LECTURE 7

Forecasting with ARMA Models

Minimum Mean-Square Error Prediction

Imagine that $y(t)$ is a stationary stochastic process with $E\{y(t)\} = 0$. We may be interested in predicting values of this process several periods into the future on the basis of its observed history. This history is contained in the so-called information set. In practice, the latter is always a finite set $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$ representing the recent past. Nevertheless, in developing the theory of prediction, it is also useful to consider an infinite information set $\mathcal{I}_t = \{y_t, y_{t-1}, \ldots, y_{t-p}, \ldots\}$ representing the entire past.

We shall denote the prediction of y_{t+h} which is made at the time t by $\hat{y}_{t+h|t}$ or by $\hat{y}_{t+h}$ when it is clear that we are predicting h steps ahead.

The criterion which is commonly used in judging the performance of an estimator or predictor $\hat{y}$ of a random variable y is its mean-square error defined by $E\{(y - \hat{y})^2\}$. If all of the available information on y is summarised in its marginal distribution, then the minimum-mean-square-error prediction is simply the expected value $E(y)$. However, if y is statistically related to another random variable x whose value can be observed, and if the form of the joint distribution of x and y is known, then the minimum-mean-square-error prediction of y is the conditional expectation $E(y|x)$. This proposition may be stated formally:

(1) Let $\hat{y} = \hat{y}(x)$ be the conditional expectation of y given x which is also expressed as $\hat{y} = E(y|x)$. Then $E\{(y - \hat{y})^2\} \leq E\{(y - \pi)^2\}$, where $\pi = \pi(x)$ is any other function of x.

Proof. Consider

$$(2) \qquad \begin{aligned} E\big\{(y - \pi)^2\big\} &= E\Big[\big\{(y - \hat{y}) + (\hat{y} - \pi)\big\}^2\Big] \\ &= E\big\{(y - \hat{y})^2\big\} + 2E\big\{(y - \hat{y})(\hat{y} - \pi)\big\} + E\big\{(\hat{y} - \pi)^2\big\} \end{aligned}$$

Within the second term, there is

$$\begin{aligned} E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} &= \int_x\int_y (y-\hat{y})(\hat{y}-\pi)f(x,y)\partial y\partial x \\ &= \int_x \left\{ \int_y (y-\hat{y})f(y|x)\partial y \right\}(\hat{y}-\pi)f(x)\partial x \\ &= 0. \end{aligned} \tag{3}$$

Here the second equality depends upon the factorisation $f(x,y) = f(y|x)f(x)$ which expresses the joint probability density function of x and y as the product of the conditional density function of y given x and the marginal density function of x. The final equality depends upon the fact that $\int(y-\hat{y})f(y|x)\partial y = E(y|x) - E(y|x) = 0$. Therefore $E\{(y-\pi)^2\} = E\{(y-\hat{y})^2\} + E\{(\hat{y}-\pi)^2\} \geq E\{(y-\hat{y})^2\}$, and the assertion is proved.

The definition of the conditional expectation implies that

$$\begin{aligned} E(xy) &= \int_x\int_y xyf(x,y)\partial y\partial x \\ &= \int_x x\left\{ \int_y yf(y|x)\partial y \right\}f(x)\partial x \\ &= E(x\hat{y}). \end{aligned} \tag{4}$$

When the equation $E(xy) = E(x\hat{y})$ is rewritten as

$$E\big\{x(y-\hat{y})\big\} = 0, \tag{5}$$

it may be described as an orthogonality condition. This condition indicates that the prediction error $y-\hat{y}$ is uncorrelated with x. The result is intuitively appealing; for, if the error were correlated with x, we should not using the information of x efficiently in forming $\hat{y}$.

The proposition of (1) is readily generalised to accommodate the case where, in place of the scalar x, there is a vector $x = [x_1, \ldots, x_p]'$. This generalisation indicates that the minimum-mean-square-error prediction of y_{t+h} given the information in $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$ is the conditional expectation $E(y_{t+h}|y_t, y_{t-1}, \ldots, y_{t-p})$.

In order to determine the conditional expectation of y_{t+h} given $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$, we need to known the functional form of the joint probability density function all of these variables. In lieu of precise knowledge, we are often prepared to assume that the distribution is normal. In that case, it follows that the conditional expectation of y_{t+h} is a linear function of $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$; and so the problem of predicting y_{t+h} becomes a matter of forming a linear

regression. Even if we are not prepared to assume that the joint distribution of the variables in normal, we may be prepared, nevertheless, to base the prediction of y upon a linear function of $\{y_t, y_{t-1}, \ldots, y_{t-p}\}$. In that case, the criterion of minimum-mean-square-error linear prediction is satisfied by forming $\hat{y}_{t+h} = \phi_1 y_y + \phi_2 y_{t-1} + \cdots + \phi_{p+1} y_{t-p}$ from the values $\phi_1, \ldots, \phi_{p+1}$ which minimise

$$\begin{aligned} E\left\{(y_{t+h} - \hat{y}_{t+h})^2\right\} &= E\left\{\left(y_{t+h} - \sum_{j=1}^{p+1} \phi_j y_{t-j+1}\right)^2\right\} \\ &= \gamma_0 - 2\sum_j \phi_j \gamma_{h+j-1} + \sum_i \sum_j \phi_i \phi_j \gamma_{i-j}, \end{aligned} \tag{6}$$

wherein $\gamma_{i-j} = E(\varepsilon_{t-i}\varepsilon_{t-j})$. This is a linear least-squares regression problem which leads to a set of $p+1$ orthogonality conditions described as the normal equations:

$$\begin{aligned} E\big\{(y_{t+h} - \hat{y}_{t+h})y_{t-j+1}\big\} &= \gamma_{h+j-1} - \sum_{i=1}^{p} \phi_i \gamma_{i-j} \\ &= 0 \quad ; \quad j = 1, \ldots, p+1. \end{aligned} \tag{7}$$

In matrix terms, these are

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_p \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_p & \gamma_{p-1} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{p+1} \end{bmatrix} = \begin{bmatrix} \gamma_h \\ \gamma_{h+1} \\ \vdots \\ \gamma_{h+p} \end{bmatrix}. \tag{8}$$

Notice that, for the one-step-ahead prediction of y_{t+1}, they are nothing but the Yule–Walker equations.

In the case of an optimal predictor which combines previous values of the series, it follows from the orthogonality principle that the forecast errors are uncorrelated with the previous predictions.

A result of this sort is familiar to economists in connection with the so-called efficient-markets hypothesis. A financial market is efficient if the prices of the traded assets constitute optimal forecasts of their discounted future returns, which consist of interest and dividend payments and of capital gains.

According to the hypothesis, the changes in asset prices will be uncorrelated with the past or present price levels; which is to say that asset prices will follow random walks. Moreover, it should not be possible for someone who is appraised only of the past history of asset prices to reap speculative profits on a systematic and regular basis.

Forecasting with ARMA Models

So far, we have avoided making specific assumptions about the nature of the process $y(t)$. We are greatly assisted in the business of developing practical forecasting procedures if we can assume that $y(t)$ is generated by an ARMA process such that

$$y(t) = \frac{\mu(L)}{\alpha(L)}\varepsilon(t) = \psi(L)\varepsilon(t). \tag{9}$$

We shall continue to assume, for the sake of simplicity, that the forecasts are based on the information contained in the infinite set $\{y_t, y_{t-1}, y_{t-2}, \ldots\} = \mathcal{I}_t$ comprising all values that have been taken by the variable up to the present time t. Knowing the parameters in $\psi(L)$ enables us to recover the sequence $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$ from the sequence $\{y_t, y_{t-1}, y_{t-2}, \ldots\}$ and vice versa; so either of these constitute the information set. This equivalence implies that the forecasts may be expressed in terms $\{y_t\}$ or in terms $\{\varepsilon_t\}$ or as a combination of the elements of both sets.

Let us write the realisations of equation (9) as

$$\begin{aligned} y_{t+h} = &\{\psi_0\varepsilon_{t+h} + \psi_1\varepsilon_{t+h-1} + \cdots + \psi_{h-1}\varepsilon_{t+1}\} \\ &+ \{\psi_h\varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \cdots\}. \end{aligned} \tag{10}$$

Here the first term on the RHS embodies disturbances subsequent to the time t when the forecast is made, and the second term embodies disturbances which are within the information set $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. Let us now define a forecasting function, based on the information set, which takes the form of

$$\hat{y}_{t+h|t} = \{\rho_h\varepsilon_t + \rho_{h+1}\varepsilon_{t-1} + \cdots\}. \tag{11}$$

Then, given that $\varepsilon(t)$ is a white-noise process, it follows that the mean square of the error in the forecast h periods ahead is given by

$$E\big\{(y_{t+h} - \hat{y}_{t+h})^2\big\} = \sigma_\varepsilon^2 \sum_{i=0}^{h-1} \psi_i^2 + \sigma_\varepsilon^2 \sum_{i=h}^{\infty} (\psi_i - \rho_i)^2. \tag{12}$$

Clearly, the mean-square error is minimised by setting $\rho_i = \psi_i$; and so the optimal forecast is given by

$$\hat{y}_{t+h|t} = \{\psi_h\varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \cdots\}. \tag{13}$$

This might have been derived from the equation $y(t+h) = \psi(L)\varepsilon(t+h)$, which generates the true value of y_{t+h}, simply by putting zeros in place of the unobserved disturbances $\varepsilon_{t+1}, \varepsilon_{t+2}, \ldots, \varepsilon_{t+h}$ which lie in the future when the

forecast is made. Notice that, on the assumption that the process is stationary, the mean-square error of the forecast tends to the value of

$$V\{y(t)\} = \sigma_\varepsilon^2 \sum \psi_i^2 \tag{14}$$

as the lead time h of the forecast increases. This is nothing but the variance of the process $y(t)$.

The optimal forecast of (5) may also be derived by specifying that the forecast error should be uncorrelated with the disturbances up to the time of making the forecast. For, if the forecast errors were correlated with some of the elements of the information set, then, as we have noted before, we would not be using the information efficiently, and we could not be generating optimal forecasts. To demonstrate this result anew, let us consider the covariance between the forecast error and the disturbance ε_{t-i}:

$$\begin{aligned} E\{(y_{t+h} - \hat{y}_{t+h})\varepsilon_{t-i}\} &= \sum_{k=1}^{h} \psi_{h-k} E(\varepsilon_{t+k}\varepsilon_{t-i}) \\ &\quad + \sum_{j=0}^{\infty} (\psi_{h+j} - \rho_{h+j}) E(\varepsilon_{t-j}\varepsilon_{t-i}) \\ &= \sigma_\varepsilon^2 (\psi_{h+i} - \rho_{h+i}). \end{aligned} \tag{15}$$

Here the final equality follows from the fact that

$$E(\varepsilon_{t-j}\varepsilon_{t-i}) = \begin{cases} \sigma_\varepsilon^2, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \tag{16}$$

If the covariance in (15) is to be equal to zero for all values of $i \geq 0$, then we must have $\rho_i = \psi_i$ for all i, which means that the forecasting function must be the one that has been specified already under (13).

It is helpful, sometimes, to have a functional notation for describing the process which generates the h-steps-ahead forecast. The notation provided by Whittle (1963) is widely used. To derive this, let us begin by writing

$$y(t+h) = \{L^{-h}\psi(L)\}\varepsilon(t). \tag{17}$$

On the LHS, there are not only the lagged sequences $\{\varepsilon(t), \varepsilon(t-1), \ldots\}$ but also the sequences $\varepsilon(t+h) = L^{-h}\varepsilon(t), \ldots, \varepsilon(t+1) = L^{-1}\varepsilon(t)$, which are associated with negative powers of L which serve to shift a sequence forwards in time. Let $\{L^{-h}\psi(L)\}_+$ be defined as the part of the operator containing only nonnegative powers of L. Then the forecasting function can be expressed as

$$\begin{aligned} \hat{y}(t+h|t) &= \{L^{-h}\psi(L)\}_+ \varepsilon(t), \\ &= \left\{\frac{\psi(L)}{L^h}\right\}_+ \frac{1}{\psi(L)} y(t). \end{aligned} \tag{18}$$

Example. Consider an ARMA (1, 1) process represented by the equation

$$(1 - \phi L)y(t) = (1 - \theta L)\varepsilon(t). \tag{19}$$

The function which generates the sequence of forecasts h steps ahead is given by

$$\begin{aligned}\hat{y}(t+h|t) &= \left\{L^{-h}\left[1 + \frac{(\phi-\theta)L}{1-\phi L}\right]\right\}_+ \varepsilon(t) \\ &= \phi^{h-1}\frac{(\phi-\theta)}{1-\phi L}\varepsilon(t) \\ &= \phi^{h-1}\frac{(\phi-\theta)}{1-\theta L}y(t).\end{aligned} \tag{20}$$

When $\theta = 0$, this gives the simple result that $\hat{y}(t+h|t) = \phi^h y(t)$.

Generating The Forecasts Recursively

We have already seen that the optimal (minimum-mean-square-error) forecast of y_{t+h} can be regarded as the conditional expectation of y_{t+h} given the information set $\mathcal{I}_t$ which comprises the values of $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$ or equally the values of $\{y_t, y_{t-1}, y_{t-2}, \ldots\}$. On taking expectations of $y(t)$ and $\varepsilon(t)$ conditional on $\mathcal{I}_t$, we find that

$$\begin{aligned}E(y_{t+k}|\mathcal{I}_t) &= \hat{y}_{t+k|t} \quad \text{if} \quad k > 0, \\ E(y_{t-j}|\mathcal{I}_t) &= y_{t-j} \quad \text{if} \quad j \geq 0, \\ E(\varepsilon_{t+k}|\mathcal{I}_t) &= 0 \quad \text{if} \quad k > 0, \\ E(\varepsilon_{t-j}|\mathcal{I}_t) &= \varepsilon_{t-j} \quad \text{if} \quad j \geq 0.\end{aligned} \tag{21}$$

In this notation, the forecast h periods ahead is

$$\begin{aligned}E(y_{t+h}|\mathcal{I}_t) &= \sum_{k=1}^{h}\psi_{h-k}E(\varepsilon_{t+k}|\mathcal{I}_t) + \sum_{j=0}^{\infty}\psi_{h+j}E(\varepsilon_{t-j}|\mathcal{I}_t) \\ &= \sum_{j=0}^{\infty}\psi_{h+j}\varepsilon_{t-j}.\end{aligned} \tag{22}$$

In practice, the forecasts may be generated using a recursion based on the equation

$$\begin{aligned}y(t) = &-\{\alpha_1 y(t-1) + \alpha_2 y(t-2) + \cdots + \alpha_p y(t-p)\} \\ &+ \mu_0\varepsilon(t) + \mu_1\varepsilon(t-1) + \cdots + \mu_q\varepsilon(t-q).\end{aligned} \tag{23}$$

By taking the conditional expectation of this function, we get

$$\begin{aligned}\hat{y}_{t+h} &= -\{\alpha_1\hat{y}_{t+h-1} + \cdots + \alpha_p y_{t+h-p}\} \\ &\quad + \mu_h\varepsilon_t + \cdots + \mu_q\varepsilon_{t+h-q} \quad \text{when} \quad 0 < h \leq p, q,\end{aligned} \tag{24}$$

$$\hat{y}_{t+h} = -\{\alpha_1\hat{y}_{t+h-1} + \cdots + \alpha_p y_{t+h-p}\} \quad \text{if} \quad q < h \leq p, \tag{25}$$

$$\begin{aligned}\hat{y}_{t+h} &= -\{\alpha_1\hat{y}_{t+h-1} + \cdots + \alpha_p\hat{y}_{t+h-p}\} \\ &\quad + \mu_h\varepsilon_t + \cdots + \mu_q\varepsilon_{t+h-q} \quad \text{if} \quad p < h \leq q,\end{aligned} \tag{26}$$

and

$$\hat{y}_{t+h} = -\{\alpha_1\hat{y}_{t+h-1} + \cdots + \alpha_p\hat{y}_{t+h-p}\} \quad \text{when} \quad p, q < h. \tag{27}$$

It can be from (27) that, for $h > p, q$, the forecasting function becomes a pth-order homogeneous difference equation in y. The p values of $y(t)$ from $t = r = \max(p, q)$ to $t = r - p + 1$ serve as the starting values for the equation.

The behaviour of the forecast function beyond the reach of the starting values can be characterised in terms of the roots of the autoregressive operator. It may be assumed that none of the roots of $\alpha(L) = 0$ lie inside the unit circle; for, if there were roots inside the circle, then the process would be radically unstable. If all of the roots are less than unity, then $\hat{y}_{t+h}$ will converge to zero as h increases. If one of the roots of $\alpha(L) = 0$ is unity, then we have an ARIMA$(p, 1, q)$ model; and the general solution of the homogeneous equation of (27) will include a constant term which represents the product of the unit root with an coefficient which is determined by the starting values. Hence the forecast will tend to a nonzero constant. If two of the roots are unity, then the general solution will embody a linear time trend which is the asymptote to which the forecasts will tend. In general, if d of the roots are unity, then the general solution will comprise a polynomial in t of order $d - 1$.

The forecasts can be updated easily once the coefficients in the expansion of $\psi(L) = \mu(L)/\alpha(L)$ have been obtained. Consider

$$\begin{aligned}\hat{y}_{t+h|t+1} &= \{\psi_{h-1}\varepsilon_{t+1} + \psi_h\varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \cdots\} \quad \text{and} \\ \hat{y}_{t+h|t} &= \{\psi_h\varepsilon_t + \psi_{h+1}\varepsilon_{t-1} + \psi_{h+2}\varepsilon_{t-2} + \cdots\}.\end{aligned} \tag{28}$$

The first of these is the forecast for $h - 1$ periods ahead made at time $t + 1$ whilst the second is the forecast for h periods ahead made at time t. It can be seen that

$$\hat{y}_{t+h|t+1} = \hat{y}_{t+h|t} + \psi_{h-1}\varepsilon_{t+1}, \tag{29}$$

where $\varepsilon_{t+1} = y_{t+1} - \hat{y}_{t+1}$ is the current disturbance at time $t+1$. The later is also the prediction error of the one-step-ahead forecast made at time t.

Example. For an example of the analytic form of the forecast function, we may consider the Integrated Autoregressive (IAR) Process defined by

$$\left\{1 - (1+\phi)L + \phi L^2\right\}y(t) = \varepsilon(t), \tag{30}$$

wherein $\phi \in (0,1)$. The roots of the auxiliary equation $z^2 - (1+\phi)z + \phi = 0$ are $z = 1$ and $z = \phi$. The solution of the homogeneous difference equation

$$\left\{1 - (1+\phi)L + \phi L^2\right\}\hat{y}(t+h|t) = 0, \tag{31}$$

which defines the forecast function, is

$$\hat{y}(t+h|t) = c_1 + c_2\phi^h, \tag{32}$$

where c_1 and c_2 are constants which reflect the initial conditions. These constants are found by solving the equations

$$\begin{aligned} y_{t-1} &= c_1 + c_2\phi^{-1}, \\ y_t &= c_1 + c_2. \end{aligned} \tag{33}$$

The solutions are

$$c_1 = \frac{y_t - \phi y_{t-1}}{1-\phi} \qquad \text{and} \qquad c_2 = \frac{\phi}{\phi - 1}(y_t - y_{t-1}). \tag{34}$$

The long-term forecast is $\bar{y} = c_1$ which is the asymptote to which the forecasts tend as the lead period h increases.

Ad-hoc Methods of Forecasting

There are some time-honoured methods of forecasting which, when analysed carefully, reveal themselves to be the methods which are appropriate to some simple ARIMA models which might be suggested by *a priori* reasoning. Two of the leading examples are provided by the method of exponential smoothing and the Holt–Winters trend-extrapolation method.

Exponential Smoothing. A common forecasting procedure is exponential smoothing. This depends upon taking a weighted average of past values of the time series with the weights following a geometrically declining pattern. The function generating the one-step-ahead forecasts can be written as

$$\begin{aligned} \hat{y}(t+1|t) &= \frac{(1-\theta)}{1-\theta L}y(t) \\ &= (1-\theta)\left\{y(t) + \theta y(t-1) + \theta^2 y(t-2) + \cdots\right\}. \end{aligned} \tag{35}$$

On multiplying both sides of this equation by $1 - \theta L$ and rearranging, we get

$$\hat{y}(t+1|t) = \theta\hat{y}(t|t-1) + (1-\theta)y(t), \tag{36}$$

which shows that the current forecast for one step ahead is a convex combination of the previous forecast and the value which actually transpired.

The method of exponential smoothing corresponds to the optimal forecasting procedure for the ARIMA$(0, 1, 1)$ model $(1 - L)y(t) = (1 - \theta L)\varepsilon(t)$, which is better described as an IMA$(1, 1)$ model. To see this, let us consider the ARMA$(1, 1)$ model $y(t) - \phi y(t-1) = \varepsilon(t) - \theta\varepsilon(t-1)$. This gives

$$\begin{aligned}
\hat{y}(t+1|t) &= \phi y(t) - \theta\varepsilon(t) \\
&= \phi y(t) - \theta\frac{(1-\phi L)}{1-\theta L}y(t) \\
&= \frac{\{(1-\theta L)\phi - (1-\phi L)\theta\}}{1-\theta L}y(t) \\
&= \frac{(\phi-\theta)}{1-\theta L}y(t).
\end{aligned} \tag{37}$$

On setting $\phi = 1$, which converts the ARMA$(1, 1)$ model to an IMA$(1, 1)$ model, we obtain precisely the forecasting function of (35).

The Holt–Winters Method. The Holt–Winters algorithm is useful in extrapolating local linear trends. The prediction h periods ahead of a series $y(t) = \{y_t, t = 0, \pm1, \pm2, \ldots\}$ which is made at time t is given by

$$\hat{y}_{t+h|t} = \hat{\alpha}_t + \hat{\beta}_t h, \tag{38}$$

where

$$\begin{aligned}
\hat{\alpha}_t &= \lambda y_t + (1-\lambda)(\hat{\alpha}_{t-1} + \hat{\beta}_{t-1}) \\
&= \lambda y_t + (1-\lambda)\hat{y}_{t|t-1}
\end{aligned} \tag{39}$$

is the estimate of an intercept or levels parameter formed at time t and

$$\hat{\beta}_t = \mu(\hat{\alpha}_t - \hat{\alpha}_{t-1}) + (1-\mu)\hat{\beta}_{t-1} \tag{40}$$

is the estimate of the slope parameter, likewise formed at time t. The coefficients $\lambda, \mu \in (0, 1]$ are the smoothing parameters.

The algorithm may also be expressed in error-correction form. Let

$$e_t = y_t - \hat{y}_{t|t-1} = y_t - \hat{\alpha}_{t-1} - \hat{\beta}_{t-1} \tag{41}$$

be the error at time t arising from the prediction of y_t on the basis of information available at time $t-1$. Then the formula for the levels parameter can be given as

$$\begin{aligned}\hat{\alpha}_t &= \lambda e_t + \hat{y}_{t|t-1} \\ &= \lambda e_t + \hat{\alpha}_{t-1} + \hat{\beta}_{t-1},\end{aligned} \tag{42}$$

which, on rearranging, becomes

$$\hat{\alpha}_t - \hat{\alpha}_{t-1} = \lambda e_t + \hat{\beta}_{t-1}. \tag{43}$$

When the latter is drafted into equation (40), we get an analogous expression for the slope parameter:

$$\begin{aligned}\hat{\beta}_t &= \mu(\lambda e_t + \hat{\beta}_{t-1}) + (1-\mu)\hat{\beta}_{t-1} \\ &= \lambda\mu e_t + \hat{\beta}_{t-1}.\end{aligned} \tag{44}$$

In order reveal the underlying nature of this method, it is helpful to combine the two equations (42) and (44) in a simple state-space model:

$$\begin{bmatrix}\hat{\alpha}(t) \\ \hat{\beta}(t)\end{bmatrix} = \begin{bmatrix}1 & 1 \\ 0 & 1\end{bmatrix}\begin{bmatrix}\hat{\alpha}(t-1) \\ \hat{\beta}(t-1)\end{bmatrix} + \begin{bmatrix}\lambda \\ \lambda\mu\end{bmatrix} e(t). \tag{45}$$

This can be rearranged to give

$$\begin{bmatrix}1-L & -L \\ 0 & 1-L\end{bmatrix}\begin{bmatrix}\hat{\alpha}(t) \\ \hat{\beta}(t)\end{bmatrix} = \begin{bmatrix}\lambda \\ \lambda\mu\end{bmatrix} e(t). \tag{46}$$

The solution of the latter is

$$\begin{bmatrix}\hat{\alpha}(t) \\ \hat{\beta}(t)\end{bmatrix} = \frac{1}{(1-L)^2}\begin{bmatrix}1-L & L \\ 0 & 1-L\end{bmatrix}\begin{bmatrix}\lambda \\ \lambda\mu\end{bmatrix} e(t). \tag{47}$$

Therefore, from (38), it follows that

$$\begin{aligned}\hat{y}(t+1|t) &= \hat{\alpha}(t) + \hat{\beta}(t) \\ &= \frac{(\lambda+\lambda\mu)e(t) + \lambda e(t-1)}{(1-L)^2}.\end{aligned} \tag{48}$$

This can be recognised as the forecasting function of an IMA(2, 2) model of the form

$$(I-L)^2 y(t) = \mu_0\varepsilon(t) + \mu_1\varepsilon(t-1) + \mu_2\varepsilon(t-2) \tag{49}$$

for which

$$\hat{y}(t+1|t) = \frac{\mu_1\varepsilon(t) + \mu_2\varepsilon(t-1)}{(1-L)^2}. \tag{50}$$

The Local Trend Model. There are various arguments which suggest that an IMA(2, 2) model might be a natural model to adopt. The simplest of these arguments arises from an elaboration of a second-order random walk which adds an ordinary white-noise disturbance to the tend. The resulting model may be expressed in two equations

$$\begin{aligned} (I-L)^2\xi(t) &= \nu(t), \\ y(t) &= \xi(t) + \eta(t), \end{aligned} \tag{51}$$

where $\nu(t)$ and $\eta(t)$ are mutually independent white-noise processes. Combining the equations, and using the notation $\nabla = 1 - L$, gives

$$\begin{aligned} y(t) &= \frac{\nu(t)}{\nabla^2} + \eta(t) \\ &= \frac{\nu(t) + \nabla^2\eta(t)}{\nabla^2}. \end{aligned} \tag{52}$$

Here the numerator $\nu(t)+\nabla^2\eta(t) = \{\nu(t)+\eta(t)\}-2\eta(t-1)+\eta(t-2)$ constitutes an second-order MA process.

Slightly more elaborate models with the same outcome have also been proposed. Thus the so-called structural model consists of the equations

$$\begin{aligned} y(t) &= \mu(t) + \varepsilon(t), \\ \mu(t) &= \mu(t-1) + \beta(t-1) + \eta(t), \\ \beta(t) &= \beta(t-1) + \zeta(t). \end{aligned} \tag{53}$$

Working backwards from the final equation gives

$$\begin{aligned} \beta(t) &= \frac{\zeta(t)}{\nabla}, \\ \mu(t) &= \frac{\beta(t-1)}{\nabla} + \frac{\eta(t)}{\nabla} \\ &= \frac{\zeta(t-1)}{\nabla^2} + \frac{\eta(t)}{\nabla}, \\ y(t) &= \frac{\zeta(t-1)}{\nabla^2} + \frac{\eta(t)}{\nabla} + \varepsilon(t) \\ &= \frac{\zeta(t-1) + \nabla\eta(t) + \nabla^2\varepsilon(t)}{\nabla^2}. \end{aligned} \tag{54}$$

Once more, the numerator constitutes a second-order MA process.

Equivalent Forecasting Functions

Consider a model which combines a global linear trend with an autoregressive disturbance process:

$$y(t) = \gamma_0 + \gamma_1 t + \frac{\varepsilon(t)}{I - \phi L}. \tag{55}$$

The formation of an h-step-ahead prediction is straightforward; for we can separate the forecast function into two additive parts.

The first part of the function is the extrapolation of the global linear trend. This takes the form of

$$\begin{aligned} z_{t+h|t} &= \gamma_0 + \gamma_1(t+h) \\ &= z_t + \gamma_1 h \end{aligned} \tag{56}$$

where $z_t = \gamma_0 + \gamma_1 t$.

The second part is the prediction associated with the AR(1) disturbance term $\eta(t) = (I - \phi L)^{-1}\varepsilon(t)$. The following iterative scheme is provides a recursive solution to the problem of generating the forecasts:

$$\begin{aligned} \hat{\eta}_{t+1|t} &= \phi\eta_t, \\ \hat{\eta}_{t+2|t} &= \phi\hat{\eta}_{t+1|t}, \\ \hat{\eta}_{t+3|t} &= \phi\hat{\eta}_{t+2|t}, \quad \text{etc.} \end{aligned} \tag{57}$$

Notice that the analytic solution of the associated difference equation is just

$$\hat{\eta}_{t+h|t} = \phi^h \eta_t. \tag{58}$$

This reminds us that, whenever we can express the forecast function in terms of a linear recursion, we can also express it in an analytic form embodying the roots of a polynomial lag operator. The operator in this case is the AR(1) operator $I - \phi L$. Since, by assumption, $|\phi| < 1$, it is clear that the contribution of the disturbance part to the overall forecast function

$$\hat{y}_{t+h|t} = z_{t+h|t} + \hat{\eta}_{t+h|t}, \tag{59}$$

becomes negligible when h becomes large.

Consider the limiting case when $\phi \to 1$. Now, in place of an AR(1) disturbance process, we have to consider a random-walk process. We know that the forecast function of a random walk consists of nothing more than a constant

function. On adding this constant to the linear function $z_{t+h|t} = \gamma_0 + \gamma_1(t+h)$ we continue to have a simple linear forecast function.

Another way of looking at the problem depends upon writing equation (55) as

$$(I - \phi L)\{y(t) - \gamma_0 - \gamma_1 t\} = \varepsilon(t). \tag{60}$$

Setting $\phi = 1$ turns the operator $I - \phi L$ into the difference operator $I - L = \nabla$. But $\nabla\gamma_0 = 0$ and $\nabla\gamma_1 t = \gamma_1$, so equation (60) with $\phi = 1$ can also be written as

$$\nabla y(t) = \gamma_1 + \varepsilon(t). \tag{61}$$

This is the equation of a process which is described as random walk with drift. Yet another way of expressing the process is via the equation $y(t) = y(t-1) + \gamma_1 + \varepsilon(t)$.

It is intuitively clear that, if the random walk process $\nabla z(t) = \varepsilon(t)$ is associated with a constant forecast function, and if $z(t) = y(t) - \gamma_0 - \gamma_1 t$, then $y(t)$ will be associated with a linear forecast function.

The purpose of this example has been to offer a limiting case where models with local stochastic trends—ie. random walk and unit root models—and models with global polynomial trends come together. Finally, we should notice that the model of random walk with drift has the same linear forecast function as the model

$$\nabla^2 y(t) = \varepsilon(t) \tag{62}$$

which has two unit roots in the AR operator.