CHAPTER 9

FORECASTING

One of the primary objectives of building a model for a time series is to be able to forecast the values for that series at future times. Of equal importance is the assessment of the precision of those forecasts. In this chapter, we shall consider the calculation of forecasts and their properties for both deterministic trend models and ARIMA models. Forecasts for models that combine deterministic trends with ARIMA stochastic components are considered also.

For the most part, we shall assume that the model is known *exactly*, including specific values for all the parameters. Although this is never true in practice, the use of estimated parameters for large sample sizes does not seriously affect the results.

9.1 Minimum Mean Square Error Forecasting

Based on the available history of the series up to time t, namely $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$, we would like to forecast the value of $Y_{t+\ell}$ that will occur ℓ time units into the future. We call time t the **forecast origin** and ℓ the **lead time** for the forecast, and denote the forecast itself as $\hat{Y}_t(\ell)$.

As shown in Appendix F, the minimum mean square error forecast is given by

$$\hat{Y}_t(\ell) = E(Y_{t+\ell} \mid Y_1, Y_2, \ldots, Y_t) \tag{9.1.1}$$

(Appendices E and F on page 218 review the properties of conditional expectation and minimum mean square error prediction.)

The computation and properties of this conditional expectation as related to forecasting will be our concern for the remainder of this chapter.

9.2 Deterministic Trends

Consider once more the deterministic trend model of Chapter 3,

$$Y_t = \mu_t + X_t \tag{9.2.1}$$

where the stochastic component, X_t, has a mean of zero. For this section, we shall assume that $\{X_t\}$ is in fact white noise with variance γ_0. For the model in Equation (9.2.1), we have

$$\begin{aligned}\hat{Y}_t(\ell) &= E(\mu_{t+\ell} + X_{t+\ell} | Y_1, Y_2, \ldots, Y_t) \\ &= E(\mu_{t+\ell} | Y_1, Y_2, \ldots, Y_t) + E(X_{t+\ell} | Y_1, Y_2, \ldots, Y_t) \\ &= \mu_{t+\ell} + E(X_{t+\ell})\end{aligned}$$

or

$$\hat{Y}_t(\ell) = \mu_{t+\ell} \tag{9.2.2}$$

since for $\ell \geq 1$, $X_{t+\ell}$ is independent of $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$ and has expected value zero. Thus, in this simple case, forecasting amounts to extrapolating the deterministic time trend into the future.

For the linear trend case, $\mu_t = \beta_0 + \beta_1 t$, the forecast is

$$\hat{Y}_t(\ell) = \beta_0 + \beta_1(t + \ell) \tag{9.2.3}$$

As we emphasized in Chapter 3, this model assumes that the *same* linear time trend persists into the future, and the forecast reflects that assumption. Note that it is the lack of statistical dependence between $Y_{t+\ell}$ and $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$ that prevents us from improving on $\mu_{t+\ell}$ as a forecast.

For seasonal models where, say, $\mu_t = \mu_{t+12}$, our forecast is $\hat{Y}_t(\ell) = \mu_{t+12+\ell} = \hat{Y}_t(\ell + 12)$. Thus the forecast will also be periodic, as desired.

The **forecast error**, $e_t(\ell)$, is given by

$$\begin{aligned}e_t(\ell) &= Y_{t+\ell} - \hat{Y}_t(\ell) \\ &= \mu_{t+\ell} + X_{t+\ell} - \mu_{t+\ell} \\ &= X_{t+\ell}\end{aligned}$$

so that

$$E(e_t(\ell)) = E(X_{t+\ell}) = 0$$

That is, the forecasts are **unbiased**. Also

$$Var(e_t(\ell)) = Var(X_{t+\ell}) = \gamma_0 \tag{9.2.4}$$

is the **forecast error variance** for all lead times ℓ.

The cosine trend model for the average monthly temperature series was estimated in Chapter 3 on page 35 as

$$\hat{\mu}_t = 46.2660 + (-26.7079)\cos(2\pi t) + (-2.1697)\sin(2\pi t)$$

Here time is measured in years with a starting value of January 1964, frequency $f = 1$ per year, and the final observed value is for December 1975. To forecast the June 1976 temperature value, we use $t = 1976.41667$ as the time value[†] and obtain

[†] June is the fifth month of the year, and $5/12 \approx 0.416666666\ldots$.

$$\hat{\mu}_t = 46.2660 + (-26.7079)\cos(2\pi(1976.41667)) + (-2.1697)\sin(2\pi(1976.41667))$$
$$= 68.3\ ^\circ\text{F}$$

Forecasts for other months are obtained similarly.

9.3 ARIMA Forecasting

For ARIMA models, the forecasts can be expressed in several different ways. Each expression contributes to our understanding of the overall forecasting procedure with respect to computing, updating, assessing precision, or long-term forecasting behavior.

AR(1)

We shall first illustrate many of the ideas with the simple AR(1) process with a nonzero mean that satisfies

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + e_t \tag{9.3.1}$$

Consider the problem of forecasting one time unit into the future. Replacing t by $t + 1$ in Equation (9.3.1), we have

$$Y_{t+1} - \mu = \phi(Y_t - \mu) + e_{t+1} \tag{9.3.2}$$

Given $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$, we take the conditional expectations of both sides of Equation (9.3.2) and obtain

$$\hat{Y}_t(1) - \mu = \phi[E(Y_t|Y_1, Y_2, \ldots, Y_t) - \mu] + E(e_{t+1}|Y_1, Y_2, \ldots, Y_t) \tag{9.3.3}$$

Now, from the properties of conditional expectation, we have

$$E(Y_t|Y_1, Y_2, \ldots, Y_t) = Y_t \tag{9.3.4}$$

Also, since e_{t+1} is independent of $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$, we obtain

$$E(e_{t+1}|Y_1, Y_2, \ldots, Y_t) = E(e_{t+1}) = 0 \tag{9.3.5}$$

Thus, Equation (9.3.3) can be written as

$$\hat{Y}_t(1) = \mu + \phi(Y_t - \mu) \tag{9.3.6}$$

In words, a proportion ϕ of the current deviation from the process mean is added to the process mean to forecast the next process value.

Now consider a general lead time ℓ. Replacing t by $t + \ell$ in Equation (9.3.1) and taking the conditional expectations of both sides produces

$$\hat{Y}_t(\ell) = \mu + \phi[\hat{Y}_t(\ell - 1) - \mu] \quad \text{for } \ell \geq 1 \tag{9.3.7}$$

since $E(Y_{t+\ell-1}|Y_1, Y_2, \ldots, Y_t) = \hat{Y}_t(\ell - 1)$ and, for $\ell \geq 1$, $e_{t+\ell}$ is independent of $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$.

Equation (9.3.7), which is recursive in the lead time ℓ, shows how the forecast for any lead time ℓ can be built up from the forecasts for shorter lead times by starting with the initial forecast $\hat{Y}_t(1)$ computed using Equation (9.3.6). The forecast $\hat{Y}_t(2)$ is then obtained from $\hat{Y}_t(2) = \mu + \phi[\hat{Y}_t(1) - \mu]$, then $\hat{Y}_t(3)$ from $\hat{Y}_t(2)$, and so on until the desired $\hat{Y}_t(\ell)$ is found. Equation (9.3.7) and its generalizations for other ARIMA models are most convenient for actually computing the forecasts. Equation (9.3.7) is sometimes called the **difference equation form** of the forecasts.

However, Equation (9.3.7) can also be solved to yield an explicit expression for the forecasts in terms of the observed history of the series. Iterating backward on ℓ in Equation (9.3.7), we have

$$\begin{aligned}\hat{Y}_t(\ell) &= \phi[\hat{Y}_t(\ell-1) - \mu] + \mu \\ &= \phi\{\phi[\hat{Y}_t(\ell-2) - \mu]\} + \mu \\ &\vdots \\ &= \phi^{\ell-1}[\hat{Y}_t(1) - \mu] + \mu\end{aligned}$$

or

$$\hat{Y}_t(\ell) = \mu + \phi^{\ell}(Y_t - \mu) \tag{9.3.8}$$

The current deviation from the mean is discounted by a factor ϕ^{ℓ}, whose magnitude decreases with increasing lead time. The discounted deviation is then added to the process mean to produce the lead ℓ forecast.

As a numerical example, consider the AR(1) model that we have fitted to the industrial color property time series. The maximum likelihood estimation results were partially shown in Exhibit 7.7 on page 165, but more complete results are shown in Exhibit 9.1.

Exhibit 9.1 Maximum Likelihood Estimation of an AR(1) Model for Color

Coefficients:	ar1	intercept[†]
	0.5705	74.3293
s.e.	0.1435	1.9151

sigma^2 estimated as 24.8: log-likelihood = –106.07, AIC = 216.15

[†]Remember that the intercept here is the estimate of the process mean μ—not θ_0.

```
> data(color)
> m1.color=arima(color,order=c(1,0,0))
> m1.color
```

For illustration purposes, we assume that the estimates $\phi = 0.5705$ and $\mu = 74.3293$ are true values. The final forecasts may then be rounded.

The last observed value of the color property is 67, so we would forecast one time period ahead as[†]

$$\begin{aligned}\hat{Y}_t(1) &= 74.3293 + (0.5705)(67 - 74.3293)\\ &= 74.3293 - 4.181366\\ &= 70.14793\end{aligned}$$

For lead time 2, we have from Equation (9.3.7)

$$\begin{aligned}\hat{Y}_t(2) &= 74.3293 + 0.5705(70.14793 - 74.3293)\\ &= 74.3293 - 2.385472\\ &= 71.94383\end{aligned}$$

Alternatively, we can use Equation (9.3.8):

$$\begin{aligned}\hat{Y}_t(2) &= 74.3293 + (0.5705)^2(67 - 74.3293)\\ &= 71.92823\end{aligned}$$

At lead 5, we have

$$\begin{aligned}\hat{Y}_t(5) &= 74.3293 + (0.5705)^5(67 - 74.3293)\\ &= 73.88636\end{aligned}$$

and by lead 10 the forecast is

$$\hat{Y}_t(10) = 74.30253$$

which is very nearly μ (= 74.3293). In reporting these forecasts we would probably round to the nearest tenth.

In general, since $|\phi| < 1$, we have simply

$$\hat{Y}_t(\ell) \approx \mu \quad \text{for large } \ell \tag{9.3.9}$$

Later we shall see that Equation (9.3.9) holds for *all stationary* ARMA models.

Consider now the **one-step-ahead forecast error**, $e_t(1)$. From Equations (9.3.2) and (9.3.6), we have

$$\begin{aligned}e_t(1) &= Y_{t+1} - \hat{Y}_t(1)\\ &= [\phi(Y_t - \mu) + \mu + e_{t+1}] - [\phi(Y_t - \mu) + \mu]\end{aligned}$$

or

$$e_t(1) = e_{t+1} \tag{9.3.10}$$

[†] As round off error will accumulate, you should use many decimal places when performing recursive calculations.

The white noise process $\{e_t\}$ can now be reinterpreted as a sequence of one-step-ahead forecast errors. We shall see that Equation (9.3.10) persists for completely general ARIMA models. Note also that Equation (9.3.10) implies that the forecast error $e_t(1)$ is independent of the history of the process $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$ up to time t. If this were not so, the dependence could be exploited to improve our forecast.

Equation (9.3.10) also implies that our one-step-ahead forecast error variance is given by

$$Var(e_t(1)) = \sigma_e^2 \tag{9.3.11}$$

To investigate the properties of the forecast errors for longer leads, it is convenient to express the AR(1) model in general linear process, or MA(∞), form. From Equation (4.3.8) on page 70, we recall that

$$Y_t = e_t + \phi e_{t-1} + \phi^2 e_{t-2} + \phi^3 e_{t-3} + \cdots \tag{9.3.12}$$

Then Equations (9.3.8) and (9.3.12) together yield

$$\begin{aligned} e_t(\ell) &= Y_{t+\ell} - \mu - \phi^\ell (Y_t - \mu) \\ &= e_{t+\ell} + \phi e_{t+\ell-1} + \cdots + \phi^{\ell-1} e_{t+1} + \phi^\ell e_t \\ &\qquad + \cdots - \phi^\ell (e_t + \phi e_{t-1} + \cdots) \end{aligned}$$

so that

$$e_t(\ell) = e_{t+\ell} + \phi e_{t+\ell-1} + \cdots + \phi^{\ell-1} e_{t+1} \tag{9.3.13}$$

which can also be written as

$$e_t(\ell) = e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1} e_{t+1} \tag{9.3.14}$$

Equation (9.3.14) will be shown to hold for *all* ARIMA models (see Equation (9.3.43) on page 202).

Note that $E(e_t(\ell)) = 0$; thus the forecasts are **unbiased**. Furthermore, from Equation (9.3.14), we have

$$Var(e_t(\ell)) = \sigma_e^2 (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{\ell-1}^2) \tag{9.3.15}$$

We see that the forecast error variance increases as the lead ℓ increases. Contrast this with the result given in Equation (9.2.4) on page 192, for deterministic trend models.

In particular, for the AR(1) case,

$$Var(e_t(\ell)) = \sigma_e^2 \left[\frac{1 - \phi^{2\ell}}{1 - \phi^2} \right] \tag{9.3.16}$$

which we obtain by summing a finite geometric series.

For long lead times, we have

$$Var(e_t(\ell)) \approx \frac{\sigma_e^2}{1-\phi^2} \quad \text{for large } \ell \tag{9.3.17}$$

or, by Equation (4.3.3), page 66,

$$Var(e_t(\ell)) \approx Var(Y_t) = \gamma_0 \quad \text{for large } \ell \tag{9.3.18}$$

Equation (9.3.18) will be shown to be valid for *all stationary* ARMA processes (see Equation (9.3.39) on page 201).

MA(1)

To illustrate how to solve the problems that arise in forecasting moving average or mixed models, consider the MA(1) case with nonzero mean:

$$Y_t = \mu + e_t - \theta e_{t-1}$$

Again replacing t by $t+1$ and taking conditional expectations of both sides, we have

$$\hat{Y}_t(1) = \mu - \theta E(e_t | Y_1, Y_2, \ldots, Y_t) \tag{9.3.19}$$

However, for an invertible model, Equation (4.5.2) on page 80 shows that e_t is a function of $Y_1, Y_2, \ldots, Y_t$ and so

$$E(e_t | Y_1, Y_2, \ldots, Y_t) = e_t \tag{9.3.20}$$

In fact, an approximation is involved in this equation since we are conditioning only on $Y_1, Y_2, \ldots, Y_t$ and not on the infinite history of the process. However, if, as in practice, t is large and the model is invertible, the error in the approximation will be very small. If the model is not invertible—for example, if we have overdifferenced the data—then Equation (9.3.20) is not even approximately valid; see Harvey (1981c, p.161).

Using Equations (9.3.19) and (9.3.20), we have the one-step-ahead forecast for an invertible MA(1) expressed as

$$\hat{Y}_t(1) = \mu - \theta e_t \tag{9.3.21}$$

The computation of e_t will be a by-product of estimating the parameters in the model.

Notice once more that the one-step-ahead forecast error is

$$\begin{aligned} e_t(1) &= Y_{t+1} - \hat{Y}_t(1) \\ &= (\mu + e_{t+1} - \theta e_t) - (\mu - \theta e_t) \\ &= e_{t+1} \end{aligned}$$

as in Equation (9.3.10), and thus Equation (9.3.11) also obtains.

For longer lead times, we have

$$\hat{Y}_t(\ell) = \mu + E(e_{t+\ell} | Y_1, Y_2, \ldots, Y_t) - \theta E(e_{t+\ell-1} | Y_1, Y_2, \ldots, Y_t)$$

But, for $\ell > 1$, both $e_{t+\ell}$ and $e_{t+\ell-1}$ are independent of $Y_1, Y_2, \ldots, Y_t$. Consequently, these conditional expected values are the unconditional expected values, namely zero, and we have

$$\hat{Y}_t(\ell) = \mu \quad \text{for } \ell > 1 \tag{9.3.22}$$

Notice here that Equation (9.3.9) on page 195 holds exactly for the MA(1) case when $\ell > 1$. Since for this model we trivially have $\psi_1 = -\theta$ and $\psi_j = 0$ for $j > 1$, Equations (9.3.14) and (9.3.15) also hold.

The Random Walk with Drift

To illustrate forecasting with nonstationary ARIMA series, consider the random walk with drift defined by

$$Y_t = Y_{t-1} + \theta_0 + e_t \tag{9.3.23}$$

Here

$$\hat{Y}_t(1) = E(Y_t | Y_1, Y_2, \ldots, Y_t) + \theta_0 + E(e_{t+1} | Y_1, Y_2, \ldots, Y_t)$$

so that

$$\hat{Y}_t(1) = Y_t + \theta_0 \tag{9.3.24}$$

Similarly, the difference equation form for the lead ℓ forecast is

$$\hat{Y}_t(\ell) = \hat{Y}_t(\ell - 1) + \theta_0 \quad \text{for } \ell \geq 1 \tag{9.3.25}$$

and iterating backward on ℓ yields the explicit expression

$$\hat{Y}_t(\ell) = Y_t + \theta_0 \ell \quad \text{for } \ell \geq 1 \tag{9.3.26}$$

In contrast to Equation (9.3.9) on page 195, if $\theta_0 \neq 0$, the forecast does not converge for long leads but rather follows a straight line with slope θ_0 for all ℓ.

Note that the presence or absence of the constant term θ_0 significantly alters the nature of the forecast. For this reason, constant terms should not be included in nonstationary ARIMA models unless the evidence is clear that the mean of the differenced series is significantly different from zero. Equation (3.2.3) on page 28 for the variance of the sample mean will help assess this significance.

However, as we have seen in the AR(1) and MA(1) cases, the one-step-ahead forecast error is

$$e_t(1) = Y_{t+1} - \hat{Y}_t(1) = e_{t+1}$$

Also

$$\begin{aligned} e_t(\ell) &= Y_{t+\ell} - \hat{Y}_t(\ell) \\ &= (Y_t + \ell\theta_0 + e_{t+1} + \cdots + e_{t+\ell}) - (Y_t + \ell\theta_0) \\ &= e_{t+1} + e_{t+2} + \cdots + e_{t+\ell} \end{aligned}$$

which agrees with Equation (9.3.14) on page 196 since in this model $\psi_j = 1$ for all j. (See Equation (5.2.6) on page 93 with $\theta = 0$.)

So, as in Equation (9.3.15), we have

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2 = \ell\sigma_e^2 \tag{9.3.27}$$

In contrast to the stationary case, here $Var(e_t(\ell))$ grows without limit as the forecast lead time ℓ increases. We shall see that this property is characteristic of the forecast error variance for all *nonstationary* ARIMA processes.

ARMA(*p*,*q*)

For the general stationary ARMA(p,q) model, the difference equation form for computing forecasts is given by

$$\begin{aligned} \hat{Y}_t(\ell) = {} & \phi_1 \hat{Y}_t(\ell-1) + \phi_2 \hat{Y}_t(\ell-2) + \cdots + \phi_p \hat{Y}_t(\ell-p) + \theta_0 \\ & - \theta_1 E(e_{t+\ell-1} | Y_1, Y_2, \ldots, Y_t) - \theta_2 E(e_{t+\ell-2} | Y_1, Y_2, \ldots, Y_t) \\ & - \cdots - \theta_q E(e_{t+\ell-q} | Y_1, Y_2, \ldots, Y_t) \end{aligned} \tag{9.3.28}$$

where

$$E(e_{t+j} | Y_1, Y_2, \ldots, Y_t) = \begin{cases} 0 & \text{for } j > 0 \\ e_{t+j} & \text{for } j \le 0 \end{cases} \tag{9.3.29}$$

We note that $\hat{Y}_t(j)$ is a true forecast for $j > 0$, but for $j \le 0$, $\hat{Y}_t(j) = Y_{t+j}$. As in Equation (9.3.20) on page 197, Equation (9.3.29) involves some minor approximation. For an invertible model, Equation (4.5.5) on page 80 shows that, using the π-weights, e_t can be expressed as a linear combination of the infinite sequence Y_t, Y_{t-1}, Y_{t-2},.... However, the π-weights die out exponentially fast, and the approximation assumes that π_j is negligible for $j > t - q$.

As an example, consider an ARMA(1,1) model. We have

$$\hat{Y}_t(1) = \phi Y_t + \theta_0 - \theta e_t \tag{9.3.30}$$

with

$$\hat{Y}_t(2) = \phi \hat{Y}_t(1) + \theta_0$$

and, more generally,

$$\hat{Y}_t(\ell) = \phi\hat{Y}_t(\ell-1) + \theta_0 \quad \text{for } \ell \geq 2 \tag{9.3.31}$$

using Equation (9.3.30) to get the recursion started.

Equations (9.3.30) and (9.3.31) can be rewritten in terms of the process mean and then solved by iteration to get the alternative explicit expression

$$\hat{Y}_t(\ell) = \mu + \phi^{\ell}(Y_t - \mu) - \phi^{\ell-1}e_t \quad \text{for } \ell \geq 1 \tag{9.3.32}$$

As Equations (9.3.28) and (9.3.29) indicate, the noise terms $e_{t-(q-1)},\ldots, e_{t-1}, e_t$ appear directly in the computation of the forecasts for leads $\ell = 1, 2,\ldots, q$. However, for $\ell > q$, the autoregressive portion of the difference equation takes over, and we have

$$\hat{Y}_t(\ell) = \phi_1\hat{Y}_t(\ell-1) + \phi_2\hat{Y}_t(\ell-2) + \cdots + \phi_p\hat{Y}_t(\ell-p) + \theta_0 \quad \text{for } \ell > q \tag{9.3.33}$$

Thus the general nature of the forecast for long lead times will be determined by the autoregressive parameters $\phi_1, \phi_2,\ldots, \phi_p$ (and the constant term, θ_0, which is related to the mean of the process).

Recalling from Equation (5.3.17) on page 97 that $\theta_0 = \mu(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$, we can rewrite Equation (9.3.33) in terms of deviations from μ as

$$\begin{aligned}\hat{Y}_t(\ell) - \mu = {} & \phi_1[\hat{Y}_t(\ell-1) - \mu] + \phi_2[\hat{Y}_t(\ell-2) - \mu] + \cdots \\ & + \phi_p[\hat{Y}_t(\ell-p) - \mu] \quad \text{for } \ell > q\end{aligned} \tag{9.3.34}$$

As a function of lead time ℓ, $\hat{Y}_t(\ell) - \mu$ follows the same Yule-Walker recursion as the autocorrelation function ρ_k of the process (see Equation (4.4.8), page 79). Thus, as in Section 4.3 on page 66 and Section 4.4 on page 77, the roots of the characteristic equation will determine the general behavior of $\hat{Y}_t(\ell) - \mu$ for large lead times. In particular, $\hat{Y}_t(\ell) - \mu$ can be expressed as a linear combination of exponentially decaying terms in ℓ (corresponding to the real roots) and damped sine wave terms (corresponding to the pairs of complex roots).

Thus, for any stationary ARMA model, $\hat{Y}_t(\ell) - \mu$ decays to zero as ℓ increases, and the long-term forecast is simply the process mean μ as given in Equation (9.3.9) on page 195. This agrees with common sense since for stationary ARMA models the dependence dies out as the time span between observations increases, and this dependence is the only reason we can improve on the "naive" forecast of using μ alone.

To argue the validity of Equation (9.3.15) for $e_t(\ell)$ in the present generality, we need to consider a new representation for ARIMA processes. Appendix G shows that any ARIMA model can be written in **truncated linear process** form as

$$Y_{t+\ell} = C_t(\ell) + I_t(\ell) \quad \text{for } \ell > 1 \tag{9.3.35}$$

where, for our present purposes, we need only know that $C_t(\ell)$ is a certain function of Y_t, $Y_{t-1},\ldots$ and

$$I_t(\ell) = e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1}e_{t+1} \quad \text{for } \ell \geq 1 \tag{9.3.36}$$

Furthermore, for invertible models with t reasonably large, $C_t(\ell)$ is a certain function of the finite history $Y_t, Y_{t-1},\ldots, Y_1$. Thus we have

$$\begin{aligned}\hat{Y}_t(\ell) &= E(C_t(\ell)|Y_1, Y_2, \ldots, Y_t) + E(I_t(\ell)|Y_1, Y_2, \ldots, Y_t) \\ &= C_t(\ell)\end{aligned}$$

Finally,

$$\begin{aligned}e_t(\ell) &= Y_{t+\ell} - \hat{Y}_t(\ell) \\ &= [C_t(\ell) + I_t(\ell)] - C_t(\ell) \\ &= I_t(\ell) \\ &= e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1} e_{t+1}\end{aligned}$$

Thus, for a general invertible ARIMA process,

$$E[e_t(\ell)] = 0 \quad \text{for } \ell \geq 1 \tag{9.3.37}$$

and

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2 \quad \text{for } \ell \geq 1 \tag{9.3.38}$$

From Equations (4.1.4) and (9.3.38), we see that for long lead times in stationary ARMA models, we have

$$Var(e_t(\ell)) \approx \sigma_e^2 \sum_{j=0}^{\infty} \psi_j^2$$

or

$$Var(e_t(\ell)) \approx \gamma_0 \quad \text{for large } \ell \tag{9.3.39}$$

Nonstationary Models

As the random walk shows, forecasting for nonstationary ARIMA models is quite similar to forecasting for stationary ARMA models, but there are some striking differences. Recall from Equation (5.2.2) on page 92 that an ARIMA(p,1,q) model can be written as a nonstationary ARMA(p+1,q) model, We shall write this as

$$\begin{aligned}Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \varphi_3 Y_{t-3} + \cdots + \varphi_p Y_{t-p} + \varphi_{p+1} Y_{t-p-1} \\ + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}\end{aligned} \tag{9.3.40}$$

where the script coefficients φ are directly related to the block ϕ coefficients. In particular,

$$\left.\begin{aligned} &\varphi_1 = 1 + \phi_1, \varphi_j = \phi_j - \phi_{j-1} \text{ for } j = 1, 2, \ldots, p \\ &\text{and} \\ &\varphi_{p+1} = -\phi_p \end{aligned}\right\} \tag{9.3.41}$$

For a general order of differencing d, we would have $p + d$ of the φ coefficients.

From this representation, we can immediately extend Equations (9.3.28), (9.3.29), and (9.3.30) on page 199 to cover the nonstationary cases by replacing p by $p + d$ and ϕ_j by φ_j.

As an example of the necessary calculations, consider the ARIMA(1,1,1) case. Here

$$Y_t - Y_{t-1} = \phi(Y_{t-1} - Y_{t-2}) + \theta_0 + e_t - \theta e_{t-1}$$

so that

$$Y_t = (1 + \phi)Y_{t-1} - \phi Y_{t-2} + \theta_0 + e_t - \theta e_{t-1}$$

Thus

$$\left.\begin{aligned} &\hat{Y}_t(1) = (1 + \phi)Y_t - \phi Y_{t-1} + \theta_0 - \theta e_t \\ &\hat{Y}_t(2) = (1 + \phi)\hat{Y}_t(1) - \phi Y_t + \theta_0 \\ &\text{and} \\ &\hat{Y}_t(\ell) = (1 + \phi)\hat{Y}_t(\ell - 1) - \phi \hat{Y}_t(\ell - 2) + \theta_0 \end{aligned}\right\} \tag{9.3.42}$$

For the general invertible ARIMA model, the truncated linear process representation given in Equations (9.3.35) and (9.3.36) and the calculations following these equations show that we can write

$$e_t(\ell) = e_{t+\ell} + \psi_1 e_{t+\ell-1} + \psi_2 e_{t+\ell-2} + \cdots + \psi_{\ell-1} e_{t+1} \text{ for } \ell \geq 1 \tag{9.3.43}$$

and so

$$E(e_t(\ell)) = 0 \text{ for } \ell \geq 1 \tag{9.3.44}$$

and

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2 \text{ for } \ell \geq 1 \tag{9.3.45}$$

However, for nonstationary series, the ψ_j-weights do not decay to zero as j increases. For example, for the random walk model, $\psi_j = 1$ for all j; for the IMA(1,1) model, $\psi_j = 1 - \theta$ for $j \geq 1$; for the IMA(2,2) case, $\psi_j = 1 + \theta_2 + (1 - \theta_1 - \theta_2)j$ for $j \geq 1$; and for the ARI(1,1) model, $\psi_j = (1 - \phi^{j+1})/(1 - \phi)$ for $j \geq 1$ (see Chapter 5).

Thus, for any nonstationary model, Equation (9.3.45) shows that the forecast error variance will grow without bound as the lead time ℓ increases. This fact should not be too surprising since with nonstationary series the distant future is quite uncertain.

9.4 Prediction Limits

As in all statistical endeavors, in addition to forecasting or predicting the unknown $Y_{t+\ell}$, we would like to assess the precision of our predictions.

Deterministic Trends

For the deterministic trend model with a white noise stochastic component $\{X_t\}$, we recall that

$$\hat{Y}_t(\ell) = \mu_{t+\ell}$$

and

$$Var(e_t(\ell)) = Var(X_{t+\ell}) = \gamma_0$$

If the stochastic component is normally distributed, then the forecast error

$$e_t(\ell) = Y_{t+\ell} - \hat{Y}_t(\ell) = X_{t+\ell} \tag{9.4.1}$$

is also normally distributed. Thus, for a given confidence level $1 - \alpha$, we could use a standard normal percentile, $z_{1-\alpha/2}$, to claim that

$$P\left[-z_{1-\alpha/2} < \frac{Y_{t+\ell} - \hat{Y}_t(\ell)}{\sqrt{Var(e_t(\ell))}} < z_{1-\alpha/2}\right] = 1 - \alpha$$

or, equivalently,

$$P[\hat{Y}_t(\ell) - z_{1-\alpha/2}\sqrt{Var(e_t(\ell))} < Y_{t+\ell} < \hat{Y}_t(\ell) + z_{1-\alpha/2}\sqrt{Var(e_t(\ell))}] = 1 - \alpha$$

Thus we may be $(1 - \alpha)100\%$ confident that the future observation $Y_{t+\ell}$ will be contained within the prediction limits

$$\hat{Y}_t(\ell) \pm z_{1-\alpha/2}\sqrt{Var(e_t(\ell))} \tag{9.4.2}$$

As a numerical example, consider the monthly average temperature series once more. On page 192, we used the cosine model to predict the June 1976 average temperature as 68.3°F. The estimate of $\sqrt{Var(e_t(\ell))} = \sqrt{\gamma_0}$ for this model is 3.7°F. Thus 95% prediction limits for the average June 1976 temperature are

$$68.3 \pm 1.96(3.7) = 68.3 \pm 7.252 \text{ or } 61.05°\text{F to } 75.55°\text{F}$$

Readers who are familiar with standard regression analysis will recall that since the forecast involves *estimated* regression parameters, the correct forecast error variance is given by $\gamma_0[1 + (1/n) + c_{n,\ell}]$, where $c_{n,\ell}$ is a certain function of the sample size n and the lead time ℓ. However, it may be shown that for the types of trends that we are considering (namely, cosines and polynomials in time) and for large sample sizes n, the $1/n$ and $c_{n,\ell}$ are both negligible relative to 1. For example, with a cosine trend of period 12 over $N = n/12$ years, we have that $c_{n,\ell} = 2/n$; thus the correct forecast error variance is

$\gamma_0[1 + (3/n)]$ rather than our approximate γ_0. For the linear time trend model, it can be shown that $c_{n,\ell} = 3(n + 2\ell - 1)^2/[n(n^2 - 1)] \approx 3/n$ for moderate lead ℓ and large n. Thus, again our approximation seems justified.

ARIMA Models

If the white noise terms $\{e_t\}$ in a general ARIMA series each arise independently from a normal distribution, then from Equation (9.3.43) on page 202, the forecast error $e_t(\ell)$ will also have a normal distribution, and the steps leading to Equation (9.4.2) remain valid. However, in contrast to the deterministic trend model, recall that in the present case

$$Var(e_t(\ell)) = \sigma_e^2 \sum_{j=0}^{\ell-1} \psi_j^2$$

In practice, σ_e^2 will be unknown and must be estimated from the observed time series. The necessary ψ-weights are, of course, also unknown since they are certain functions of the unknown ϕ's and θ's. For large sample sizes, these estimations will have little effect on the actual prediction limits given above.

As a numerical example, consider the AR(1) model that we estimated for the industrial color property series. From Exhibit 9.1 on page 194, we use $\phi = 0.5705$, $\mu = 74.3293$, and $\sigma_e^2 = 24.8$. For an AR(1) model, we recall Equation (9.3.16) on page 196

$$Var(e_t(\ell)) = \sigma_e^2 \left[\frac{1 - \phi^{2\ell}}{1 - \phi^2}\right]$$

For a one-step-ahead prediction, we have

$$70.14793 \pm 1.96\sqrt{24.8} = 70.14793 \pm 9.760721 \text{ or } 60.39 \text{ to } 79.91$$

Two steps ahead, we obtain

$$71.86072 \pm 11.88343 \text{ or } 60.71 \text{ to } 83.18$$

Notice that this prediction interval is wider than the previous interval. Forecasting ten steps ahead leads to

$$74.173934 \pm 11.88451 \text{ or } 62.42 \text{ to } 86.19$$

By lead 10, both the forecast and the forecast limits have settled down to their long-lead values.

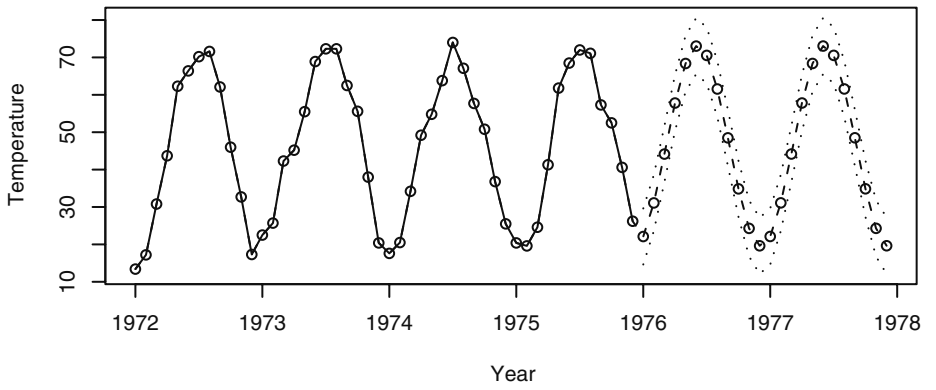
9.5 Forecasting Illustrations

Rather than showing forecast and forecast limit calculations, it is often more instructive to display appropriate plots of the forecasts and their limits.

Deterministic Trends

Exhibit 9.2 displays the last four years of the average monthly temperature time series together with forecasts and 95% forecast limits for two additional years. Since the model fits quite well with a relatively small error variance, the forecast limits are quite close to the fitted trend forecast.

Exhibit 9.2 Forecasts and Limits for the Temperature Cosine Trend

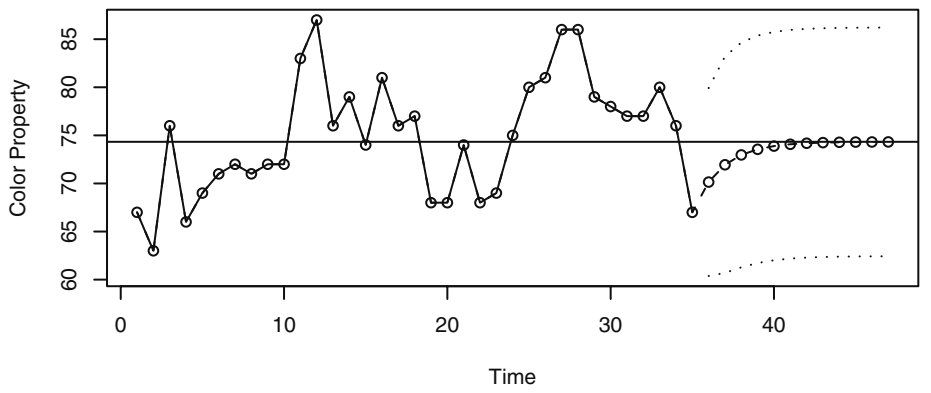



```
> data(tempdub)
> tempdub1=ts(c(tempdub,rep(NA,24)),start=start(tempdub),
    freq=frequency(tempdub))
> har.=harmonic(tempdub,1)
> m5.tempdub=arima(tempdub,order=c(0,0,0),xreg=har.)
> newhar.=harmonic(ts(rep(1,24), start=c(1976,1),freq=12),1)
> win.graph(width=4.875, height=2.5,pointsize=8)
> plot(m5.tempdub,n.ahead=24,n1=c(1972,1),newxreg=newhar.,
    type='b',ylab='Temperature',xlab='Year')
```

ARIMA Models

We use the industrial color property series as our first illustration of ARIMA forecasting. Exhibit 9.3 displays this series together with forecasts out to lead time 12 with the upper and lower 95% prediction limits for those forecasts. In addition, a horizontal line at the estimate for the process mean is shown. Notice how the forecasts approach the mean exponentially as the lead time increases. Also note how the prediction limits increase in width.

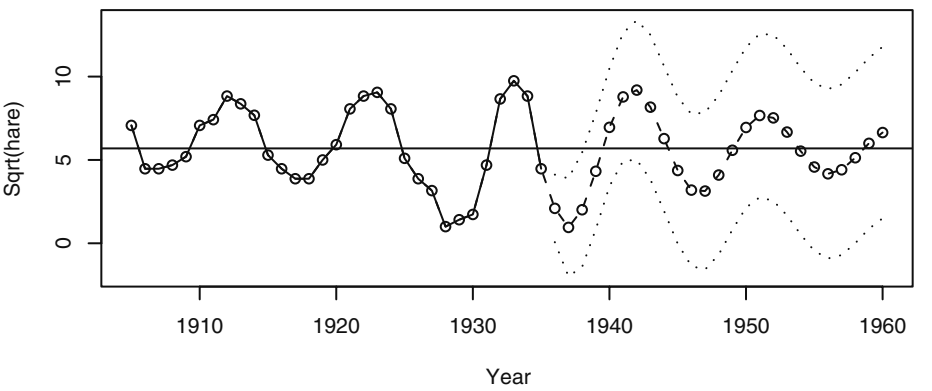
Exhibit 9.3 Forecasts and Forecast Limits for the AR(1) Model for Color




```
> data(color)
> m1.color=arima(color,order=c(1,0,0))
> plot(m1.color,n.ahead=12,type='b',xlab='Time',
    ylab='Color Property')
> abline(h=coef(m1.color)[names(coef(m1.color))=='intercept'])
```

The Canadian hare abundance series was fitted by working with the square root of the abundance numbers and then fitting an AR(3) model. Notice how the forecasts mimic the approximate cycle in the actual series even when we forecast with a lead time out to 25 years in Exhibit 9.4.

Exhibit 9.4 Forecasts from an AR(3) Model for Sqrt(Hare)




```
> data(hare)
> m1.hare=arima(sqrt(hare),order=c(3,0,0))
> plot(m1.hare, n.ahead=25,type='b',
    xlab='Year',ylab='Sqrt(hare)')
> abline(h=coef(m1.hare)[names(coef(m1.hare))=='intercept'])
```

9.6 Updating ARIMA Forecasts

Suppose we are forecasting a monthly time series. Our last observation is, say, for February, and we forecast for March, April, and May. As time goes by, the actual value for March becomes available. With this new value in hand, we would like to update or revise (and, one hopes, improve) our forecasts for April and May. Of course, we could compute new forecasts from scratch. However, there is a simpler way.

For a general forecast origin t and lead time $\ell + 1$, our original forecast is denoted $\hat{Y}_t(\ell + 1)$. Once the observation at time $t + 1$ becomes available, we would like to update our forecast as $\hat{Y}_{t+1}(\ell)$. Equations (9.3.35) and (9.3.36) on page 200 yield

$$Y_{t+\ell+1} = C_t(\ell + 1) + e_{t+\ell+1} + \psi_1 e_{t+\ell} + \psi_2 e_{t+\ell-1} + \cdots + \psi_\ell e_{t+1}$$

Since $C_t(\ell+1)$ and e_{t+1} are functions of $Y_{t+1}, Y_t, \ldots$, whereas $e_{t+\ell+1}, e_{t+\ell}, \ldots, e_{t+2}$ are independent of $Y_{t+1}, Y_t, \ldots$, we quickly obtain the expression

$$\hat{Y}_{t+1}(\ell) = C_t(\ell + 1) + \psi_\ell e_{t+1}$$

However, $\hat{Y}_t(\ell + 1) = C_t(\ell + 1)$, and, of course, $e_{t+1} = Y_{t+1} - \hat{Y}_t(1)$. Thus we have the general **updating equation**

$$\hat{Y}_{t+1}(\ell) = \hat{Y}_t(\ell + 1) + \psi_\ell[Y_{t+1} - \hat{Y}_t(1)] \tag{9.6.1}$$

Notice that $[Y_{t+1} - \hat{Y}_t(1)]$ is the actual forecast error at time $t + 1$ once Y_{t+1} has been observed.

As a numerical example, consider the industrial color property time series. Following Exhibit 9.1 on page 194, we fit an AR(1) model to forecast one step ahead as $\hat{Y}_{35}(1) = 70.096$ and two steps ahead as $\hat{Y}_{35}(2) = 71.86072$. If now the next color value becomes available as $Y_{t+1} = Y_{36} = 65$, then we update the forecast for time $t = 37$ as

$$\hat{Y}_{t+1}(1) = \hat{Y}_{36}(1) = 71.86072 + 0.5705(65 - 70.096) = 68.953452$$

9.7 Forecast Weights and Exponentially Weighted Moving Averages

For ARIMA models without moving average terms, it is clear how the forecasts are explicitly determined from the observed series $Y_t, Y_{t-1}, \ldots, Y_1$. However, for any model with $q > 0$, the noise terms appear in the forecasts, and the nature of the forecasts explicitly in terms of $Y_t, Y_{t-1}, \ldots, Y_1$ is hidden. To bring out this aspect of the forecasts, we return to the inverted form of any invertible ARIMA process, namely

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + e_t$$

(See Equation (4.5.5) on page 80.) Thus we can also write

$$Y_{t+1} = \pi_1 Y_t + \pi_2 Y_{t-1} + \pi_3 Y_{t-2} + \cdots + e_{t+1}$$

Taking conditional expectations of both sides, given Y_t, Y_{t-1}, …, Y_1, we obtain

$$\hat{Y}_t(1) = \pi_1 Y_t + \pi_2 Y_{t-1} + \pi_3 Y_{t-2} + \cdots \tag{9.7.1}$$

(We are assuming the t is sufficiently large and/or that the π-weights die out sufficiently quickly so that π_t, π_{t+1},… are all negligible.)

For any invertible ARIMA model, the π-weights can be calculated recursively from the expressions

$$\pi_j = \begin{cases} \sum_{i=1}^{min(j,q)} \theta_i \pi_{j-i} + \varphi_j \quad \text{for } 1 \le j \le p+d \\ \sum_{i=1}^{min(j,q)} \theta_i \pi_{j-i} \quad \text{for } j > p+d \end{cases} \tag{9.7.2}$$

with initial value $\pi_0 = -1$. (Compare this with Equations (4.4.7) on page 79 for the ψ-weights.)

Consider in particular the nonstationary IMA(1,1) model

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1}$$

Here $p = 0$, $d = 1$, $q = 1$, with $\varphi_1 = 1$; thus

$$\pi_1 = \theta\pi_0 + 1 = 1 - \theta$$

$$\pi_2 = \theta\pi_1 = \theta(1-\theta)$$

and, generally,

$$\pi_j = \theta\pi_{j-1} \quad \text{for } j > 1$$

Thus we have explicitly

$$\pi_j = (1-\theta)\theta^{j-1} \quad \text{for } j \ge 1 \tag{9.7.3}$$

so that, from Equation (9.7.1), we can write

$$\hat{Y}_t(1) = (1-\theta)Y_t + (1-\theta)\theta Y_{t-1} + (1-\theta)\theta^2 Y_{t-2} + \cdots \tag{9.7.4}$$

In this case, the π-weights *decrease exponentially*, and furthermore,

$$\sum_{j=1}^{\infty} \pi_j = (1-\theta)\sum_{j=1}^{\infty} \theta^{j-1} = \frac{1-\theta}{1-\theta} = 1$$

Thus $\hat{Y}_t(1)$ is called an **exponentially weighted moving average** (**EWMA**).

Simple algebra shows that we can also write

$$\hat{Y}_t(1) = (1-\theta)Y_t + \theta\hat{Y}_{t-1}(1) \tag{9.7.5}$$

and

$$\hat{Y}_t(1) = \hat{Y}_{t-1}(1) + (1-\theta)[Y_t - \hat{Y}_{t-1}(1)] \tag{9.7.6}$$

Equations (9.7.5) and (9.7.6) show how to update forecasts from origin $t-1$ to origin t, and they express the result as a linear combination of the new observation and the old forecast or in terms of the old forecast and the last observed forecast error.

Using EWMA to forecast time series has been advocated, mostly on an ad hoc basis, for a number of years; see Brown (1962) and Montgomery and Johnson (1976).

The parameter $1-\theta$ is called the **smoothing constant** in EWMA literature, and its selection (estimation) is often quite arbitrary. From the ARIMA model-building approach, we let the data indicate whether an IMA(1,1) model is appropriate for the series under consideration. If so, we then estimate θ in an efficient manner and compute an EWMA forecast that we are confident is the minimum mean square error forecast. A comprehensive treatment of exponential smoothing methods and their relationships with ARIMA models is given in Abraham and Ledolter (1983).

9.8 Forecasting Transformed Series

Differencing

Suppose we are interested in forecasting a series whose model involves a first difference to achieve stationarity. Two methods of forecasting can be considered:

1. forecasting the original nonstationary series, for example by using the difference equation form of Equation (9.3.28) on page 199, with ϕ's replaced by φ's throughout, or
2. forecasting the stationary differenced series $W_t = Y_t - Y_{t-1}$ and then "undoing" the difference by summing to obtain the forecast in original terms.

We shall show that both methods lead to the same forecasts. This follows essentially because differencing is a *linear* operation and because conditional expectation of a linear combination is the same linear combination of the conditional expectations.

Consider in particular the IMA(1,1) model. Basing our work on the original nonstationary series, we forecast as

$$\hat{Y}_t(1) = Y_t - \theta e_t \tag{9.8.1}$$

and

$$\hat{Y}_t(\ell) = \hat{Y}_t(\ell-1) \quad \text{for } \ell > 1 \tag{9.8.2}$$

Consider now the differenced stationary MA(1) series $W_t = Y_t - Y_{t-1}$. We would forecast $W_{t+\ell}$ as

$$\hat{W}_t(1) = -\theta e_t \tag{9.8.3}$$

and

$$\hat{W}_t(\ell) = 0 \quad \text{for } \ell > 1 \tag{9.8.4}$$

However, $\hat{W}_t(1) = \hat{Y}_t(1) - Y_t$; thus $\hat{W}_t(1) = -\theta e_t$ is equivalent to $\hat{Y}_t(1) = Y_t - \theta e_t$ as before. Similarly, $\hat{W}_t(\ell) = \hat{Y}_t(\ell) - \hat{Y}_t(\ell - 1)$, and Equation (9.8.4) becomes Equation (9.8.2), as we have claimed.

The same result would apply to any model involving differences of any order and indeed to any type of *linear* transformation with constant coefficients. (Certain linear transformations other than differencing may be applicable to seasonal time series. See Chapter 10.)

Log Transformations

As we saw earlier, it is frequently appropriate to model the logarithms of the original series—a nonlinear transformation. Let Y_t denote the original series value and let $Z_t = \log(Y_t)$. It can be shown that we always have

$$E(Y_{t+\ell} | Y_t, Y_{t-1}, \ldots, Y_1) \geq \exp[E(Z_{t+\ell} | Z_t, Z_{t-1}, \ldots, Z_1)] \tag{9.8.5}$$

with equality holding only in trivial cases. Thus, the naive forecast $\exp[\hat{Z}_t(\ell)]$ is *not* the minimum mean square error forecast of $Y_{t+\ell}$. To evaluate the minimum mean square error forecast in original terms, we shall find the following fact useful: If X has a normal distribution with mean μ and variance σ^2, then

$$E[\exp(X)] = \exp\left[\mu + \frac{\sigma^2}{2}\right]$$

(This follows, for example, from the moment-generating function for X.) In our application

$$\mu = E(Z_{t+\ell} | Z_t, Z_{t-1}, \ldots, Z_1)$$

and

$$\begin{aligned}
\sigma^2 &= Var(Z_{t+\ell} | Z_t, Z_{t-1}, \ldots, Z_1) \\
&= Var[e_t(\ell) + C_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] \\
&= Var[e_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] + Var[C_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] \\
&= Var[e_t(\ell) | Z_t, Z_{t-1}, \ldots, Z_1] \\
&= Var[e_t(\ell)]
\end{aligned}$$

These follow from Equations (9.3.35) and (9.3.36) (applied to Z_t) and the fact that $C_t(l)$ is a function of $Z_t, Z_{t-1}, \ldots$, whereas $e_t(\ell)$ is independent of $Z_t, Z_{t-1}, \ldots$. Thus the minimum mean square error forecast in the original series is given by

$$\exp\left\{\hat{Z}_t(\ell) + \frac{1}{2} Var[e_t(\ell)]\right\} \tag{9.8.6}$$

Throughout our discussion of forecasting, we have assumed that minimum mean square forecast error is the criterion of choice. For normally distributed variables, this is an

excellent criterion. However, if Z_t has a normal distribution, then $Y_t = \exp(Z_t)$ has a log-normal distribution, for which a different criterion may be desirable. In particular, since the log-normal distribution is asymmetric and has a long right tail, a criterion based on the mean absolute error may be more appropriate. For this criterion, the optimal forecast is the **median** of the distribution of $Z_{t+\ell}$ conditional on $Z_t, Z_{t-1},\ldots, Z_1$. Since the log transformation preserves medians and since, for a normal distribution, the mean and median are identical, the naive forecast $\exp[\hat{Z}_t(\ell)]$ is the optimal forecast for $Y_{t+\ell}$ in the sense that it minimizes the mean absolute forecast error.

9.9 Summary of Forecasting with Certain ARIMA Models

Here we bring together various forecasting results for special ARIMA models.

AR(1): $Y_t = \mu + \phi(Y_{t-1} - \mu) + e_t$

$$\hat{Y}_t(\ell) = \mu + \phi[\hat{Y}_t(\ell-1) - \mu] \quad \text{for } \ell \geq 1$$

$$= \mu + \phi^{\ell}(Y_t - \mu) \quad \text{for } l \geq 1$$

$$\hat{Y}_t(\ell) \approx \mu \quad \text{for large } \ell$$

$$e_t(\ell) = e_{t+\ell} + \phi e_{t+\ell-1} + \cdots + \phi^{\ell-1} e_{t+1}$$

$$Var(e_t(\ell)) = \sigma_e^2\left[\frac{1-\phi^{2\ell}}{1-\phi^2}\right]$$

$$Var(e_t(\ell)) \approx \frac{\sigma_e^2}{1-\phi^2} = \gamma_0 \quad \text{for large } \ell$$

$$\psi_j = \phi^j \quad \text{for } j > 0$$

MA(1): $Y_t = \mu + e_t - \theta e_{t-1}$

$$\hat{Y}_t(1) = \mu - \theta e_t$$

$$\hat{Y}_t(\ell) = \mu \quad \text{for } l > 1$$

$$e_t(1) = e_{t+1}$$

$$e_t(\ell) = e_{t+\ell} - \theta e_{t+\ell-1} \quad \text{for } \ell > 1$$

$$Var(e_t(\ell)) = \begin{cases} \sigma_e^2 & \text{for } \ell = 1 \\ \sigma_e^2(1+\theta^2) & \text{for } \ell > 1 \end{cases}$$

$$\psi_j = \begin{cases} -\theta & \text{for } j = 1 \\ 0 & \text{for } j > 1 \end{cases}$$

IMA (1,1) with Constant Term: $Y_t = Y_{t-1} + \theta_0 + e_t - \theta e_{t-1}$

$$\begin{aligned} \hat{Y}_t(\ell) &= \hat{Y}_t(\ell-1) + \theta_0 - \theta e_t \\ &= Y_t + \ell\theta_0 - \theta e_t \end{aligned}$$

$$\hat{Y}_t(1) = (1-\theta)Y_t + (1-\theta)\theta Y_{t-1} + (1-\theta)\theta^2 Y_{t-2} + \cdots (\text{the EWMA for } \theta_0 = 0)$$

$$e_t(\ell) = e_{t+\ell} + (1-\theta)e_{t+\ell-1} + (1-\theta)e_{t+\ell-2} + \cdots + (1-\theta)e_{t+1} \quad \text{for } \ell \geq 1$$

$$Var(e_t(\ell)) = \sigma_e^2[1 + (\ell-1)(1-\theta)^2]$$

$$\psi_j = 1 - \theta \quad \text{for } j > 0$$

Note that if $\theta_0 \neq 0$, the forecasts follow a straight line with slope θ_0, but if $\theta_0 = 0$, which is the usual case, then the forecast is the same for all lead times, namely

$$\hat{Y}_t(\ell) = Y_t - \theta e_t$$

IMA(2,2): $Y_t = 2Y_{t-1} - Y_{t-2} + \theta_0 + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$

$$\left.\begin{aligned} \hat{Y}_t(1) &= 2Y_t - Y_{t-1} + \theta_0 - \theta_1 e_t - \theta_2 e_{t-1} \\ \hat{Y}_t(2) &= 2\hat{Y}_t(1) - Y_t + \theta_0 - \theta_2 e_t \\ \hat{Y}_t(\ell) &= 2\hat{Y}_t(\ell-1) - \hat{Y}_t(\ell-2) + \theta_0 \quad \text{for } \ell > 2 \end{aligned}\right\} \tag{9.9.1}$$

$$\hat{Y}_t(\ell) = A + B\ell + \frac{\theta_0}{2}\ell^2 \tag{9.9.2}$$

where

$$A = 2\hat{Y}_t(1) - \hat{Y}_t(2) + \theta_0 \tag{9.9.3}$$

and

$$B = \hat{Y}_t(2) - \hat{Y}_t(1) - \frac{3}{2}\theta_0 \tag{9.9.4}$$

If $\theta_0 \neq 0$, the forecasts follow a quadratic curve in ℓ, but if $\theta_0 = 0$, the forecasts form a straight line with slope $\hat{Y}_t(2) - \hat{Y}_t(1)$ and will pass through the two initial forecasts $\hat{Y}_t(1)$ and $\hat{Y}_t(2)$. It can be shown that $Var(e_t(\ell))$ is a certain cubic function of ℓ; see Box, Jenkins, and Reinsel (1994, p. 156). We also have

$$\psi_j = 1 + \theta_2 + (1 - \theta_1 - \theta_2)j \quad \text{for } j > 0 \tag{9.9.5}$$

It can also be shown that forecasting the special case with $\theta_1 = 2\omega$ and $\theta_2 = -\omega^2$ is equivalent to so-called **double exponential smoothing** with smoothing constant $1 - \omega$; see Abraham and Ledolter (1983).

9.10 Summary

Forecasting or predicting future as yet unobserved values is one of the main reasons for developing time series models. Methods discussed in this chapter are all based on minimizing the mean square forecasting error. When the model is simply deterministic trend plus zero mean white noise error, forecasting amounts to extrapolating the trend. However, if the model contains autocorrelation, the forecasts exploit the correlation to produce better forecasts than would otherwise be obtained. We showed how to do this with ARIMA models and investigated the computation and properties of the forecasts. In special cases, the computation and properties of the forecasts are especially interesting and we presented them separately. Prediction limits are especially important to assess the potential accuracy (or otherwise) of the forecasts. Finally, we addressed the problem of forecasting time series for which the models involve transformation of the original series.

EXERCISES

9.1 For an AR(1) model with $Y_t = 12.2$, $\phi = -0.5$, and $\mu = 10.8$,
- **(a)** Find $\hat{Y}_t(1)$.
- **(b)** Calculate $\hat{Y}_t(2)$ in two different ways.
- **(c)** Calculate $\hat{Y}_t(10)$.

9.2 Suppose that annual sales (in millions of dollars) of the Acme Corporation follow the AR(2) model $Y_t = 5 + 1.1Y_{t-1} - 0.5Y_{t-2} + e_t$ with $\sigma_e^2 = 2$.
- **(a)** If sales for 2005, 2006, and 2007 were \$9 million, \$11 million, and \$10 million, respectively, forecast sales for 2008 and 2009.
- **(b)** Show that $\psi_1 = 1.1$ for this model.
- **(c)** Calculate 95% prediction limits for your forecast in part (a) for 2008.
- **(d)** If sales in 2008 turn out to be \$12 million, update your forecast for 2009.

9.3 Using the estimated cosine trend on page 192:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for April 1976.
- **(b)** Find a 95% prediction interval for that April forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.719°F.)
- **(c)** What is the forecast for April, 1977? For April 2009?

9.4 Using the estimated cosine trend on page 192:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for May 1976.
- **(b)** Find a 95% prediction interval for that May 1976 forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.719°F.)

9.5 Using the seasonal means model *without* an intercept shown in Exhibit 3.3 on page 32:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for April, 1976.
- **(b)** Find a 95% prediction interval for that April forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.419°F.)
- **(c)** Compare your forecast with the one obtained in Exercise 9.3.
- **(d)** What is the forecast for April 1977? April 2009?

9.6 Using the seasonal means model *with* an intercept shown in Exhibit 3.4 on page 33:
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for April 1976.
- **(b)** Find a 95% prediction interval for that April forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.419°F.)
- **(c)** Compare your forecast with the one obtained in Exercise 9.5.

9.7 Using the seasonal means model *with* an intercept shown in Exhibit 3.4 on page 33
- **(a)** Forecast the average monthly temperature in Dubuque, Iowa, for January 1976.
- **(b)** Find a 95% prediction interval for that January forecast. (The estimate of $\sqrt{\gamma_0}$ for this model is 3.419°F.)

9.8 Consider the monthly electricity generation time series shown in Exhibit 5.8 on page 99. The data are in the file named electricity.
- **(a)** Fit a deterministic trend model containing seasonal means together with a linear time trend to the logarithms of the electricity values.
- **(b)** Plot the last five years of the series together with two years of forecasts and the 95% forecast limits. Interpret the plot.

9.9 Simulate an AR(1) process with $\phi = 0.8$ and $\mu = 100$. Simulate 48 values but set aside the last 8 values to compare forecasts to actual values.
- **(a)** Using the first 40 values of the series, find the values for the maximum likelihood estimates of ϕ and μ.
- **(b)** Using the estimated model, forecast the next eight values of the series. Plot the series together with the eight forecasts. Place a horizontal line at the estimate of the process mean.
- **(c)** Compare the eight forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and the same sample size.

9.10 Simulate an AR(2) process with $\phi_1 = 1.5$, $\phi_2 = -0.75$, and $\mu = 100$. Simulate 52 values but set aside the last 12 values to compare forecasts to actual values.
- **(a)** Using the first 40 values of the series, find the values for the maximum likelihood estimates of the ϕ's and μ.
- **(b)** Using the estimated model, forecast the next 12 values of the series. Plot the series together with the 12 forecasts. Place a horizontal line at the estimate of

the process mean.

(c) Compare the 12 forecasts with the actual values that you set aside.

(d) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(e) Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.11 Simulate an MA(1) process with $\theta = 0.6$ and $\mu = 100$. Simulate 36 values but set aside the last 4 values to compare forecasts to actual values.

(a) Using the first 32 values of the series, find the values for the maximum likelihood estimates of the θ and μ.

(b) Using the estimated model, forecast the next four values of the series. Plot the series together with the four forecasts. Place a horizontal line at the estimate of the process mean.

(c) Compare the four forecasts with the actual values that you set aside.

(d) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(e) Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.12 Simulate an MA(2) process with $\theta_1 = 1$, $\theta_2 = -0.6$, and $\mu = 100$. Simulate 36 values but set aside the last 4 values with compare forecasts to actual values.

(a) Using the first 32 values of the series, find the values for the maximum likelihood estimates of the θ's and μ.

(b) Using the estimated model, forecast the next four values of the series. Plot the series together with the four forecasts. Place a horizontal line at the estimate of the process mean.

(c) What is special about the forecasts at lead times 3 and 4?

(d) Compare the four forecasts with the actual values that you set aside.

(e) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(f) Repeat parts (a) through (e) with a new simulated series using the same values of the parameters and same sample size.

9.13 Simulate an ARMA(1,1) process with $\phi = 0.7$, $\theta = -0.5$, and $\mu = 100$. Simulate 50 values but set aside the last 10 values to compare forecasts with actual values.

(a) Using the first 40 values of the series, find the values for the maximum likelihood estimates of ϕ, θ, and μ.

(b) Using the estimated model, forecast the next ten values of the series. Plot the series together with the ten forecasts. Place a horizontal line at the estimate of the process mean.

(c) Compare the ten forecasts with the actual values that you set aside.

(d) Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?

(e) Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.14 Simulate an IMA(1,1) process with $\theta = 0.8$ and $\theta_0 = 0$. Simulate 35 values, but set aside the last five values to compare forecasts with actual values.
- **(a)** Using the first 30 values of the series, find the value for the maximum likelihood estimate of θ.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about the forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.15 Simulate an IMA(1,1) process with $\theta = 0.8$ and $\theta_0 = 10$. Simulate 35 values, but set aside the last five values to compare forecasts to actual values.
- **(a)** Using the first 30 values of the series, find the values for the maximum likelihood estimates of θ and θ_0.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about these forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.16 Simulate an IMA(2,2) process with $\theta_1 = 1$, $\theta_2 = -0.75$, and $\theta_0 = 0$. Simulate 45 values, but set aside the last five values to compare forecasts with actual values.
- **(a)** Using the first 40 values of the series, find the value for the maximum likelihood estimate of θ_1 and θ_2.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about the forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.17 Simulate an IMA(2,2) process with $\theta_1 = 1$, $\theta_2 = -0.75$, and $\theta_0 = 10$. Simulate 45 values, but set aside the last five values to compare forecasts with actual values.
- **(a)** Using the first 40 values of the series, find the values for the maximum likelihood estimates of θ_1, θ_2, and θ_0.
- **(b)** Using the estimated model, forecast the next five values of the series. Plot the series together with the five forecasts. What is special about these forecasts?
- **(c)** Compare the five forecasts with the actual values that you set aside.
- **(d)** Plot the forecasts together with 95% forecast limits. Do the actual values fall within the forecast limits?
- **(e)** Repeat parts (a) through (d) with a new simulated series using the same values of the parameters and same sample size.

9.18 Consider the model $Y_t = \beta_0 + \beta_1 t + X_t$, where $X_t = \phi X_{t-1} + e_t$. We assume that β_0, β_1, and ϕ are known. Show that the minimum mean square error forecast ℓ steps ahead can be written as $\hat{Y}_t(\ell) = \beta_0 + \beta_1(t+\ell) + \phi^{\ell}(Y_t - \beta_0 - \beta_1 t)$.

9.19 Verify Equation (9.3.16) on page 196.

9.20 Verify Equation (9.3.32) on page 200.

9.21 The data file named deere3 contains 57 consecutive values from a complex machine tool process at Deere & Co. The values given are deviations from a target value in units of ten millionths of an inch. The process employs a control mechanism that resets some of the parameters of the machine tool depending on the magnitude of deviation from target of the last item produced.

(a) Using an AR(1) model for this series, forecast the next ten values.

(b) Plot the series, the forecasts, and 95% forecast limits, and interpret the results.

9.22 The data file named days contains accounting data from the Winegard Co. of Burlington, Iowa. The data are the number of days until Winegard receives payment for 130 consecutive orders from a particular distributor of Winegard products. (The name of the distributor must remain anonymous for confidentiality reasons.) The time series contains outliers that are quite obvious in the time series plot. Replace each of the unusual values at "times" 63, 106, and 129 with the much more typical value of 35 days.

(a) Use an MA(2) model to forecast the next ten values of this modified series.

(b) Plot the series, the forecasts, and 95% forecast limits, and interpret the results.

9.23 The time series in the data file robot gives the final position in the "*x*-direction" after an industrial robot has finished a planned set of exercises. The measurements are expressed as deviations from a target position. The robot is put through this planned set of exercises in the hope that its behavior is repeatable and thus predictable.

(a) Use an IMA(1,1) model to forecast five values ahead. Obtain 95% forecast limits also.

(b) Display the forecasts, forecast limits, and actual values in a graph and interpret the results.

(c) Now use an ARMA(1,1) model to forecast five values ahead and obtain 95% forecast limits. Compare these results with those obtained in part (a).

9.24 Exhibit 9.4 on page 206 displayed the forecasts and 95% forecast limits for the square root of the Canadian hare abundance. The data are in the file named hare. Produce a similar plot in original terms. That is, plot the original abundance values together with the squares of the forecasts and squares of the forecast limits.

9.25 Consider the seasonal means plus linear time trend model for the logarithms of the monthly electricity generation time series in Exercise 9.8. (The data are in the file named electricity.)

(a) Find the two-year forecasts and forecast limits in original terms. That is, exponentiate (antilog) the results obtained in Exercise 9.8.

(b) Plot the last five years of the original time series together with two years of forecasts and the 95% forecast limits, all in original terms. Interpret the plot.

Appendix E: Conditional Expectation

If X and Y have joint pdf $f(x,y)$ and we denote the marginal pdf of X by $f(x)$, then the **conditional pdf** of Y given $X = x$ is given by

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

For a given value of x, the conditional pdf has all of the usual properties of a pdf. In particular, the **conditional expectation** of Y given $X = x$ is defined as

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f(y|x) dy$$

As an expected value or mean, the **conditional expectation** of Y given $X = x$ has all of the usual properties. For example,

$$E(aY + bZ + c|X=x) = aE(Y|X=x) + bE(Z|X=x) + c \tag{9.E.1}$$

and

$$E[h(Y)|X = x] = \int_{-\infty}^{\infty} y f(y|x) dx \tag{9.E.2}$$

In addition, several new properties arise:

$$E[h(X)|X=x] = h(x) \tag{9.E.3}$$

That is, given $X = x$, the random variable $h(X)$ can be treated like a constant $h(x)$. More generally,

$$E[h(X, Y)|X=x] = E(h(x, Y)|X=x) \tag{9.E.4}$$

If we set $E(Y|X=x) = g(x)$, then $g(X)$ is a random variable and we can consider $E[g(X)]$. It can be shown that

$$E[g(X)] = E(Y)$$

which is often written as

$$E[E(Y|X)] = E(Y) \tag{9.E.5}$$

If Y and X are independent, then

$$E(Y|X) = E(Y) \tag{9.E.6}$$

Appendix F: Minimum Mean Square Error Prediction

Suppose Y is a random variable with mean μ_Y and variance σ_Y^2. If our object is to predict Y using only a constant c, what is the *best* choice for c? Clearly, we must first define *best*. A common (and convenient) criterion is to choose c to minimize the **mean square error of prediction**, that is, to minimize

$$g(c) = E[(Y - c)^2]$$

If we expand $g(c)$, we have

$$g(c) = E(Y^2) - 2cE(Y) + c^2$$

Since $g(c)$ is quadratic in c and opens upward, solving $g'(c) = 0$ will produce the required minimum. We have

$$g'(c) = -2E(Y) + 2c$$

so that the optimal c is

$$c = E(Y) = \mu \tag{9.F.1}$$

Note also that

$$\min_{-\infty < c < \infty} g(c) = E(Y-\mu)^2 = \sigma_Y^2 \tag{9.F.2}$$

Now consider the situation where a second random variable X is available and we wish to use the observed value of X to help predict Y. Let $\rho = Corr\,(X,Y)$. We first suppose, for simplicity, that only *linear* functions $a + bX$ can be used for the prediction. The mean square error is then given by

$$g(a, b) = E(Y - a - bX)^2$$

and expanding we gave

$$g(a, b) = E(Y^2) + a^2 + b^2E(X^2) - 2aE(Y) + 2abE(X) - 2bE(XY)$$

This is also quadratic in a and b and opens upward. Thus we can find the point of minimum by solving simultaneous linear equations $\partial g(a, b)/\partial a = 0$ and $\partial g(a, b)/\partial b = 0$. We have

$$\partial g(a, b)/\partial a = 2a - 2E(Y) + 2bE(X) = 0$$
$$\partial g(a, b)/\partial b = 2bE(X^2) + 2aE(X) - 2E(XY) = 0$$

which we rewrite as

$$a + E(X)b = E(Y)$$
$$E(X)a + E(X^2)b = EXY$$

Multiplying the first equation by $E(X)$ and subtracting yields

$$b = \frac{E(XY) - E(X)E(Y)}{E(X^2) - [E(X)]^2} = \frac{Cov(X, Y)}{Var(X)} = \rho\frac{\sigma_Y}{\sigma_X} \tag{9.F.3}$$

Then

$$a = E(Y) - bE(X) = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X \tag{9.F.4}$$

If we let $\hat{Y}$ be the minimum mean square error prediction of Y based on a linear function of X, then we can write

$$\hat{Y} = \left[\mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X\right] + \left[\rho\frac{\sigma_Y}{\sigma_X}\mu_X\right]X \tag{9.F.5}$$

or

$$\left[\frac{\hat{Y}-\mu_Y}{\sigma_Y}\right] = \rho\left[\frac{X-\mu_X}{\sigma_X}\right] \tag{9.F.6}$$

In terms of standardized variables $\hat{Y}^*$ and X^*, we have simply $\hat{Y}^* = \rho X^*$.

Also, using Equations (9.F.3) and (9.F.4), we find

$$min\ g(a,b) = \sigma_Y^2(1-\rho^2) \tag{9.F.7}$$

which provides a proof that $-1 \le \rho \le +1$ since $g(a,b) \ge 0$.

If we compare Equation (9.F.7) with Equation (9.F.2), we see that the minimum mean square error obtained when we use a linear function of X to predict Y is reduced by a factor of $1-\rho^2$ compared with that obtained by ignoring X and simply using the constant μ_Y for our prediction.

Let us now consider the more general problem of predicting Y with an arbitrary function of X. Once more our criterion will be to minimize the mean square error of prediction. We need to choose the function $h(X)$, say, that minimizes

$$E[Y-h(X)]^2 \tag{9.F.8}$$

Using Equation (9.E.5), we can write this as

$$E[Y-h(X)]^2 = E(E\{[Y-h(X)]^2 \big| X\}) \tag{9.F.9}$$

Using Equation (9.E.4), the inner expectation can be written as

$$E\{[Y-h(X)]^2 \big| X=x\} = E\{[Y-h(x)]^2 \big| X=x\} \tag{9.F.10}$$

For each value of x, $h(x)$ is a constant, and we can apply the result of Equation (9.F.1) to the conditional distribution of Y given $X=x$. Thus, for each x, the best choice of $h(x)$ is

$$h(x) = E(Y|X=x) \tag{9.F.11}$$

Since this choice of $h(x)$ minimizes the inner expectation in Equation (9.F.9), it must also provide the overall minimum of Equation (9.F.8). Thus

$$h(X) = E(Y|X) \tag{9.F.12}$$

is the best predictor of Y of *all* functions of X.

If X and Y have a bivariate normal distribution, it is well-known that

$$E(Y|X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)$$

so that the solutions given in Equations (9.F.12) and (9.F.5) coincide. In this case, the linear predictor is the best of all functions.

More generally, if Y is to be predicted by a function of $X_1, X_2, \ldots, X_n$, then it can be easily argued that the minimum square error predictor is given by

$$E(Y|X_1, X_2, \ldots, X_n) \tag{9.F.13}$$

Appendix G: The Truncated Linear Process

Suppose $\{Y_t\}$ satisfies the general ARIMA(p,d,q) model with AR characteristic polynomial $\phi(x)$, MA characteristic polynomial $\theta(x)$, and constant term θ_0. Then the **truncated linear process** representation for $\{Y_t\}$ is given by

$$Y_{t+l} = C_t(\ell) + I_t(\ell) \quad \text{for } \ell \geq 1 \tag{9.G.1}$$

where

$$I_t(\ell) = \sum_{j=0}^{\ell-1} \psi_j e_{t+\ell-j} \quad \text{for } \ell \geq 1 \tag{9.G.2}$$

$$C_t(\ell) = \sum_{i=0}^{d} A_i \ell^i + \sum_{i=1}^{r} \sum_{j=0}^{p_i-1} B_{ij} \ell^j (G_i)^\ell \tag{9.G.3}$$

and A_i, B_{ij}, $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, p_i$, are constant in ℓ and depend only on Y_t, $Y_{t-1}, \ldots$.[†] As always, the ψ-weights are defined by the identity

$$\phi(x)(1-x)^d(1 + \psi_1 x + \psi_2 x^2 + \cdots) = \theta(x) \tag{9.G.4}$$

or

$$\varphi(x)(1 + \psi_1 x + \psi_2 x^2 + \cdots) = \theta(x) \tag{9.G.5}$$

We shall show that the representation given by Equation (9.G.1) is valid by arguing that, for fixed t, $C_t(\ell)$ is essentially the **complementary function** of the defining difference equation, that is,

$$C_t(\ell) - \varphi_1 C_t(\ell-1) - \varphi_2 C_t(\ell-2) - \cdots - \varphi_{p+d} C_t(\ell-p-d) = \theta_0 \quad \text{for } \ell \geq 0 \tag{9.G.6}$$

and that $I_t(\ell)$ is a **particular solution** (without θ_0):

$$\begin{aligned} I_t(\ell) - \varphi_1 I_t(\ell-1) - \varphi_2 I_t(\ell-2) - \cdots - \varphi_{p+d} I_t(\ell-p-d) \\ = e_{t+\ell} - \theta_1 e_{t+l-1} - \theta_2 e_{t+l-2} - \cdots - \theta_q e_{t+l-q} \quad \text{for } \ell > q \end{aligned} \tag{9.G.7}$$

Since $C_t(\ell)$ contains $p + d$ arbitrary constants (the A's and the B's), summing $C_t(\ell)$ and $I_t(\ell)$ yields the general solution of the ARIMA equation. Specific values for the A's and B's will be determined by initial conditions on the $\{Y_t\}$ process.

We note that A_d is not arbitrary. We have

$$A_d = \frac{\theta_0}{(1 - \phi_1 - \phi_2 - \cdots - \phi_p)d!} \tag{9.G.8}$$

The proof that $C_t(\ell)$ as given by Equation (9.G.2) is the complementary function and satisfies Equation (9.G.6) is a standard result from the theory of difference equations

[†] The only property of the $C_t(\ell)$ that we need is that it depends only on $Y_t, Y_{t-1}, \ldots$.

(see, for example, Goldberg, 1958). We shall show that the particular solution $I_t(\ell)$ defined by Equation (9.G.2) does satisfy Equation (9.G.7).

For convenience of notation, we let $\varphi_j = 0$ for $j > p + d$. Consider the left-hand side of Equation (9.G.7). It can be written as:

$$\left.\begin{aligned}(\psi_0 e_{t+\ell} + \psi_1 e_{t+\ell-1} + \cdots + \psi_{\ell-1} e_{t+1}) - \varphi_1(\psi_0 e_{t+\ell-1} + \psi_1 e_{t+\ell-2} + \cdots \\ + \psi_{\ell-2} e_{t+1}) - \cdots - \varphi_{p+d}(\psi_0 e_{t+\ell-p-d} \\ + \psi_1 e_{t+\ell-p-d-1} + \cdots + \psi_{\ell-p-d-1} e_{t+1})\end{aligned}\right\} \quad (9.G.9)$$

Now grouping together common e_t terms and picking off their coefficients, we obtain

$$\begin{aligned}&\text{Coefficient of } e_{t+\ell-1} : \psi_0 \\ &\text{Coefficient of } e_{t+\ell-2} : \psi_1 - \varphi_1\psi_0 \\ &\text{Coefficient of } e_{t+\ell-3} : \psi_2 - \varphi_1\psi_1 - \varphi_2\psi_0 \\ &\qquad\qquad\vdots \\ &\text{Coefficient of } e_{t+1} \quad : \psi_{\ell-1} - \varphi_1\psi_{\ell-2} - \varphi_2\psi_{\ell-3} - \cdots - \varphi_{p+d}\psi_{\ell-p-d-1}\end{aligned}$$

If $\ell > q$, we can match these coefficients to the corresponding coefficients on the right-hand side of Equation (9.G.7) to obtain the relationships

$$\left.\begin{aligned}\psi_0 &= 1 \\ \psi_1 - \varphi_1\psi_0 &= -\theta_1 \\ \psi_2 - \varphi_1\psi_1 - \varphi_2\psi_0 &= -\theta_2 \\ &\vdots \\ \psi_q - \varphi_1\psi_{q-1} - \varphi_2\psi_{q-2} - \cdots - \varphi_q\psi_0 &= -\theta_q \\ \psi_{\ell-1} - \varphi_1\psi_{\ell-2} - \varphi_2\psi_{\ell-3} - \cdots - \varphi_{p+d}\psi_{\ell-p-d-1} &= 0 \quad \text{for } \ell > q\end{aligned}\right\} \quad (9.G.10)$$

However, by comparing these relationships with Equation (9.G.5), we see that Equations (9.G.10) are precisely the equations defining the ψ-weights and thus Equation (9.G.7) is established as required.

Appendix H: State Space Models

Control theory engineers have developed and successfully used so-called **state space models** and **Kalman filtering** since Kalman published his seminal work in 1960. Recent references include Durbin and Koopman (2001) and Harvey et al. (2004).

Consider a general stationary and invertible ARMA(p,q) process $\{Z_t\}$. Put $m = max(p, q + 1)$ and define the **state** of the process at time t as the column vector $\boldsymbol{Z}(t)$ of length m whose jth element is the forecast $\hat{Z}(j)$ for $j = 0, 1, 2,\ldots, m - 1$, based on Z_t, $Z_{t-1},\ldots$. Note that the lead element of $\boldsymbol{Z}(t)$ is just $\hat{Z}(0) = Z_t$.

Recall the updating Equation (9.6.1) on page 207, which in the present context can

be written

$$\hat{Z}_{t+1}(\ell) = \hat{Z}_t(\ell+1) + \psi_\ell e_{t+1} \tag{9.H.1}$$

We shall use this expression directly for $\ell = 0, 1, 2, \ldots, m-2$. For $\ell = m-1$, we have

$$\begin{aligned}\hat{Z}_{t+1}(m-1) &= \hat{Z}_t(m) + \psi_{m-1} e_{t+1} \\ &= \phi_1 \hat{Z}_t(m-1) + \phi_2 \hat{Z}_t(m-2) + \ldots + \phi_p \hat{Z}_t(m-p) + \psi_{m-1} e_{t+1}\end{aligned} \tag{9.H.2}$$

where the last expression comes from Equation (9.3.34) on page 200, with $\mu = 0$.

The matrix formulation of Equations (9.H.1) and (9.H.2) relating $\boldsymbol{Z}(t+1)$ to $\boldsymbol{Z}(t)$ and e_{t+1}, called the **equations of state** (or **Akaike's Markovian representation**), is given as

$$\boldsymbol{Z}(t+1) = \boldsymbol{F}\boldsymbol{Z}(t) + \boldsymbol{G}e_{t+1} \tag{9.H.3}$$

where

$$\boldsymbol{F} = \begin{bmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 \\ & & & & & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \\ \phi_m & \phi_{m-1} & \cdot & \cdot & \cdot & \phi_1 \end{bmatrix} \tag{9.H.4}$$

and

$$\boldsymbol{G} = \begin{bmatrix} 1 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_{m-1} \end{bmatrix} \tag{9.H.5}$$

with $\phi_j = 0$ for $j > p$. Note that the simplicity of Equation (9.H.3) is obtained at the expense of having to deal with vector-valued processes. Because the state space formulation also usually allows for measurement error, we do not observe Z_t directly but only observe Y_t through the **observational equation**

$$Y_t = \boldsymbol{H}\boldsymbol{Z}(t) + \varepsilon_t \tag{9.H.6}$$

where $\boldsymbol{H} = [1, 0, 0, \ldots, 0]$ and $\{\varepsilon_t\}$ is another zero-mean white noise process independent of $\{e_t\}$. The special case of *no* measurement error is obtained by setting $\varepsilon_t = 0$ in Equation (9.H.6). Equivalently, this case is obtained by taking $\sigma_\varepsilon^2 = 0$ in subsequent equations. More general state space models allow $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{H}$ to be more general, possibly also depending on time.

Evaluation of the Likelihood Function and Kalman Filtering

First a definition: The **covariance matrix** for a vector of random variables $\boldsymbol{X}$ of dimension $n \times 1$ is defined to be the $n \times n$ matrix whose ijth element is the covariance between the ith and jth components of $\boldsymbol{X}$.

If $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{B}$, then it is easily shown that the covariance matrix for $\boldsymbol{Y}$ is $\boldsymbol{AVA}^T$, where $\boldsymbol{V}$ is the covariance matrix for $\boldsymbol{X}$ and the superscript T denotes matrix transpose.

Getting back to the Kalman filter, we let $\boldsymbol{Z}(t+1|t)$ denote the $m \times 1$ vector whose jth component is $E[\hat{Z}_{t+1}(j)|Y_t, Y_{t-1}, \ldots, Y_1]$ for $j = 0, 1, 2, \ldots, m-1$. Similarly, let $\boldsymbol{Z}(t|t)$ be the vector whose jth component is $E[\hat{Z}_t(j)|Y_t, Y_{t-1}, \ldots, Y_1]$ for $j = 0, 1, 2, \ldots, m-1$.

Then, since e_{t+1} is independent of $Z_t, Z_{t-1}, \ldots$, and hence also of $Y_t, Y_{t-1}, \ldots$, we see from Equation (9.H.3) that

$$\boldsymbol{Z}(t+1|t) = \boldsymbol{F}\boldsymbol{Z}(t|t) \tag{9.H.7}$$

Also letting $\boldsymbol{P}(t+1|t)$ be the covariance matrix for the "forecast error" $\boldsymbol{Z}(t+1) - \boldsymbol{Z}(t+1|t)$ and $\boldsymbol{P}(t|t)$ be the covariance matrix for the "forecast error" $\boldsymbol{Z}(t) - \boldsymbol{Z}(t|t)$, we have from Equation (9.H.3) that

$$\boldsymbol{P}(t+1|t) = \boldsymbol{F}[\boldsymbol{P}(t|t)]\boldsymbol{F}^T + \sigma_e^2 \boldsymbol{G}\boldsymbol{G}^T \tag{9.H.8}$$

From the observational equation (Equation (9.H.6)) and then replacing $t+1$ by t,

$$Y(t+1|t) = \boldsymbol{H}\boldsymbol{Z}(t+1|t) \tag{9.H.9}$$

where $Y(t+1|t) = E(Y_{t+1}|Y_t, Y_{t-1}, \ldots, Y_1)$.

It can now be shown that the following relationships hold (see, for example, Harvey, 1981c):

$$\boldsymbol{Z}(t+1|t+1) = \boldsymbol{Z}(t+1|t) + \boldsymbol{K}(t+1)[Y_{t+1} - Y(t+1|t)] \tag{9.H.10}$$

where

$$\boldsymbol{K}(t+1) = \boldsymbol{P}(t+1|t)\boldsymbol{H}^T[\boldsymbol{H}\boldsymbol{P}(t+1|t)\boldsymbol{H}^T + \sigma_\varepsilon^2]^{-1} \tag{9.H.11}$$

and

$$\boldsymbol{P}(t+1|(t+1)) = \boldsymbol{P}(t+1|t) - \boldsymbol{K}(t+1)\boldsymbol{H}\boldsymbol{P}(t+1|t) \tag{9.H.12}$$

Collectively, Equations (9.H.10), (9.H.11), and (9.H.12) are referred to as the **Kalman filter equations**. The quantity

$$err_{t+1} = Y_{t+1} - Y(t+1|t) \tag{9.H.13}$$

in Equation (9.H.10) is the prediction error and is independent of (or at least uncorrelated with) the past observations $Y_t, Y_{t-1}, \ldots$. Since we are allowing for measurement error, err_{t+1} is not, in general, the same as e_{t+1}.

From Equations (9.H.13) and (9.H.6), we have

$$v_{t+1} = Var(err_{t+1}) = \boldsymbol{H}\boldsymbol{P}(t+1|t)\boldsymbol{H}^T + \sigma_\varepsilon^2 \tag{9.H.14}$$

Now consider the likelihood function for the observed series $Y_1, Y_2,\ldots, Y_n$. From the definition of the conditional probability density function, we can write

$$f(y_1, y_2, \ldots, y_n) = f(y_n|y_1, y_2, \ldots, y_{n-1})f(y_1, y_2, \ldots, y_{n-1})$$

or, by taking logs,

$$\log f(y_1, y_2, \ldots, y_n) = \log f(y_1, y_2, \ldots, y_{n-1}) + \log f(y_n|y_1, y_2, \ldots, y_{n-1}) \quad (9.H.15)$$

Assume now that we are dealing with normal distributions, that is, that $\{e_t\}$ and $\{\varepsilon_t\}$ are normal white noise processes. Then it is known that the distribution of Y_n conditional on $Y_1 = y_1, Y_2 = y_2,\ldots, Y_{n-1} = y_{n-1}$, is also normal with mean $y(n|n-1)$ and variance v_n. In the remainder of this section and the next, we write $y(n|n-1)$ for the observed value of $Y(n|n-1)$.The second term on the right-hand side of Equation (9.H.15) can then be written

$$\log f(y_n|y_1, y_2, \ldots, y_{n-1}) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log v_n - \frac{1}{2}\frac{[y_n - y(n|n-1)]^2}{v_n}$$

Furthermore, the first term on the right-hand side of Equation (9.H.15) can be decomposed similarly again and again until we have

$$\log f(y_1, y_2, \ldots, y_n) = \sum_{t=2}^{n} \log f(y_t|y_1, y_2, \ldots, y_{t-1}) + \log f(y_1) \quad (9.H.16)$$

which then becomes the prediction error decomposition of the likelihood, namely

$$\log f(y_1, y_2, \ldots, y_n) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{n} v_t - \frac{1}{2}\sum_{t=1}^{n} \frac{[y_t - y(t|t-1)]^2}{v_t} \quad (9.H.17)$$

with $y(1|0) = 0$ and $v_1 = Var(Y_1)$.

The overall strategy for computing the likelihood for a given set of parameter values is to use the Kalman filter equations to generate recursively the prediction errors and their variances and then use the prediction error decomposition of the likelihood function. Only one point remains: We need initial values $\boldsymbol{Z}(0|0)$ and $\boldsymbol{P}(0|0)$ to get the recursions started.

The Initial State Covariance Matrix

The initial state vector $\boldsymbol{Z}(0|0)$ will be a vector of zeros for a zero-mean process, and $\boldsymbol{P}(0|0)$ is the covariance matrix for $\boldsymbol{Z}(0) - \boldsymbol{Z}(0|0) = \boldsymbol{Z}(0)$. Now, because $\boldsymbol{Z}(0)$ is the column vector with elements $[Z_0, \hat{Z}_0(1), \ldots, \hat{Z}_0(m-1)]$, it is necessary for us to evaluate

$$Cov[\hat{Z}_0(i), \hat{Z}_0(j)] \qquad \text{for } i, j = 0, 1, \ldots, m-1$$

From the truncated linear process form, Equation (9.3.35) on page 200 with $C_t(\ell) = \hat{Z}_t(\ell)$, we may write, for $j > 0$

$$Z_j = \hat{Z}_0(j) + \sum_{k=-j}^{-1} \psi_{j+k}\, e_{-k} \tag{9.H.18}$$

Multiplying Equation (9.H.18) by Z_0 and taking expected values yields

$$\gamma_j = E(Z_0 Z_j) = E[\hat{Z}_0(0)(\hat{Z}_0(j))] \qquad \text{for } j \geq 0 \tag{9.H.19}$$

Now multiply Equation (9.H.18) by itself with j replaced by i and take expected values. Recalling that the e's are independent of past Z's and assuming $0 < i \leq j$, we obtain

$$\gamma_{j-i} = Cov[\hat{Z}_0(i), \hat{Z}_0(j)] + \sigma_e^2 \sum_{k=0}^{i-1} \psi_k \psi_{k+j-i} \tag{9.H.20}$$

Combining Equations (9.H.19) and (9.H.20), we have as the required elements of $\boldsymbol{P}(0|0)$

$$Cov[\hat{Z}_0(i), \hat{Z}_0(j)] = \begin{cases} \gamma_i & 0 = i \leq j \leq m-1 \\ \gamma_{j-i} - \sigma_e^2 \sum_{k=0}^{i-1} \psi_k \psi_{k+j-i} & 1 \leq i \leq j \leq m-1 \end{cases} \tag{9.H.21}$$

where the ψ-weights are obtained from the recursion of Equation (4.4.7) on page 79, and γ_k, the autocovariance function for the $\{Z_t\}$ process, is obtained as in Appendix C on page 85.

The variance σ_e^2 can be removed from the problem by dividing σ_ε^2 by σ_e^2. The prediction error variance v_t is then replaced by $\sigma_e^2 v_t$ in the log-likelihood of Equation (9.H.17), and we set $\sigma_e^2 = 1$ in Equation (9.H.8). Dropping unneeded constants, we get the new log-likelihood

$$\ell = \sum_{t=1}^{n} \left\{ \log(\sigma_e^2 v_t) + \frac{[y_t - y(t|t-1)]^2}{v_t} \right\} \tag{9.H.22}$$

which can be minimized analytically with respect to σ_e^2. We obtain

$$\sigma_e^2 = \sum_{t=1}^{n} \left\{ \frac{[y_t - y(t|t-1)]^2}{\sigma_e^2 v_t} \right\} \tag{9.H.23}$$

Substituting this back into Equation (9.H.22), we now find that

$$\ell = \sum_{t=1}^{n} \log v_t + n \log \sum_{t=1}^{n} \frac{[y_t - y(t|t-1)]^2}{v_t} \tag{9.H.24}$$

which must be minimized numerically with respect to $\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$, and σ_e^2. Having done so, we return to Equation (9.H.23) to estimate σ_e^2. The function defined by Equation (9.H.24) is sometimes called the **concentrated log-likelihood function**.