8 Diagnostic checking and Introduction to Forecasting

8.1 Overfitting

Overfitting is a method of diagnostic checking, in which, having identified what we believe to be an adequate model, we fit a more general model with an extra parameter and check whether the estimate of the extra parameter differs significantly from zero. If it does not, this provides evidence in support of the original model.

In the present case, from our earlier examination of the pacf, we might well suppose that an AR(1) model might be adequate.

Using the method of overfitting, we fit the AR(2) model, and find $\hat{\phi}_2 = -0.1235387$, with standard error $\sqrt{0.0079488694} = 0.089156$. The corresponding t-statistic with 120 degrees of freedom is given by

$$\frac{-0.1235387}{0.089156} = -1.386,$$

with corresponding p-value 0.168. Thus $\hat{\phi}_2$ does not differ significantly from zero, and we may conclude that the inclusion of a second autoregressive parameter does not significantly improve the fit of the model, i.e. the use of the AR(2) model does not give a significant improvement in fit over the use of the AR(1) model.

We may note that this conclusion is in line with the increase of the AIC in moving from the AR(1) to the AR(2) model.

8.2 Diagnostic checking of the residuals

To investigate the adequacy of the fitted model, we may examine the residuals or standardized residuals, which should appear to be similar to observations from a white noise process, uncorrelated with each other and identically distributed.

Furthermore, we may wish to assume that the white noise terms are normally distributed. Indeed, the use of the method of unconditional least squares for estimation is to a large degree justified by its close relationship to the method of maximum likelihood when it is assumed that the white noise terms are normally distributed. We may also wish to make the normality assumption for the purposes of forecasting. If this assumption is valid, the residuals should appear to be approximately normally distributed.

The residuals e_t are given by

$$e_t = y_t - \hat{y}_t, \tag{1}$$

where $\hat{y}_t$ is the fitted value, given for the AR(1) model by

$$\hat{y}_t - \hat{\mu} = \hat{\phi}(y_{t-1} - \hat{\mu}),$$

i.e.,

$$\hat{y}_t = (1 - \hat{\phi})\hat{\mu} + \hat{\phi} y_{t-1}, \qquad 2 \le t \le T.$$

For a general ARIMA model the residuals (1) are not always so easily expressed in simple terms. The residuals e_t or standardized residuals d_t,

$$d_t = \frac{e_t}{\hat{\sigma}},$$

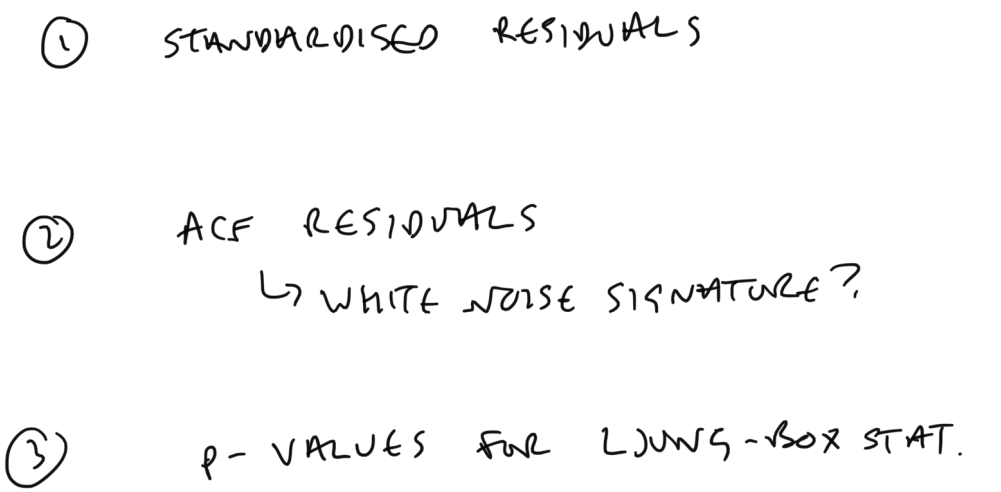
should then be plotted and their acf examined to see whether it appears to be the acf of a white noise process. The distribution of the standardized residuals may be examined to see whether it appears to be standard normal, using the usual techniques: descriptive statistics, histogram, plot of the normal scores.

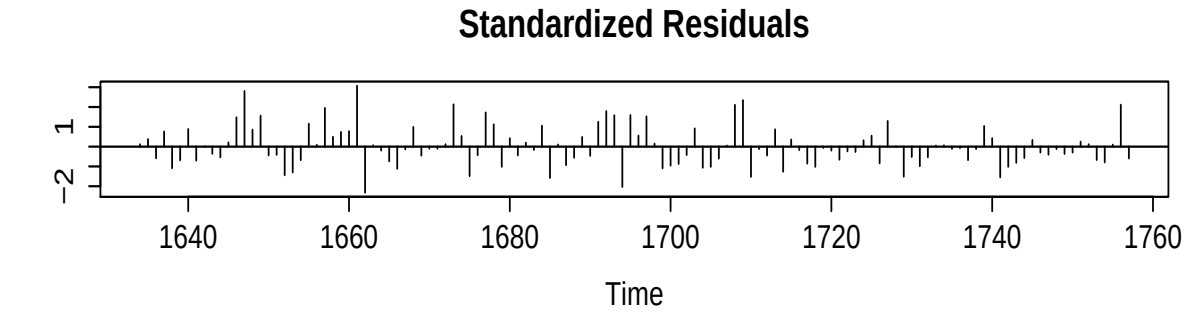
In examining the acf of the residuals, it should be noted that the joint distribution of the sample autocorrelations of the residuals after the fitting of an appropriate ARIMA model is not the same as the joint distribution of the sample autocorrelations from a white noise process. Recall that in the latter case the sample autocorrelations are approximately NID$(0, 1/T)$. A substantial reduction in the variance can occur at low lags for the sample autocorrelations of the residuals after the fitting of an ARIMA model, and, furthermore, the sample autocorrelations at low lags can be highly correlated with each other. However, these effects usually disappear rather quickly at higher lags. For example, for the residuals from a fitted AR(1) model, the only change worth noting is that $\text{var}(r_1) \approx \phi^2/T$. In general, however, $1/T$ must be regarded as an upper bound for the variances of the r_τ. For low lags, use of the standard error $1/\sqrt{T}$ may seriously underestimate the significance of apparent differences from zero.

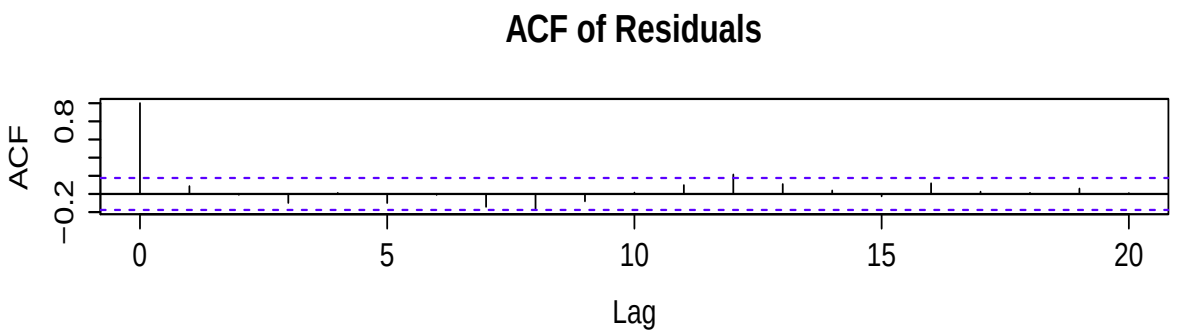
8.3 Bread price example: diagnostic checking

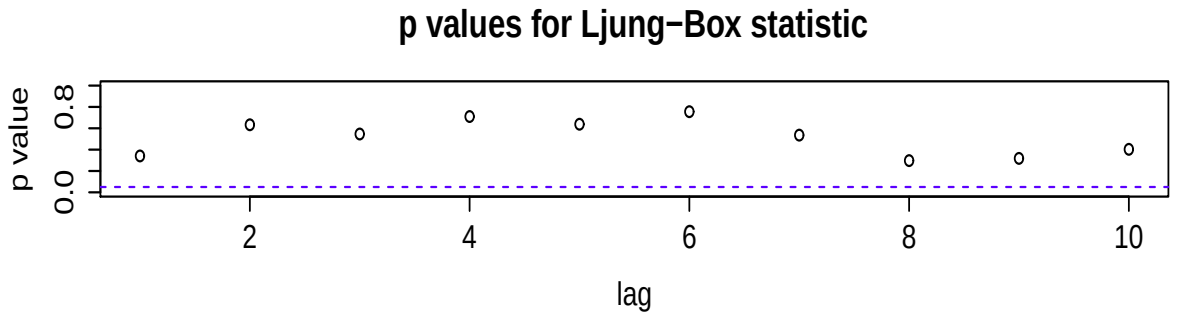
To investigate the adequacy of the fitted model, we may examine the standardized residuals, which should appear to be similar to observations from a white noise process, uncorrelated with each other and identically distributed. For the bread price data, we apply the function `tsdiag` (for ARIMA *diagnostics*) to the object `BP.ar1m`. There is no object created or any value to the function `tsdiag`: just plots of the diagnostics are produced.

Plots of the standardized residuals, the acf of the residuals, and the p-values of the portmanteau statistics (which will be discussed in the next section) are produced.




Standardized Residuals
1
−2
1640
1660
1680
1700
1720
1740
1760
Time


ACF of Residuals
ACF
0.8
−0.2
0
5
10
15
20
Lag


p values for Ljung−Box statistic
p value
0.8
0.0
2
4
6
8
10
lag

8.4 The portmanteau statistics

Given a lag K, the Box-Pierce test statistic based upon the first K autocorrelations is

$$Q_K^* = T\sum_{\tau=1}^{K} r_\tau^2,$$

where the r_τ here are the sample autocorrelations of the residuals. If the r_τ are the sample autocorrelations of the residuals from fitting an ARIMA(p, d, q) model, and this model is correct, then it turns out that Q_K^* has approximately the χ^2_{K-p-q} distribution.

An improved statistic, whose distribution, assuming that the model is correct, is closer to the χ^2_{K-p-q} distribution, is the *modified Box-Pierce statistic* or *Ljung-Box statistic*,

$$Q_K = T(T+2)\sum_{\tau=1}^{K} \frac{r_\tau^2}{T-\tau}.$$

These are known as the portmanteau statistics.

The `tsdiag` function computes the Q_K for values of K up to a value set by the `gof.lag` parameter within the function (which defaults to 10). If any of the chi-square values is significant, we may reject the null hypothesis that the chosen model is correct. In our example of fitting the AR(1) model to the bread price data, none of the chi-square values is significant at the 5% level as indicated by the corresponding p-values in the diagnostic plot. Thus the AR(1) model appears quite adequate.

8.5 Modelling the airline passenger data

We consider the series $\{y_t\}$, where y_t is the logarithm of the number of passengers in thousands in month t. In the following R output, the object `lair.ts` contains the values of the series $\{y_t\}$. The acf of this series is calculated. Then the series is differenced to create the object `laird.ts` that contains the differenced series $\{\triangle y_t\}$. The acf of this series is calculated. Then the series is differenced again, this time using seasonal differencing at lag 12, to create the object `lairdd.rts` that contains the series $\{\triangle_{12}\triangle y_t\}$. The acf of this series too is calculated. Finally, for the purposes of presentation, a data frame `airacf` is created that contains a tabulation of all three acf's. (In doing this we have to bear in mind that the component `lair.acf$acf` of the object `lair.acf` is a $41 \times 1 \times 1$ array, and similarly for the other `acf` objects.)

```
> lair.ts <- log(air.ts)
> plot(lair.ts, xlab="", ylab="",
+           main = "Monthly Log Totals of Airline Passengers", las = 1)
> lair.acf <- acf(lair.ts,40)
> laird.ts <- diff(lair.ts)
> laird.acf <- acf(laird.ts,40)
> lairdd.ts <- diff(laird.ts, lag=12)
> lairdd.acf <- acf(lairdd.ts,40)
> lag <- 0:40
> lair <- lair.acf$acf[,,1]
> laird <- laird.acf$acf[,,1]
> lairdd <- lairdd.acf$acf[,,1]
> airacf <- data.frame(lag,lair,laird,lairdd)
```

Recall: $\Delta_s = (1-L^s)$ $\Delta^d = (1-L)^d$

$\Delta_s^D = (1-L^s)^D$

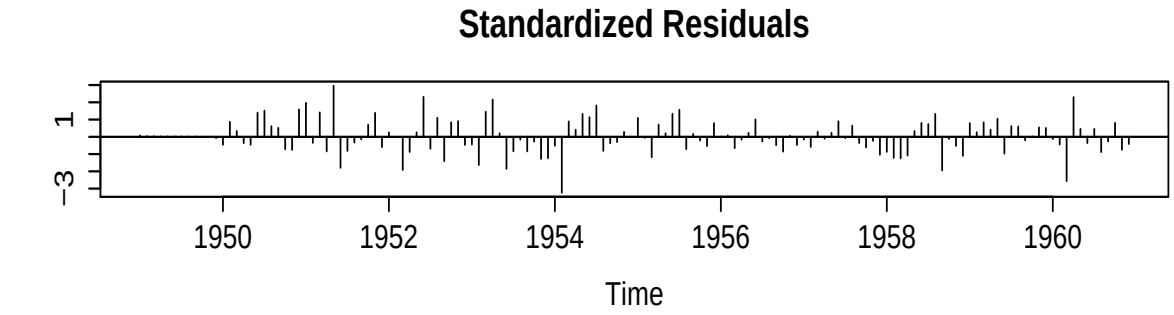
Difference operator

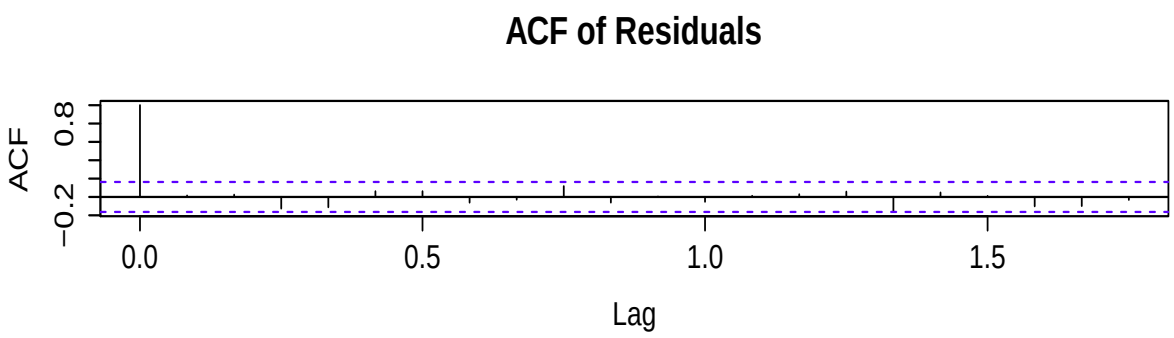
$\{y_t\}$ $\{\Delta y_t\}$ $\{\Delta_{12}\Delta y_t\}$

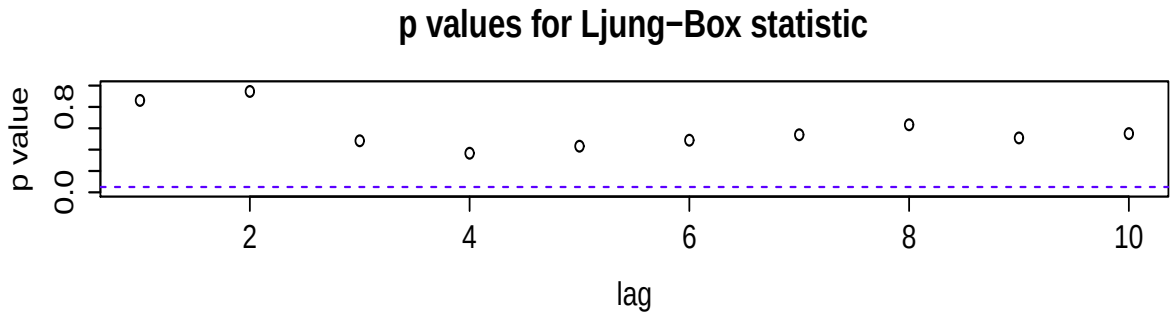
```
> airacf[1:27,]
   lag      lair       laird        lairdd
1    0 1.0000000  1.00000000  1.0000000000
2    1 0.9537034  0.19975134 -0.3411237983
3    2 0.8989159 -0.12010433  0.1050467496
4    3 0.8508025 -0.15077204 -0.2021386642
5    4 0.8084252 -0.32207432  0.0213592288
6    5 0.7788994 -0.08397453  0.0556543435
7    6 0.7564422  0.02577843  0.0308036696
8    7 0.7376017 -0.11096075 -0.0555785695
9    8 0.7271313 -0.33672146 -0.0007606578
10   9 0.7336487 -0.11558631  0.1763686815
11  10 0.7442552 -0.10926704 -0.0763581912
12  11 0.7580266  0.20585223  0.0643839399
13  12 0.7619429  0.84142998 -0.3866128596
14  13 0.7165045  0.21508704  0.1516020121
15  14 0.6630428 -0.13955394 -0.0576067980
16  15 0.6183629 -0.11599576  0.1495652202
17  16 0.5762087 -0.27894284 -0.1389421819
18  17 0.5438013 -0.05170646  0.0704823385
19  18 0.5194561  0.01245814  0.0156307241
20  19 0.5007029 -0.11435760 -0.0106106130
21  20 0.4904028 -0.33717439 -0.1167285978
22  21 0.4981819 -0.10738490  0.0385542023
23  22 0.5061666 -0.07521120 -0.0913645276
24  23 0.5167434  0.19947518  0.2232689055
25  24 0.5204897  0.73692070 -0.0184181674
26  25 0.4835237  0.19726236 -0.1002881161
27  26 0.4373983 -0.12388430  0.0485657567
```

After the two differencing operations the resulting acf looks like that of a stationary process. There are a still a number of autocorrelations that are significantly different from zero, most notably the ones at lags 1 and 12. This suggests that we need to introduce moving average terms and, in particular, that we might try to fit an ARIMA$(0,1,1) \times (0,1,1)_{12}$ model. This is done in the following R output.

```
> lair.arima <- arima(lair.ts, order = c(0, 1, 1), list(order = c(0, 1, 1), period = 12))
> lair.arima

Call:
arima(x = lair.ts, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ma1     sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,  aic = -483.4
> tsdiag(lair.arima)
```


Standardized Residuals
1
−3
1950
1952
1954
1956
1958
1960
Time


ACF of Residuals
ACF
0.8
−0.2
0.0
0.5
1.0
1.5
Lag


p values for Ljung−Box statistic
p value
0.8
0.0
2
4
6
8
10
lag

The model turns out to be very satisfactory. When we compare the moving average parameter estimates with their standard errors, we see that they are both highly significant, so that we cannot drop either of them without significant loss of fit, in the sense of a significant increase in the estimated error variance.

In some other statistical packages, an alternative specification of the moving average parameters in the model is used, in which they have a negative sign in front of them. Thus the ARIMA$(0,1,1)\times(0,1,1)_{12}$ model is taken in the form

$$(1-L)(1-L^{12})Y_t = (1-\theta L)(1-\Theta L^{12})\epsilon_t.$$

However, when using the `arima` function in R, our usual specification with the positive signs in front of the moving average parameters is used. Thus ARIMA$(0,1,1)\times(0,1,1)_{12}$ is stated as

$$(1-L)(1-L^{12})Y_t = (1+\theta L)(1+\Theta L^{12})\epsilon_t.$$

In the present case, the fitted model is

$$Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} + \epsilon_t - 0.402\epsilon_{t-1} - 0.557\epsilon_{t-12} + 0.224\epsilon_{t-13}.$$

8.6 Forecasting - Introductory Concepts

Suppose that we have observed an ARIMA process $\{Y_t\}$ up to time T, so that we have the observations $y_1, y_2, \ldots, y_T$, and that we wish to forecast (predict) the value y_{T+h} of the process at time $T+h$, where $h \geq 1$. The time T is known as the *origin* of the forecast and the time difference h as the *lead time* of the forecast.

We shall make the following two assumptions.

1. The parameter values of the ARIMA process are known. In practice, we shall have only estimates of the parameter values, but if the estimates are based upon a long series of observations then they will be accurate enough for us to be able for practical purposes to regard the parameter values as known.

2. The underlying white noise process $\{\epsilon_t\}$ is such that the ϵ_t are NID$(0, \sigma^2)$. It follows that the Y_t are also normally distributed.

In this section, the distinction between the use of upper case letters for process random variables and lower case letters for the values that they take will be maintained but will sometimes be obscured.

Let $\mathcal{H}_T$ denote the observations $y_1, y_2, \ldots, y_T$ up to time T. Often, the theory that we shall develop ideally requires that we have observed the process back into the infinite past, in which case $\mathcal{H}_T$ denotes the observations $\ldots, y_{-1}, y_0, y_1, \ldots, y_T$. If T is large, this makes little practical difference. We may think of $\mathcal{H}_T$ as representing the history of the process up to time T.

The *minimum mean square error forecast*, $\hat{y}_T(h)$, is the function $\hat{y}$ of $\mathcal{H}_T$ that minimizes

$$E\left[(Y_{T+h} - \hat{y})^2 | \mathcal{H}_T\right].$$

It follows that $\hat{y}_T(h)$ is given by the conditional expectation,

$$\hat{y}_T(h) = E[Y_{T+h}|\mathcal{H}_T]. \tag{2}$$

The *forecast error*, $e_T(h)$, is defined by

$$e_T(h) = Y_{T+h} - \hat{y}_T(h).$$

From Equation (2),

$$E[e_T(h)|\mathcal{H}_T] = E[Y_{T+h} - \hat{y}_T(h)|\mathcal{H}_T] = E[Y_{T+h}|\mathcal{H}_T] - \hat{y}_T(h) = 0.$$

Thus $\hat{y}_T(h)$ is an *unbiased* forecast of Y_{T+h}. It follows that, unconditionally,

$$E[e_T(h)] = 0.$$

In fact it turns out that $e_T(h)$ is independent of $\mathcal{H}_T$. The *mean square error* or the *forecast error variance*, $V(h)$, depends essentially only on h and not on T and $\mathcal{H}_T$. It is given by

$$V(h) = \text{var}(e_T(h)).$$

Using the normality assumption, $100(1-\alpha)\%$ *prediction limits* or *probability limits* for Y_{T+h} are given by

$$\hat{y}_T(h) \pm z_{\alpha/2}\sqrt{V(h)},$$

where $z_{\alpha/2}$ is a percentage point of the standard normal distribution.

8.7 Forecasting for ARMA models

Consider the ARMA(p, q) model as specified in Chapter 6,

$$Y_t = \mu + \sum_{k=1}^{p} \phi_k(Y_{t-k} - \mu) + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}. \tag{3}$$

The infinite moving average expression of the corresponding stationary process $\{Y_t\}$ is given by

$$Y_t = \mu + \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}. \tag{4}$$

Setting $t = T + h$ in Equation (4),

$$\begin{aligned} Y_{T+h} &= \mu + \sum_{i=0}^{\infty} \psi_i \epsilon_{T+h-i} \\ &= \hat{y}_T(h) + \sum_{i=0}^{h-1} \psi_i \epsilon_{T+h-i}, \end{aligned} \tag{5}$$

where $\hat{y}_T(h)$ is the forecast with origin T and lead time h,

$$\hat{y}_T(h) = E[Y_{T+h}|\mathcal{H}_T] = \mu + \sum_{i=h}^{\infty} \psi_i \epsilon_{T+h-i}.$$

Note that $\hat{y}_T(h)$ is a function of $\mathcal{H}_T$. The second term on the right hand side of Equation (5) gives the forecast error $e_T(h)$,

$$e_T(h) = \sum_{i=0}^{h-1} \psi_i \epsilon_{T+h-i}, \tag{6}$$

from which it follows that

$$V(h) = \text{var}(e_T(h)) = \sum_{i=0}^{h-1} \psi_i^2 \sigma^2. \tag{7}$$

In particular, $\psi_0 = 1, e_T(1) = \epsilon_{T+1}$ and $V(1) = \sigma^2$. From Equation (7) it further follows that, as $h \to \infty$,

$$V(h) \uparrow \sum_{i=0}^{\infty} \psi_i^2 \sigma^2 = \gamma_0.$$

Given T, the forecasts $\hat{y}_T(h)$ may be obtained recursively for $h = 1, 2, \ldots$, using the following approach. Setting $t = T + h$ in Equation (3),

$$Y_{T+h} = \mu + \sum_{k=1}^{p} \phi_k (Y_{T+h-k} - \mu) + \epsilon_{T+h} + \sum_{i=1}^{q} \theta_i \epsilon_{T+h-i}. \tag{8}$$

Now note that

$$E[\epsilon_{T+j} | \mathcal{H}_T] = \begin{cases} 0, & j > 0 \\ \epsilon_{T+j}, & j \leq 0 \end{cases}$$

and

$$E[Y_{T+j} | \mathcal{H}_T] = \begin{cases} \hat{y}_T(j), & j > 0 \\ y_{T+j}, & j \leq 0 \end{cases}.$$

Taking expectations conditional upon $\mathcal{H}_T$ in Equation (8),

$$\hat{y}_T(h) = \mu + \sum_{k=1}^{\min(h-1,p)} \phi_k (\hat{y}_T(h-k) - \mu) + \sum_{k=h}^{p} \phi_k (y_{T+h-k} - \mu) + \sum_{i=h}^{q} \theta_i \epsilon_{T+h-i}, \qquad h \geq 1. \tag{9}$$

Solving Equation (9) for $h = 1, 2, \ldots$ will yield the *forecast function* $\{\hat{y}_T(h) : h \geq 1\}$. In particular, for $h > \max(p, q)$, Equation (9) reduces to

$$\hat{y}_T(h) = \mu + \sum_{k=1}^{p} \phi_k (\hat{y}_T(h-k) - \mu). \tag{10}$$

Equation (10) is a linear difference equation of the type studied in Chapter 5 with a general solution of the form

$$\hat{y}_T(h) = \mu + \sum_{k=1}^{p} A_k \alpha_k^h,$$

where the α_k satisfy $|\alpha_k| < 1,\ 1 \leq k \leq p$. The values of the A_k will depend upon $\mathcal{H}_T$, but, whatever their values, as $h \to \infty$

$$\hat{y}_T(h) \to \mu.$$