, is

$$(1-0.2779B+0.0884B^2)(y_t+0.0015)=\varepsilon_t.$$

The two models are essentially the same. AIC gives preference to the MA(1) model, which is very similar to the previous MA(1) fit obtained without deseasonalizing and adjusting for the January 1975, October 1987, and August 1998 outliers. The residual acf and pacf plots show no significant values for this MA(1) model.

Here is the normal quantile plot of the residuals from the MA(1) fit:

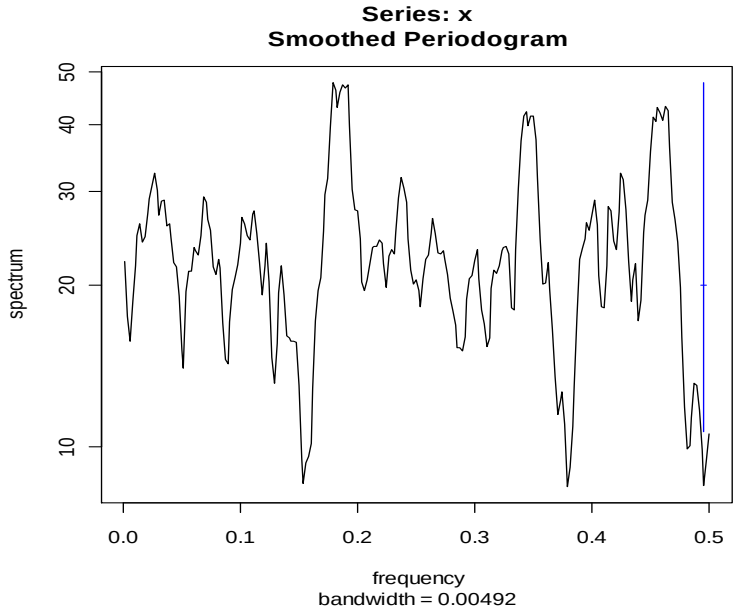
```
> qqnorm(resid(ew2ma1))
> qqline(resid(ew2ma1))
```



There are four somewhat large negative residuals.

And here is the spectral density of the residuals from the MA(1) fit.

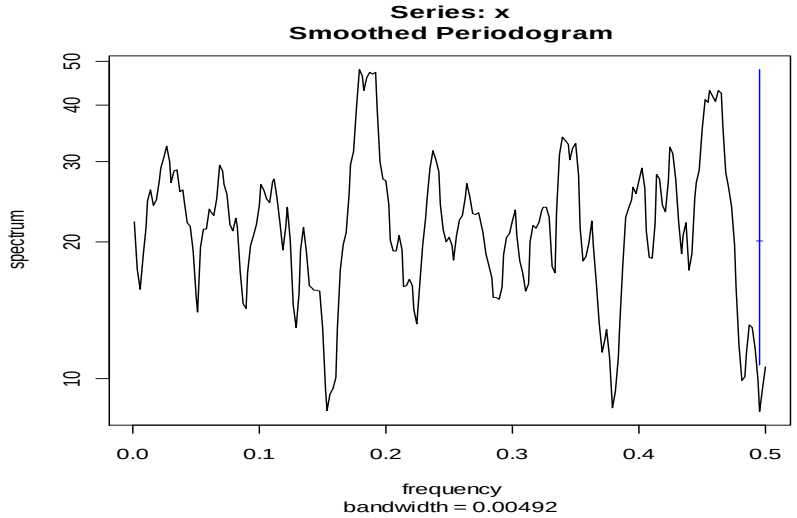
```
> spectrum(ts(resid(ew2ma1)),span=8)
```



There are three prominent spectral peaks. One is at the calendar frequency 0.348. In addition, there is a small peak at the calendar frequency 0.220. Let's add trigonometric calendar terms to the initial regression and then fit the MA(1) model to the new regression residuals.

```
> time<-1:456
> c220<-cos(0.44*pi*time);s220<-sin(0.44*pi*time)
> c348<-cos(0.696*pi*time);s348<-sin(0.696*pi*time)

> seasmodel2<-
lm(crspewreturn~SJ1+SJ2+SJ3+SJ4+SJ5+SJ6+SJ7+SJ8+SJ9+SJ10+SJ11+obs157+ob
s310+obs440+c220+s220+c348+s348);summary(seasmodel2)

Call:
lm(formula = crspewreturn ~ SJ1 + SJ2 + SJ3 + SJ4 + SJ5 + SJ6 +
    SJ7 + SJ8 + SJ9 + SJ10 + SJ11 + obs157 + obs310 + obs440 +
    c220 + s220 + c348 + s348)

Residuals:
     Min       1Q   Median       3Q      Max
-21.4642  -2.7364   0.1313   3.0434  12.6982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.13698    0.23336   4.872 1.55e-06 ***
SJ1           4.51727    0.78120   5.782 1.40e-08 ***
SJ2           0.51546    0.77177   0.668  0.50456
SJ3           0.26182    0.77175   0.339  0.73458
SJ4          -0.06899    0.77177  -0.089  0.92881
SJ5          -0.75588    0.77179  -0.979  0.32793
SJ6          -1.31170    0.77173  -1.700  0.08990 .
SJ7          -0.68575    0.77175  -0.889  0.37472
SJ8          -0.25175    0.78112  -0.322  0.74739
SJ9          -0.73723    0.77175  -0.955  0.33997
SJ10         -1.61082    0.78114  -2.062  0.03978 *
SJ11          0.17915    0.77177   0.232  0.81655
obs157       19.62035    5.05559   3.881  0.00012 ***
obs310      -30.91070    5.05518  -6.115 2.15e-09 ***
obs440      -22.89695    5.05489  -4.530 7.63e-06 ***
c220         -0.40429    0.33022  -1.224  0.22149
s220         -0.77432    0.33006  -2.346  0.01942 *
c348          0.27992    0.33068   0.847  0.39773
s348         -0.66297    0.32982  -2.010  0.04503 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.966 on 437 degrees of freedom
Multiple R-squared: 0.2446,     Adjusted R-squared: 0.2135
F-statistic: 7.861 on 18 and 437 DF,  p-value: < 2.2e-16

> ew3.ts<-ts(resid(seasmodel2,start=c(1962,1),freq=12))
> ew3ma1<-arima(ew3.ts,order=c(0,0,1))
> ew3ma1

Call:
arima(x = ew3.ts, order = c(0, 0, 1))
```

```
Coefficients:
         ma1  intercept
      0.2802    -0.0006
s.e.  0.0441     0.2802

sigma^2 estimated as 21.87:  log likelihood = -1350.47,  aic = 2706.93

> coeftest(ew3ma1)

z test of coefficients:

             Estimate  Std. Error z value  Pr(>|z|)
ma1        0.28020982  0.04408885  6.3556 2.077e-10 ***
intercept -0.00058308  0.28021908 -0.0021    0.9983
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> spectrum(ts(resid(ew3ma1)),span=8)
```



This plot is slightly better than the residual spectral plot on page 16. The difference is that here we added adjustments for two calendar frequencies.

Instead of the analysis presented, which is a two-step procedure, with regression followed by ARMA fitting, an ARMAX model can be employed. Still another approach to account for the seasonality is to fit a seasonal ARMA model, rather than an original regression model. The regression model provides estimation of only a static seasonal structure, while a seasonal ARMA model permits estimation of both static and dynamic seasonal structure. Seasonal ARMA models will be discussed in detail in future notes.

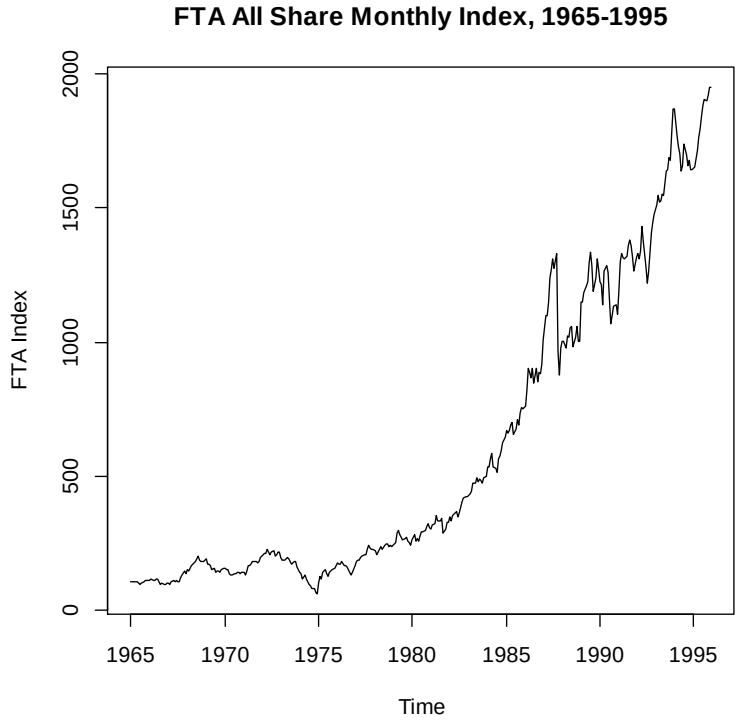
To add to this discussion we look at another broadly-based stock index for a similar stretch of time. It is the *Financial Times*-Actuaries (FTA) All Share monthly index, from

January 1965 to December 1995, 372 observations altogether. The data are in FTAPRICE.txt.

```
> fta<-read.csv("F:/Stat71122Spring/FTAPRICE.txt")
> attach(fta)
> head(fta)
  year month ftapriceindex logindex   logreturn revlogreturn
1 1965     1        109.14 4.692631          NA           NA
2 1965     2        107.80 4.680278 -0.01235380  -0.01235380
3 1965     3        106.18 4.665136 -0.01514190  -0.01514190
4 1965     4        107.12 4.673950  0.00881393   0.00881393
5 1965     5        106.21 4.665418 -0.00853140  -0.00853140
6 1965     6        100.39 4.609063 -0.05635570  -0.05635570
```

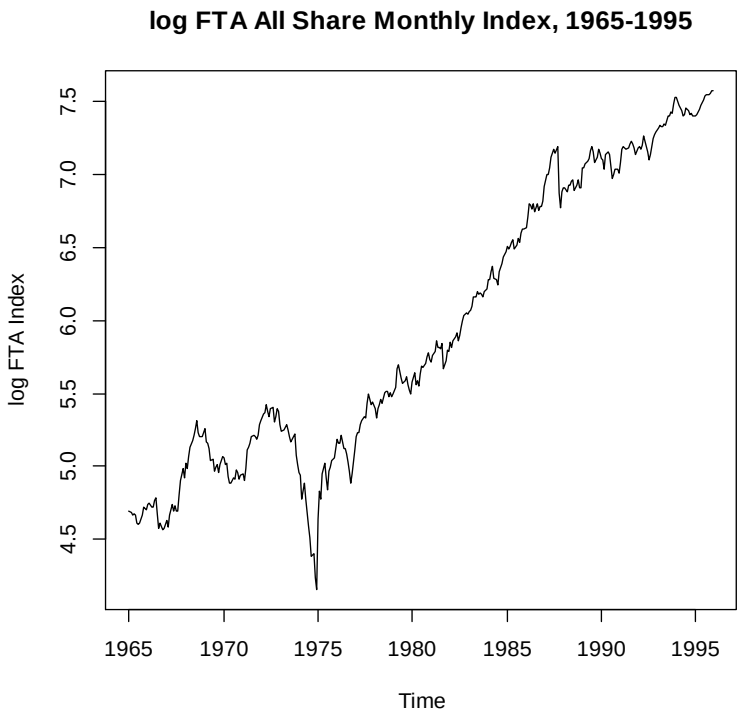
Here is a plot of the index itself. There is an upward trend which is not linear, and the fluctuations about the trend become more intense as the level rises.

```
> fta.ts<-ts(ftapriceindex,start=c(1965,1),freq=12)
> plot(fta.ts,xlab="Time",ylab="FTA Index",main="FTA All Share Monthly
Index, 1965-1995")
```

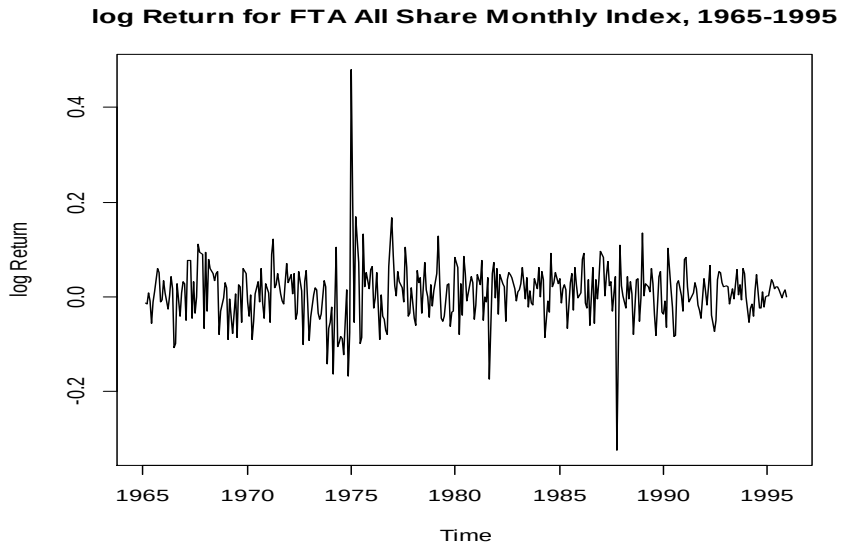


The logarithmic transformation produces, more or less, a linear trend, and it tends to stabilize the volatility.

```
> fta2.ts<-ts(logindex,start=c(1965,1),freq=12)
> plot(fta2.ts,xlab="Time",ylab="log FTA Index",main="log FTA All Share
Monthly Index, 1965-1995")
```



The difference of this log series is the log return, which is what we analyze.



There are two very prominent outliers, points 121 and 274, corresponding to January 1975 and October 1987, respectively. Another outlier is visible at point 201, for September 1981.

The January 1975 value is noticeably more extreme than the same monthly reading for the CRSP equal-weighted index. The log return for January 1975 is 0.4796, a rise of

61.54 per cent, and that for October 1987 is –0.3225, a drop of 27.57 per cent. We note that January 1975 followed a 14-month stretch during which the FTA index dropped substantially, with decreases in 12 of the 14 months. Here are the data for this 14-month period, and three previous months:

Year	Month	Price	Difflnprice
1973	8	175.29	-0.0366854
1973	9	181.3	0.03371137
1973	10	185.13	0.02090516
1973	11	160.77	-0.1410835
1973	12	150.61	-0.0652811
1974	1	142.63	-0.0544398
1974	2	139.64	-0.0211862
1974	3	118.73	-0.1622157
1974	4	131.83	0.10466121
1974	5	118.74	-0.104577
1974	6	107.79	-0.0967513
1974	7	99.19	-0.0831477
1974	8	90.91	-0.0871672
1974	9	80.39	-0.1229802
1974	10	81.63	0.01530705
1974	11	69.01	-0.1679454
1974	12	63.98	-0.0756809
1975	1	103.35	0.47955075

We begin with an initial regression to test for seasonality and address the outliers. We define the following monthly dummies.

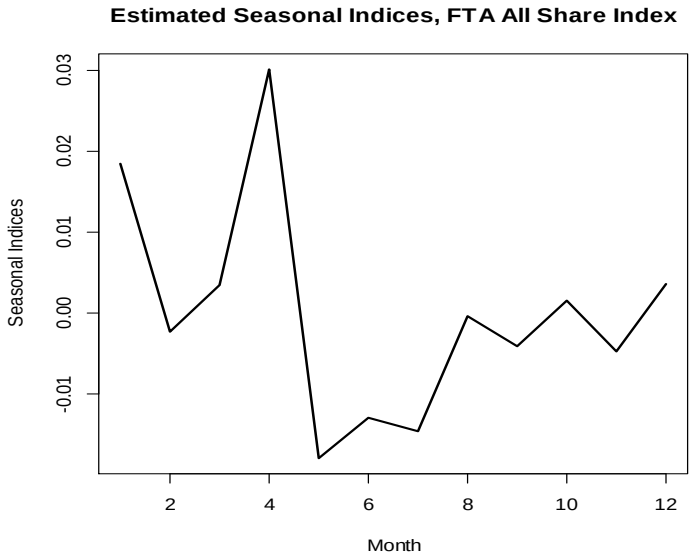
```
SJ1<-c(rep(c(1,rep(0,10),-1),31))
SJ2<-c(rep(c(0,1,rep(0,9),-1),31))
SJ3<-c(rep(c(rep(0,2),1,rep(0,8),-1),31))
SJ4<-c(rep(c(rep(0,3),1,rep(0,7),-1),31))
SJ5<-c(rep(c(rep(0,4),1,rep(0,6),-1),31))
SJ6<-c(rep(c(rep(0,5),1,rep(0,5),-1),31))
SJ7<-c(rep(c(rep(0,6),1,rep(0,4),-1),31))
SJ8<-c(rep(c(rep(0,7),1,rep(0,3),-1),31))
SJ9<-c(rep(c(rep(0,8),1,rep(0,2),-1),31))
SJ10<-c(rep(c(rep(0,9),1,0,-1),31))
SJ11<-c(rep(c(rep(0,10),1,-1),31))
obs121<-c(rep(0,120),1,rep(0,251))
obs201<-c(rep(0,200),1,rep(0,171))
obs274<-c(rep(0,273),1,rep(0,98))

> seasmodelfta<-
lm(logreturn~SJ1+SJ2+SJ3+SJ4+SJ5+SJ6+SJ7+SJ8+SJ9+SJ10+SJ11+obs121+obs20
1+obs274);summary(seasmodelfta)
```

```
Call:
lm(formula = logreturn ~ SJ1 + SJ2 + SJ3 + SJ4 + SJ5 + SJ6 +
    SJ7 + SJ8 + SJ9 + SJ10 + SJ11 + obs121 + obs201 + obs274)

Residuals:
      Min        1Q    Median        3Q       Max
-0.173655 -0.031361  0.002817  0.033671  0.189675

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0079737  0.0027030   2.950 0.003388 **
SJ1          0.0185104  0.0091941   2.013 0.044836 *
SJ2         -0.0022693  0.0089190  -0.254 0.799309
SJ3          0.0034655  0.0089190   0.389 0.697839
SJ4          0.0301018  0.0089190   3.375 0.000819 ***
SJ5         -0.0179147  0.0089190  -2.009 0.045335 *
SJ6         -0.0130188  0.0089190  -1.460 0.145265
SJ7         -0.0145998  0.0089190  -1.637 0.102529
SJ8         -0.0004285  0.0089190  -0.048 0.961705
SJ9         -0.0041712  0.0090530  -0.461 0.645254
SJ10         0.0014974  0.0090530   0.165 0.868720
SJ11        -0.0047509  0.0089190  -0.533 0.594598
obs121       0.4530666  0.0527268   8.593 2.71e-16 ***
obs201      -0.1781077  0.0526975  -3.380 0.000806 ***
obs274      -0.3319414  0.0526975  -6.299 8.84e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05184 on 356 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.3085,    Adjusted R-squared:  0.2813
F-statistic: 11.35 on 14 and 356 DF,  p-value: < 2.2e-16
```

Here is the estimate of the December seasonal index.

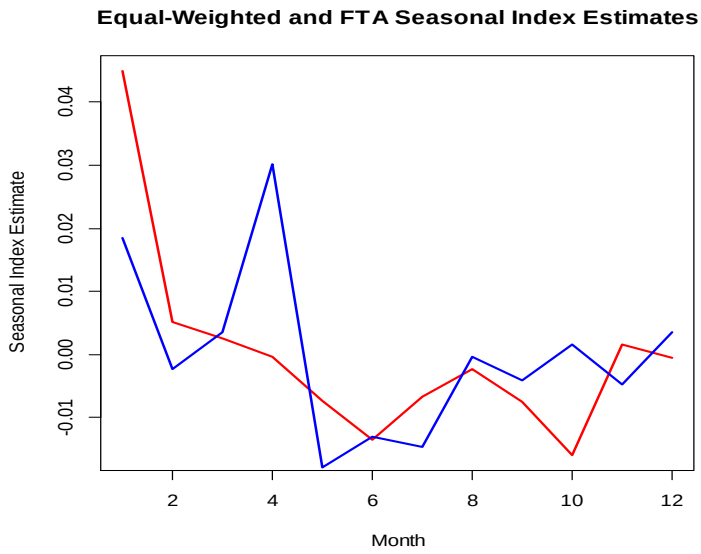
```
> Decfta<--sum(coef(seasmodelfta)[2:12])
> Decfta
[1] 0.003578039
```

The January, April, and May indices are statistically significant. For January the estimated return is 1.9 per cent greater than the overall average monthly return, and for April the estimate is 3.0 per cent greater. The May estimated return is1.8 per cent less than the overall average monthly return. A plot of the estimated seasonal indices follows.



Let's compare the sets of seasonal index estimates for the equal-weighted and FTA indices. First note that there is a scale difference between the two data sets. The equal-weighted returns are on a scale 100 times that of the FTA returns. We divide the estimated seasonal indices for the equal-weighted index by 100 for the table and plot below.

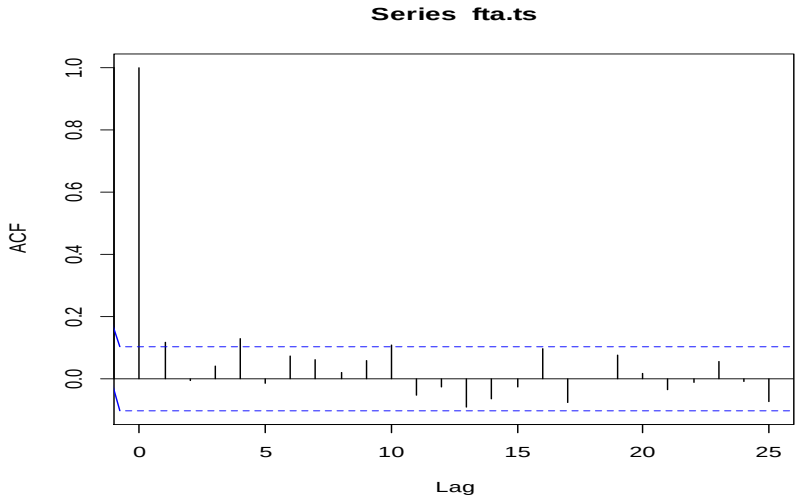
```
> ewseas<-c(coef(seasmodel)[2:12],Dec)
> ftaseas<-c(coef(seasmodelfta)[2:12],Decfta)
> cbind(1:12,ewseas,ftaseas)

         ewseas  ftaseas

SJ1   1  0.0449  0.0185
SJ2   2  0.0052 -0.0023
SJ3   3  0.0025  0.0035
SJ4   4 -0.0003  0.0301
SJ5   5 -0.0074 -0.0179
SJ6   6 -0.0136 -0.0130
SJ7   7 -0.0067 -0.0146
SJ8   8 -0.0023 -0.0004
SJ9   9 -0.0075 -0.0042
SJ10 10 -0.0159  0.0015
SJ11 11  0.0016 -0.0048
     12 -0.0006  0.0036
```



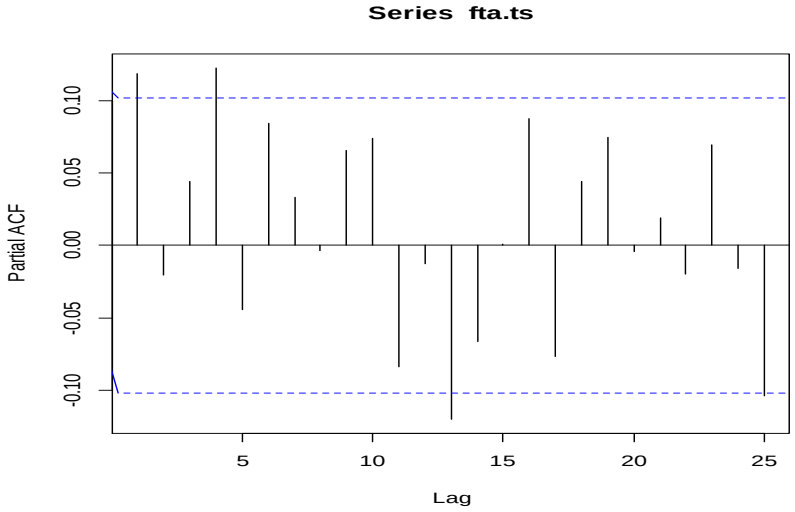
The equal-weighted estimates are in red and the FTA estimates are in blue. The January effect is less prominent for the FTA index, and the April effect present in the FTA index does not appear for the equal-weighted index.

Next, we fit an ARMA model to the FTA regression residuals. That is, we work with a time series which has been deseasonalized and adjusted for several outlier values.

```
> fta.ts<-ts(resid(seasmodelfta),start=c(1965,1))
> acf(fta.ts)
```



```
> pacf(fta.ts)
```



The acf and pacf estimates suggest an AR(1) or MA(1) fit, with AR(4) and MA(4) also possibilities. Let's try all of these models. The signal is very weak.

```
> ftaar1<-arima(fta.ts,order=c(1,0,0))
> ftaar1

Call:
arima(x = fta.ts, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
       0.118      0.000
s.e.   0.051      0.003

sigma^2 estimated as 0.00254:  log likelihood = 581.85,  aic = -1157.7

> coeftest(ftaar1)

z test of coefficients:

             Estimate Std. Error z value Pr(>|z|)
ar1         1.182e-01  5.149e-02   2.295   0.0217 *
intercept -9.723e-06  2.968e-03  -0.003   0.9974
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> ftama1<-arima(fta.ts,order=c(0,0,1))
> ftama1

Call:
arima(x = fta.ts, order = c(0, 0, 1))

Coefficients:
        ma1  intercept
      0.122      0.000
s.e.  0.052      0.003

sigma^2 estimated as 0.00254:  log likelihood = 581.94,  aic = -1157.88

> coeftest(ftama1)

z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
ma1        1.223e-01  5.193e-02   2.355   0.0185 *
intercept -5.333e-06  2.937e-03  -0.002   0.9986
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> ftaar4<-arima(fta.ts,order=c(4,0,0))
> ftaar4

Call:
arima(x = fta.ts, order = c(4, 0, 0))

Coefficients:
        ar1     ar2    ar3    ar4  intercept
      0.116  -0.022  0.029  0.121      0.000
s.e.  0.051   0.052  0.052  0.051      0.003

sigma^2 estimated as 0.0025:  log likelihood = 585.06,  aic = -1158.13

> coeftest(ftaar4)

z test of coefficients:

            Estimate Std. Error z value Pr(>|z|)
ar1        1.161e-01  5.146e-02   2.256   0.0241 *
ar2       -2.232e-02  5.171e-02  -0.432   0.6660
ar3        2.949e-02  5.169e-02   0.571   0.5683
ar4        1.213e-01  5.129e-02   2.364   0.0181 *
intercept -5.138e-05  3.428e-03  -0.015   0.9880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> ftama4<-arima(fta.ts,order=c(0,0,4))
> ftama4

Call:
arima(x = fta.ts, order = c(0, 0, 4))

Coefficients:
        ma1     ma2    ma3    ma4  intercept
      0.131  -0.016  0.005  0.137      0.000
s.e.  0.052   0.052  0.052  0.050      0.003

sigma^2 estimated as 0.00249:  log likelihood = 585.55,  aic = -1159.1
> coeftest(ftama4)

z test of coefficients:

             Estimate Std. Error z value Pr(>|z|)
ma1         1.308e-01  5.153e-02   2.538  0.01114 *
ma2        -1.558e-02  5.243e-02  -0.297  0.76631
ma3         5.170e-03  5.200e-02   0.099  0.92081
ma4         1.371e-01  5.005e-02   2.739  0.00616 **
intercept -2.862e-05  3.255e-03  -0.009  0.99298
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The AIC values are −1159.10, −1158.13, −1157.88, and −1157.70 for the MA(4), AR(4), MA(1), and AR(1) models, respectively. The MA(4) fit, applied to the deseasonalized and outlier adjusted time series, is

$$y_t = -0.00003 + (1 + 0.1308B - 0.0156B^2 + 0.0052B^3 + 0.1371B^4)\varepsilon_t.$$

In autoregressive form, this is

$$(1 - 0.1308B + 0.0327B^2 - 0.0115B^3 - 0.1344B^4 + \cdots)(y_t + 0.00003) = \varepsilon_t.$$

The mathematics and R code to obtain this autoregressive representation are explained in the appendix to these notes.

The AR(4) fit, applied to the deseasonalized and outlier adjusted series, is

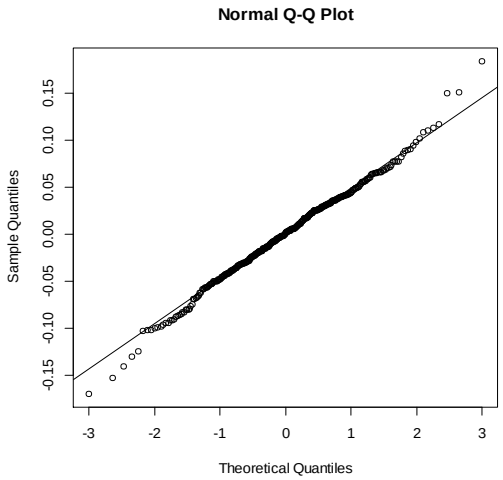
$$(1 - 0.1161B + 0.0223B^2 - 0.0295B^3 - 0.1213B^4)(y_t + 0.00005) = \varepsilon_t.$$

The two models are seen to be very similar.

The residual acf plots for the AR(4) and MA(4) fits have very slightly significant values at lags 10 and 16, and the residual pacf plots have very slightly significant values at lags 10 and 25. (The plots are not shown here.)

The normal quantile plot of the MA(4) residuals is given next.

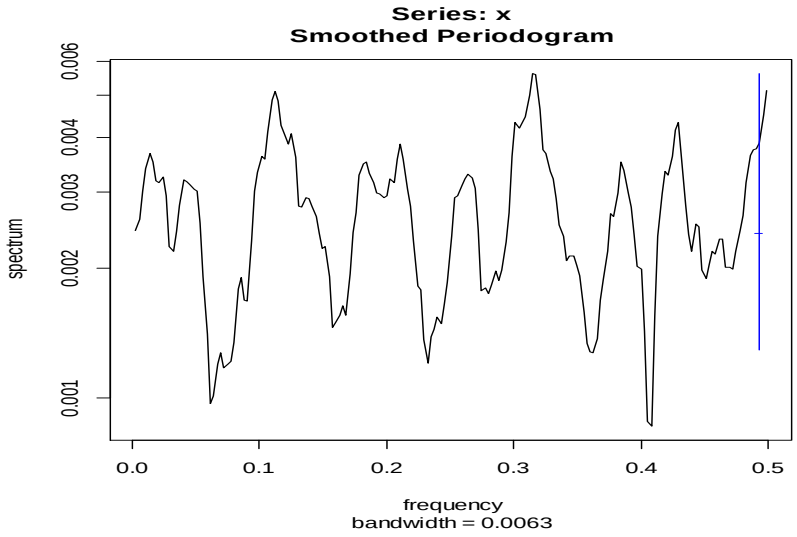
```
> qqnorm(resid(ftama4))
> qqline(resid(ftama4))
```



There are several mild outliers in each tail, but overall the plot is good. The picture for the AR(4) fit is similar.

The residual spectral plots follow. We start with the AR(4) fit.

```
> spectrum(ts(resid(ftaar4)),span=8)
```



Bartlett's Kolmogorov–Smirnov test for reduction to white noise is here:

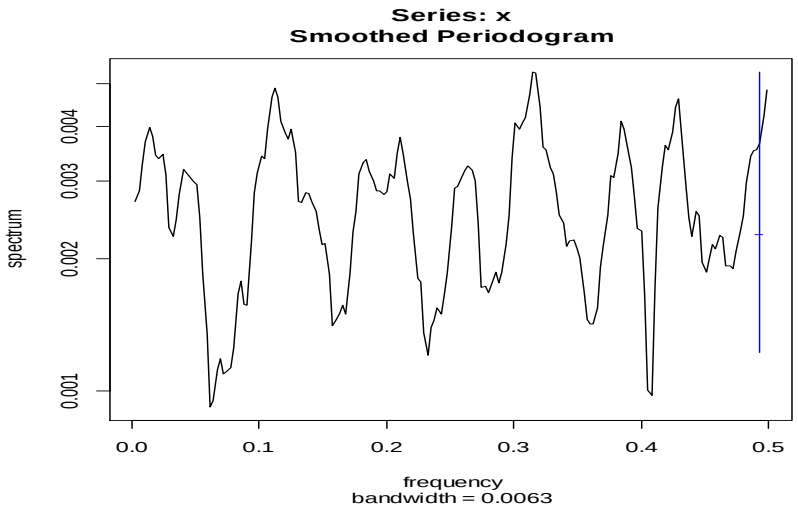
```
> library("hwwntest")
> bartlettB.test(ts(resid(ftaar4)))

        Bartlett B Test for white noise
```

```
data:
= 0.506065, p-value = 0.95993
```

The null hypothesis of reduction to white noise is not rejected—the *p*-value is much greater than 0.05.

Next, the results for the MA(4) fit.

```
> spectrum(ts(resid(ftama4)),span=8)
```




```
> bartlettB.test(ts(resid(ftama4)))

        Bartlett B Test for white noise

data:
= 0.344834, p-value = 0.99977
```

Again, the null hypothesis of reduction to white noise is not rejected.

The residual spectra for the AR(4) and MA(4) model fits have the same appearance.

Next, we explore the zeros of AR fits for both the equal-weighted and FTA series, in each case after deseasonalization and adjustment for prominent outliers. The AR(2) model on page 15 for the equal-weighted series is

$$(1-0.2779B+0.0884B^2)(y_t+0.0015)=\varepsilon_t.$$

Note this model does not adjust for calendar frequencies.

```
> #zeros of AR(2) polynomial
> zeros<-polyroot(c(1,-0.2779,0.0884))
> zeros
[1] 1.5718326+2.973476i 1.5718326-2.973476i
```

```
> #amplitude
> 1/Mod(zeros)[1]
[1] 0.29732137

> #period
> 2*pi/Arg(zeros)[1]
[1] 5.7935312
```

The model estimates a period of length 5.8 months. However, the amplitude is only 0.297, too weak to conclude that the estimated period is meaningful.

Calculations for the FTA time series follow, with use of the AR(4) model fit.

```
> zeros<-polyroot(c(1,-coef(ftaar4)[1:4]))
> zeros
[1]  0.0243101+1.6695521i -1.8719420-0.0000000i  0.0243101-1.6695521i
[4]  1.5801266-0.0000000i

> #amplitude
> 1/Mod(zeros)[1]
[1] 0.59889955

> #period
> 2*pi/Arg(zeros)[1]
[1] 4.0374232
```

There are two real-valued zeros and a complex conjugate pair. The model estimates a period of four months. The amplitude estimate is 0.599, large enough to suggest approximate periodic behavior. Is it meaningful? I am reluctant to conclude it is.

An approximate estimate of the annual growth rate for the FTA index follows.

```
> 12*mean(logreturn[2:372])
[1] 0.093242849
```

The annual estimate is 0.093243, which translates to an estimate of $100(\exp(0.093243) - 1) = 11.0$ per cent annual growth. By contrast, the estimate for the equal-weighted index on page 9 is 13.6 per cent.

Summary and additional remarks

1. Monthly log returns of the CRSP equal-weighted index are analyzed. The series has several prominent outliers, one in January 1975; a second in October 1987, stemming from 19 October, so-called Black Monday; and a third in August 1998, arising from the bombing of U.S. embassies in Nairobi and Dar es Salaam. AR(1) and MA(1) models both fit the data. The MA(1) model is essentially equivalent to an AR(2) fit.

2. Monthly log returns of the *Financial Times*-Actuaries (FTA) All Share index are also studied. This series ends in December 1995, and it also has outliers in January 1975 and October 1987, and an outlier in September 1981. Four models are fit to the data, AR(1), MA(1), AR(4), and MA(4). AIC values for the four are close together.

3. Each of the indices has some significant seasonal structure. The main estimated seasonal feature for the equal-weighted index is a significant effect in January. For the FTA index the estimated seasonal has significant positive values in April and January, and a significant negative value in May.

4. The growth rate for the equal-weighted index is estimated to be 13.6 per cent per year. The estimate for the FTA index is 11.0 per cent per year.

5. Both indices have very weak signals. While the AR and MA models fit to them are significant, the values of the parameters are small. That is, the signal-to-noise ratio is low for both indices, and the models fit to them have weak predictive power.

Appendix

The MA(4) model fit to the FTA data, from page 27 of these notes, is

```
            Estimate Std. Error z value Pr(>|z|)
ma1         1.308e-01  5.153e-02   2.538  0.01114 *
ma2        -1.558e-02  5.243e-02  -0.297  0.76631
ma3         5.170e-03  5.200e-02   0.099  0.92081
ma4         1.371e-01  5.005e-02   2.739  0.00616 **
intercept -2.862e-05  3.255e-03  -0.009  0.99298
```

We write the model as

$$y_t + 0.00003 = (1 + 0.1308B - 0.01558B^2 + 0.00517B^3 + 0.1371B^4)\varepsilon_t.$$

To obtain this in autoregressive form, we write

$$(1 + 0.1308B - 0.01558B^2 + 0.00517B^3 + 0.1371B^4)^{-1}(y_t + 0.00003) = \varepsilon_t.$$

To invert the polynomial, write

$$(1) \quad 1 = (1 + 0.1308B - 0.01558B^2 + 0.00517B^3 + 0.1371B^4)(\delta_0 + \delta_1 B + \delta_2 B^2 + \delta_3 B^3 + \cdots),$$

and determine the deltas, which will be the coefficients in the AR representation of the MA(4) model. The steps to obtain the deltas are the same those we have used to obtain 90 percent duration intervals for the Lydia Pinkham data.

Let ma_1, ma_2, ma_3, ma_4 denote the estimated MA(4) coefficients. Then from (1) we determine

$$\delta_0 = 1$$
$$\delta_1 = -ma_1 * \delta_0$$
$$\delta_2 = -ma_1 * \delta_1 - ma_2 * \delta_0$$
$$\delta_3 = -ma_1 * \delta_2 - ma_2 * \delta_1 - ma_3 * \delta_0$$
$$\delta_j = -ma_1 * \delta_{j-1} - ma_2 * \delta_{j-2} - ma_3 * \delta_{j-3} - ma_4 * \delta_{j-4},\ j = 4,5,\ldots.$$

The R code I have used to calculate these values is as follows.

```
ma4coef<-coef(ftama4)[1:4]
delta<-rep(0,10)
delta[1]<-1
for(i in 2:10){
for(j in 1:min(i-1,4)){
delta[i]<-delta[i]-ma4coef[j]*delta[i-j]
}
}
delta
```

And here are the calculations in R.

```
> ma4coef<-coef(ftama4)[1:4]
> ma4coef
         ma1          ma2          ma3          ma4
 0.130798467 -0.015583809  0.005169631  0.137079588

> ma4coef<-coef(ftama4)[1:4]
> delta<-rep(0,10)
> delta[1]<-1
> for(i in 2:10){
+ for(j in 1:min(i-1,4)){
+ delta[i]<-delta[i]-ma4coef[j]*delta[i-j]
+ }
+ }

> delta
 [1]  1.000000000 -0.130798467  0.032692048 -0.011484039 -0.134391847
 [6]  0.035160077 -0.011115265  0.004270771  0.017508786 -0.006985835
```