$$\frac{\partial^2 l}{\partial(\sigma_a^2)\partial\phi_j} \simeq -\sigma_a^{-2}\frac{\partial l}{\partial\phi_j} \tag{A7.4.22b}$$

$$-\frac{\partial^2 l}{\partial\phi_i\partial\phi_j} \simeq \frac{n}{\sigma_a^2}\frac{D_{i+1,j+1}}{n-i-j} \tag{A7.4.22c}$$

Now, since

$$E\left[\frac{\partial l}{\partial\phi_j}\right] = 0$$

it follows that for moderate or large samples,

$$E\left[-\frac{\partial^2 l}{\partial(\sigma_a^2)\partial\phi_j}\right] \simeq 0$$

and

$$|\mathbf{I}(\boldsymbol{\phi}, \sigma_a^2)| \simeq |\mathbf{I}(\boldsymbol{\phi})| I(\sigma_a^2)$$

where

$$I(\sigma_a^2) = E\left[-\frac{\partial^2 l}{\partial(\sigma_a^2)^2}\right] = \frac{n}{2(\sigma_a^2)^2}$$

Now, using (A7.4.22c), we have

$$\mathbf{I}(\boldsymbol{\phi}) = -E\left[\frac{\partial^2 l}{\partial\phi_i\partial\phi_j}\right] \simeq \frac{n}{\sigma_a^2}\boldsymbol{\Gamma}_p = \frac{n\boldsymbol{\gamma}_0}{\sigma_a^2}\mathbf{P}_p = n(\mathbf{M}_p^{(p)})^{-1} \tag{A7.4.23}$$

Hence,

$$|\mathbf{I}(\boldsymbol{\phi}, \sigma_a^2)| \simeq \frac{n^{p+1}}{2(\sigma_a^2)^2}|\mathbf{M}_p^{(p,0)}|^{-1}$$

Variances and Covariances of Estimates of Autoregressive Parameters. Now, in circumstances fully discussed by Whittle (1953), the inverse of the information matrix supplies the asymptotic variance–covariance matrix of the maximum likelihood (ML) estimates. Moreover, if the log-likelihood is approximately quadratic and the maximum is not close to a boundary, even if the sample size is only moderate, the elements of this matrix will normally provide adequate approximations to the variances and covariances of the estimates.

Thus, using (A7.4.23) and (A7.4.21b) gives

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\phi}}) = \mathbf{I}^{-1}(\boldsymbol{\phi}) &\simeq n^{-1}\mathbf{M}_p^{(p)} = n^{-1}\sigma_a^2\boldsymbol{\Gamma}_p^{-1} \\ &= n^{-1}(1-\boldsymbol{\rho}'\mathbf{P}_p^{-1}\boldsymbol{\rho})\mathbf{P}_p^{-1} \\ &= n^{-1}(1-\boldsymbol{\phi}'\mathbf{P}_p\boldsymbol{\phi})\mathbf{P}_p^{-1} = n^{-1}(1-\boldsymbol{\rho}'\boldsymbol{\phi})\mathbf{P}_p^{-1}\end{aligned} \tag{A7.4.24}$$

In particular, for autoregressive process of first and second order,

$$V(\hat{\phi}) \simeq n^{-1}(1-\phi^2)$$
$$V(\hat{\phi}_1, \hat{\phi}_2) \simeq n^{-1} \begin{bmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{bmatrix} \tag{A7.4.25}$$

Estimates of the variances and covariances may be obtained by substituting estimates for the parameters in (A7.4.25). For example, we may substitute r_j's for ρ_j's and $\hat{\boldsymbol{\phi}}$ for $\boldsymbol{\phi}$ in (A7.4.24) to obtain

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\phi}}) = n^{-1}(1 - \mathbf{r}'\hat{\boldsymbol{\phi}})\mathbf{R}^{-1} \tag{A7.4.26}$$

APPENDIX A7.5 ASYMPTOTIC DISTRIBUTION OF ESTIMATORS FOR AUTOREGRESSIVE MODELS

We provide details on the asymptotic distribution of least-squares estimator of the parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ for a *stationary* AR(p) model [i.e., all roots of $\phi(B) = 0$ lie outside the unit circle],

$$w_t = \sum_{i=1}^{p} \phi_i w_{t-i} + a_t$$

based on a sample of n observations, where the w_t are assumed to have mean $\mu = 0$ for simplicity, and the a_t are assumed to be independent random variates, with zero means, variances σ_a^2, and finite fourth moments. It is then established that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})\} \tag{A7.5.1}$$

as $n \to \infty$, where $\boldsymbol{\Gamma}_p(\boldsymbol{\phi})$ is the $p \times p$ autocovariance matrix of p successive values from the AR(p) process. Hence, for large n the distribution of $\hat{\boldsymbol{\phi}}$ is approximately normal with mean vector $\boldsymbol{\phi}$ and covariance matrix $\mathbf{V}(\hat{\boldsymbol{\phi}}) \simeq n^{-1}\sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})$, that is, $N\{\boldsymbol{\phi}, n^{-1}\sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})\}$.

We can write the AR(p) model as

$$w_t = \mathbf{w}_{t-1}'\boldsymbol{\phi} + a_t \tag{A7.5.2}$$

where $\mathbf{w}_{t-1}' = (w_{t-1}, \ldots, w_{t-p})$. For convenience, assume that observations $w_{1-p}, \ldots, w_0$ are available in addition to $w_1, \ldots, w_n$, so that the (conditional) least-squares estimator of $\boldsymbol{\phi}$ is obtained by minimizing the sum of squares:

$$S(\boldsymbol{\phi}) = \sum_{t=1}^{n} (w_t - \mathbf{w}_{t-1}'\boldsymbol{\phi})^2$$

As $n \to \infty$, the treatment of the p initial observations becomes negligible, so that conditional and unconditional LS estimators are asymptotically equivalent. From the standard results on LS estimates for regression models, we know that the LS estimate of $\boldsymbol{\phi}$ in the AR(p)

model (A7.5.2) is then given by

$$\hat{\boldsymbol{\phi}} = \left(\sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} w_t \tag{A7.5.3}$$

Substituting the expression for w_t from (A7.5.2) in (A7.5.3), we see that

$$\hat{\boldsymbol{\phi}} = \boldsymbol{\phi} + \left(\sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t$$

so that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) = \left(n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} n^{-1/2} \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t \tag{A7.5.4}$$

Notice that the information matrix for this model situation is simply

$$\mathbf{I}(\boldsymbol{\phi}) = -\frac{1}{2\sigma_a^2} E\left[\frac{\partial^2 S(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} \right] = \frac{1}{\sigma_a^2} \sum_{t=1}^{n} E[\mathbf{w}_{t-1} \mathbf{w}'_{t-1}] = \frac{n}{\sigma_a^2} \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$$

so that $n\mathbf{I}^{-1}(\boldsymbol{\phi}) \equiv \mathbf{I}_*^{-1}(\boldsymbol{\phi}) = \sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})$ as appears in (A7.5.1).

We let $U_t = \mathbf{w}_{t-1} a_t$ and argue that these terms have zero mean, covariance matrix $\sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$, and are mutually uncorrelated. That is, noting that $\mathbf{w}_{t-1}$ and a_t are independent (e.g., elements of $\mathbf{w}_{t-1}$ are functions of $a_{t-1}, a_{t-2}, \ldots$, independent of a_t), we have $E[\mathbf{w}_{t-1} a_t] = E[\mathbf{w}_{t-1}] E[a_t] = 0$, and again by independence of the terms a_t^2 and $\mathbf{w}_{t-1} \mathbf{w}'_{t-1}$,

$$\operatorname{cov}[\mathbf{w}_{t-1} a_t] = E[\mathbf{w}_{t-1} a_t a_t \mathbf{w}'_{t-1}] = E[a_t^2] E[\mathbf{w}_{t-1} \mathbf{w}'_{t-1}] = \sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$$

In addition, for any $l > 0$,

$$\begin{aligned} \operatorname{cov}[\mathbf{w}_{t-1} a_t, \mathbf{w}_{t+l-1} a_{t+l}] &= E[\mathbf{w}_{t-1} a_t a_{t+l} \mathbf{w}'_{t+l-1}] \\ &= E[a_t \mathbf{w}_{t-1} \mathbf{w}'_{t+l-1}] E[a_{t+l}] = 0 \end{aligned}$$

because a_{t+l} is independent of the other terms. By similar reasoning,

$$\operatorname{cov}[\mathbf{w}_{t-1} a_t, \mathbf{w}_{t+l-1} a_{t+l}] = 0$$

for any $l < 0$. Hence, the quantity $\sum_{t=1}^{n} \mathbf{w}_{t-1} a_t$ in (A7.5.4) is the sum of n uncorrelated terms each with zero mean and covariance matrix $\sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$.

Now, in fact, the partial sums

$$\mathrm{S}_n = \sum_{t=1}^{n} \mathrm{U}_t \equiv \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t \qquad n = 1, 2, \ldots$$

form a *martingale* sequence (with respect to the σ fields generated by the collection of random variables $\{a_n, a_{n-1}, \ldots\}$), characterized by the property that $E[\mathrm{S}_{n+1} | a_n, a_{n-1}, \ldots] = \mathrm{S}_n$. This clearly holds since $\mathrm{S}_{n+1} = \mathbf{w}_n a_{n+1} + \mathrm{S}_n$,

$$E[\mathbf{w}_n a_{n+1} | a_n, a_{n-1}, \ldots] = \mathbf{w}_n E[a_{n+1} | a_n, a_{n-1}, \ldots] = \mathbf{w}_n E[a_{n+1}] = 0$$

and $S_n = \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t$ is a function of $a_n, a_{n-1}, \ldots$ so that $E[S_n | a_n, a_{n-1}, \ldots] = S_n$. In this context, the terms $U_t = \mathbf{w}_{t-1} a_t$ are referred to as *a martingale difference* sequence. Then, by a martingale central limit theorem (e.g., Billingsley, 1999),

$$n^{-1/2} \mathbf{c}' S_n \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \sigma_a^2 \mathbf{c}' \boldsymbol{\Gamma}_p(\boldsymbol{\phi}) \mathbf{c}\}$$

for any vector or constants $\mathbf{c}' = (c_1, \ldots, c_p)$, and by use of the Cramer–Wold device, it follows that

$$n^{-1/2} S_n \equiv n^{-1/2} \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})\} \tag{A7.5.5}$$

as $n \to \infty$. Also, we know that the matrix $n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \xrightarrow{\mathcal{P}} \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$, as an $n \to \infty$, by a weak law of large numbers, since the (i, j) th element of the matrix is $\hat{\gamma}(i-j) = n^{-1} \sum_{t=1}^{n} w_{t-1} w_{t-j}$, which converges in probability to $\gamma(i-j)$ by consistency of sample autocovariances $\hat{\gamma}(i-j)$. Hence, it follows by continuity that

$$\left(n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} \xrightarrow{\mathcal{P}} \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) \tag{A7.5.6}$$

Therefore, by a standard limit theory result, applying (A7.5.5) and (A7.5.6) in (A7.5.4), we obtain that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{\mathcal{D}} \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) N\{\mathbf{0}, \sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})\} \tag{A7.5.7}$$

which leads to the result (A7.5.1).

In addition, it is easily shown that the Yule–Walker (YW) estimator $\tilde{\boldsymbol{\phi}} = \mathbf{R}^{-1} \mathbf{r}$, discussed in Section 7.3.1, is asymptotically equivalent to the LS estimator considered here, in the sense that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}) \xrightarrow{\mathcal{P}} \mathbf{0}$$

as $n \to \infty$. For instance, we can write the YW estimate as $\tilde{\boldsymbol{\phi}} = \tilde{\boldsymbol{\Gamma}}_p^{-1} \tilde{\boldsymbol{\gamma}}_p$ where $\tilde{\boldsymbol{\Gamma}}_p = \hat{\gamma}_0 \mathbf{R}$ and $\tilde{\boldsymbol{\gamma}}_p = \hat{\gamma}_0 \mathbf{r}$. For notational convenience, we write the LS estimate in (A7.5.3) as $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p$ where we denote $\hat{\boldsymbol{\Gamma}}_p = n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1}$ and $\hat{\boldsymbol{\gamma}}_p = n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} w_t$. Then, we have

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}) &= n^{1/2}(\hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p - \tilde{\boldsymbol{\Gamma}}_p^{-1} \tilde{\boldsymbol{\gamma}}_p) \\ &= n^{1/2} \hat{\boldsymbol{\Gamma}}_p^{-1} (\hat{\boldsymbol{\gamma}}_p - \tilde{\boldsymbol{\gamma}}_p) + n^{1/2} (\hat{\boldsymbol{\Gamma}}_p^{-1} - \tilde{\boldsymbol{\Gamma}}_p^{-1}) \tilde{\boldsymbol{\gamma}}_p \end{aligned} \tag{A7.5.8}$$

and we can readily determine that both $n^{1/2}(\hat{\boldsymbol{\gamma}}_p - \tilde{\boldsymbol{\gamma}}_p) \xrightarrow{\mathcal{P}} \mathbf{0}$ and $n^{1/2}(\hat{\boldsymbol{\Gamma}}_p - \tilde{\boldsymbol{\Gamma}}_p) \xrightarrow{\mathcal{P}} \mathbf{0}$ as $n \to \infty$, and consequently also

$$n^{1/2}(\hat{\boldsymbol{\Gamma}}_p^{-1} - \tilde{\boldsymbol{\Gamma}}_p^{-1}) = \tilde{\boldsymbol{\Gamma}}_p^{-1} n^{1/2} (\tilde{\boldsymbol{\Gamma}}_p - \hat{\boldsymbol{\Gamma}}_p) \hat{\boldsymbol{\Gamma}}_p^{-1} \xrightarrow{\mathcal{P}} \mathbf{0}$$

Therefore, $n^{1/2}(\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}) \xrightarrow{\mathcal{P}} \mathbf{0}$ follows directly from (A7.5.8).

APPENDIX A7.6 EXAMPLES OF THE EFFECT OF PARAMETER ESTIMATION ERRORS ON VARIANCES OF FORECAST ERRORS AND PROBABILITY LIMITS FOR FORECASTS

The variances and probability limits for the forecasts given in Section 5.2.4 are based on the assumption that the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ in the ARIMA model are known exactly. In practice, it is necessary to replace these by their estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$. To gain some insight into the effect of estimation errors on the variance of the forecast errors, we consider the special cases of the nonstationary IMA(0, 1, 1) and the stationary first-order autoregressive processes. It is shown that for these processes and for parameter estimates based on series of moderate length, the effect of such estimation errors is small.

IMA(0, 1, 1) Processes. Writing the model $\nabla z_t = a_t - \theta a_{t-1}$ for $t + l, t + l - 1, \ldots, t + 1$, and summing, we obtain

$$z_{t+l} - z_t = a_{t+l} + (1 - \theta)(a_{t+l-1} + \cdots + a_{t+1}) - \theta a_t$$

Denote by $\hat{z}_t(l|\theta)$ the lead l forecast when the parameter θ is known exactly. On taking conditional expectations at time t, for $l = 1, 2, \ldots$, we obtain

$$\begin{aligned} \hat{z}_t(1|\theta) &= z_t - \theta a_t \\ \hat{z}_t(l|\theta) &= \hat{z}_t(1|\theta) \qquad l \geq 2 \end{aligned}$$

Hence, the lead l forecast error is

$$\begin{aligned} e_t(l|\theta) &= z_{t+l} - \hat{z}_t(l|\theta) \\ &= a_{t+l} + (1 - \theta)(a_{t+l-1} + \cdots + a_{t+1}) \end{aligned}$$

and the variance of the forecast error at lead time l is

$$V(l) = E_t[e_t^2(l|\theta)] = \sigma_a^2[1 + (l - 1)\lambda^2] \tag{A7.6.1}$$

where $\lambda = 1 - \theta$.

However, if θ is replaced by its estimate $\hat{\theta}$, obtained from a time series consisting of n values of $w_t = \nabla z_t$, then,

$$\begin{aligned} \hat{z}_t(1|\hat{\theta}) &= z_t - \hat{\theta}\hat{a}_t \\ \hat{z}_t(l|\hat{\theta}) &= \hat{z}_t(1|\hat{\theta}) \qquad l \geq 2 \end{aligned}$$

where $\hat{a}_t = z_t - \hat{z}_{t-1}(1|\hat{\theta})$. Hence, the lead l forecast error using $\hat{\theta}$ is

$$\begin{aligned} e_t(l|\hat{\theta}) &= z_{t+l} - \hat{z}(l|\hat{\theta}) \\ &= z_{t+l} - z_t + \hat{\theta}\hat{a}_t \\ &= e_t(l|\theta) - (\theta a_t - \hat{\theta}\hat{a}_t) \end{aligned} \tag{A7.6.2}$$

Since $\nabla z_t = (1 - \theta B)a_t = (1 - \hat{\theta}B)\hat{a}_t$, it follows that

$$\hat{a}_t = \left(\frac{1 - \theta B}{1 - \hat{\theta}B}\right) a_t$$

and on eliminating $\hat{a}_t$ from (A7.6.2), we obtain

$$e_t(l|\hat{\theta}) = e_t(l|\theta) - \frac{\theta - \hat{\theta}}{1 - \hat{\theta}B} a_t$$

Now,

$$\begin{aligned}
\frac{\theta - \hat{\theta}}{1 - \hat{\theta}B} a_t &= \frac{\theta - \hat{\theta}}{1 - \theta B} \left[1 + \frac{(\theta - \hat{\theta})B}{1 - \theta B} \right]^{-1} a_t \\
&\simeq \frac{\theta - \hat{\theta}}{1 - \theta B} \left[1 - \frac{(\theta - \hat{\theta})B}{1 - \theta B} \right] a_t \\
&= (\theta - \hat{\theta})(a_t + \theta a_{t-1} + \theta^2 a_{t-2} + \cdots) \\
&\quad - (\theta - \hat{\theta})^2 (a_{t-1} + 2\theta a_{t-2} + 3\theta^2 a_{t-3} + \cdots)
\end{aligned} \tag{A7.6.3}$$

On the assumption that the forecast and the estimate $\hat{\theta}$ are based on essentially nonoverlapping data, $\hat{\theta}$ and $a_t, a_{t-1}, \ldots$ are independent. Also, $\hat{\theta}$ will be approximately normally distributed about θ with variance $(1 - \theta^2)/n$, for moderate-sized samples. On these assumptions the variance of the expression in (A7.6.3) may be shown to be

$$\frac{\sigma_a^2}{n} \left(1 + \frac{3}{n} \frac{1 + \theta^2}{1 - \theta^2} \right)$$

Thus, provided that $|\theta|$ is not close to unity,

$$\text{var}[e_t(l|\hat{\theta})] \simeq \sigma_a^2 [1 + (l - 1)\lambda^2] + \frac{\sigma_a^2}{n} \tag{A7.6.4}$$

Clearly, the proportional change in the variance will be greatest for $l = 1$, when the exact forecast error variance reduces to σ_a^2. In this case, for parameter estimates based on a series of moderate length, the probability limits will be increased by a factor $(n + 1)/n$.

First-Order Autoregressive Processes. Writing the AR(1) model $\tilde{z}_t = \phi \tilde{z}_{t-1} + a_t$ at time $t + l$ and taking conditional expectations at time t, the lead l forecast, given the true value of the parameter ϕ, is

$$\hat{\tilde{z}}_t(l|\phi) = \phi \hat{\tilde{z}}_t(l - 1|\phi) = \phi^l \tilde{z}_t$$

Similarly,

$$\hat{\tilde{z}}_t(l|\hat{\phi}) = \hat{\phi} \hat{\tilde{z}}_t(l - 1|\hat{\phi}) = \hat{\phi}^l \tilde{z}_t$$

and hence

$$e_t(l|\hat{\phi}) = \tilde{z}_{t+l} - \hat{\tilde{z}}_t(l|\hat{\phi}) = e_t(l|\phi) + (\phi^l - \hat{\phi}^l)\tilde{z}_t \tag{A7.6.5}$$

Because $e_t(l|\phi) = \tilde{z}_{t+1} - \hat{\tilde{z}}_t(l|\phi) = a_{t+1} + \phi a_{t+l-1} + \cdots + \phi^{l-1} a_{t+1}$ is independent of $\hat{\phi}$ and $\tilde{z}_t$, it follows from (A7.6.5) that

$$E[e_t^2(l|\hat{\phi})] = E[e_t^2(l|\phi)] + E[\tilde{z}_t^2(\phi^l - \hat{\phi}^l)^2]$$

Again, as in the MA(1) case, the estimate $\hat{\phi}$ is assumed to be essentially independent of $\tilde{z}_t$, and for sufficiently large n, $\hat{\phi}$ will be approximately normally distributed about a mean ϕ with variance $(1-\phi^2)/n$. So using (5.4.16) and $E[\tilde{z}_t^2(\phi^l - \hat{\phi}^l)^2] \simeq E[\tilde{z}_t^2]E[(\phi^l - \hat{\phi}^l)^2]$, with $E[\tilde{z}_t^2] = \gamma_0 = \sigma_a^2/(1-\phi^2)$, *on the average*

$$\text{var}[e_t(l|\hat{\phi})] \simeq \sigma_a^2 \frac{1-\phi^{2l}}{1-\phi^2} + \sigma_a^2 \frac{E[(\phi^l - \hat{\phi}^l)^2]}{1-\phi^2} \tag{A7.6.6}$$

When $l = 1$, using $E[(\phi - \hat{\phi})^2] \simeq (1-\phi^2)/n$,

$$\begin{aligned} \text{var}[e_t(1|\hat{\phi})] &\simeq \sigma_a^2 + \frac{\sigma_a^2}{1-\phi^2}\frac{1-\phi^2}{n} \\ &= \sigma_a^2\left(1 + \frac{1}{n}\right) \end{aligned} \tag{A7.6.7}$$

For $l > 1$, we have

$$\phi^l - \hat{\phi}^l = \phi^l - \{\phi - (\phi - \hat{\phi})\}^l \simeq \phi^l - \{\phi^l - l\phi^{l-1}(\phi - \hat{\phi})\} = l\phi^{l-1}(\phi - \hat{\phi})$$

since the remaining terms involving $(\phi - \hat{\phi})^j$ for $j = 2, \ldots, l$ are of smaller order. Thus, on the average, from (A7.6.6) we obtain

$$\begin{aligned} \text{var}[e_t(l|\hat{\phi})] &\simeq \text{var}[e_t(l|\phi)] + \frac{\sigma_a^2}{1-\phi^2} E[l^2\phi^{2(l-1)}(\phi - \hat{\phi})^2] \\ &= \text{var}[e_t(l|\phi)] + \frac{l^2\phi^{2(l-1)}}{n}\sigma_a^2 \end{aligned}$$

and the discrepancy is again of order n^{-1}.

General-Order Autoregressive Processes. Related approximation results for the effect of parameter estimation errors on forecast error variances have been given by Yamamoto (1976) for the general AR(p) model. In particular, we briefly consider the approximation for one-step-ahead forecasts in the AR(p) case. Write the model at time $t+1$ as

$$\tilde{z}_{t+1} = \phi_1\tilde{z}_t + \phi_2\tilde{z}_{t-1} + \cdots + \phi_p\tilde{z}_{t+1-p} + a_{t+1} = \tilde{\mathbf{z}}_t'\boldsymbol{\phi} + a_{t+1}$$

where $\tilde{\mathbf{z}}_t' = (\tilde{z}_t, \tilde{z}_{t-1}, \ldots, \tilde{z}_{t+1-p})$ and $\boldsymbol{\phi}' = (\phi_1, \phi_2, \ldots, \phi_p)$. Then,

$$\hat{\tilde{z}}_t(1|\boldsymbol{\phi}) = \phi_1\tilde{z}_t + \phi_2\tilde{z}_{t-1} + \cdots + \phi_p\tilde{z}_{t+1-p} = \tilde{\mathbf{z}}_t'\boldsymbol{\phi}$$

and similarly, $\hat{\tilde{z}}_t(1|\hat{\boldsymbol{\phi}}) = \tilde{\mathbf{z}}_t'\hat{\boldsymbol{\phi}}$, where $\hat{\boldsymbol{\phi}}$ is the ML estimate of $\boldsymbol{\phi}$ based on n observations. Hence,

$$e_t(1|\hat{\boldsymbol{\phi}}) = e_t(1|\boldsymbol{\phi}) + \tilde{\mathbf{z}}_t'(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) \tag{A7.6.8}$$

Using similar independence properties as above, as well as $\text{cov}[\tilde{\mathbf{z}}_t] = \boldsymbol{\Gamma}_p$ and the asymptotic distribution approximation for $\hat{\boldsymbol{\phi}}$ (see, e.g., [7.2.19] and [A7.4.23]) that $\text{cov}[\hat{\boldsymbol{\phi}}] \simeq n^{-1}\sigma_a^2\boldsymbol{\Gamma}_p^{-1}$,

it follows that

$$\begin{aligned} E[e_t^2(1|\hat{\boldsymbol{\phi}})] &= E[e_t^2(1|\boldsymbol{\phi})] + E[\{\tilde{\mathbf{z}}_t'(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})\}^2] \\ &= \sigma_a^2 + \text{tr}\{E[\tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t']E[(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})']\} \\ &= \sigma_a^2 + \text{tr}\{\boldsymbol{\Gamma}_p n^{-1}\sigma_a^2\boldsymbol{\Gamma}_p^{-1}\} \end{aligned}$$

Thus, the approximation for one-step-ahead forecast error variance,

$$\text{var}[e_t(1|\hat{\boldsymbol{\phi}})] \simeq \sigma_a^2\left(1 + \frac{p}{n}\right) \tag{A7.6.9}$$

is readily obtained for the AR model of order p.

APPENDIX A7.7 SPECIAL NOTE ON ESTIMATION OF MOVING AVERAGE PARAMETERS

If the least-squares iteration that involves moving average parameters is allowed to stray outside the invertibility region, parameter values can readily be found that apparently provide sums of squares smaller than the true minimum. However, these do not provide appropriate estimates and are quite meaningless. To illustrate, suppose that a series has been generated by the first-order moving average model $w_t = (1 - \theta B)a_t$ with $-1 < \theta < 1$. Then, the series could equally well have been generated by the corresponding backward process $w_t = (1 - \theta F)e_t$ with $\sigma_e^2 = \sigma_a^2$. Now, the latter process can also be written as $w_t = (1 - \theta^{-1}B)\alpha_t$, where now θ^{-1} is *outside* the invertibility region. However, in this representation $\sigma_\alpha^2 = \sigma_a^2\theta^2$ and is itself a function of θ. Therefore, a valid estimate of θ^{-1} will not be provided by minimizing $\sum_t \alpha_t^2 = \theta^2 \sum_t a_t^2$. Indeed, this has its minimum at $\theta^{-1} = \infty$.

The difficulty may be avoided:

1. By using as starting values rough preliminary estimates within the invertibility region obtained at the identification stage.
2. By checking that all moving average estimates, obtained after convergence has apparently occurred, lie within the invertibility region.

It is also possible to write least-squares programs such that estimates are constrained to lie within the invertibility region, and to check that moving average estimates lie within the invertibility region after each step of the iterative least-squares estimation procedure.

EXERCISES

7.1. The following table shows calculations for an (unrealistically short) series z_t for which the (0, 1, 1) model $w_t = \nabla z_t = (1 - \theta B)a_t$ is being considered with $\theta = -0.5$ and with an unknown starting value a_0.

t	z_t	$w_t = \nabla z_t$	$a_t = w_t - 0.5a_{t-1}$
0	40		a_0
1	42	2	$2 - 0.50a_0$
2	47	5	$4 + 0.25a_0$
3	47	0	$-2 - 0.13a_0$
4	52	5	$6 + 0.06a_0$
5	51	-1	$-4 - 0.03a_0$
6	57	6	$8 + 0.02a_0$
7	59	2	$-2 - 0.01a_0$

(a) Confirm the entries in the table.

(b) Show that the conditional sum of squares is

$$\sum_{t=1}^{7}(a_t| - 0.5, a_0 = 0)^2 = S_*(-0.5|0) = 144.00$$

7.2. Using the data in Exercise 7.1:

(a) Show (using least-squares) that the value $\hat{a}_0$ of a_0 that minimizes $S_*(-0.5|0)$ is

$$\hat{a}_0 = \frac{(2)(0.50) + (4)(-0.25) + \cdots + (-2)(0.0078)}{1^2 + 0.5^2 + \cdots + 0.0078^2} = \frac{-\sum_{t=0}^{n}\theta^t a_t^0}{\sum_{t=0}^{n}\theta^{2t}}$$

where $a_t^0 = (a_t|\theta, a_0 = 0)$ are the conditional values. Compare this expression for $\hat{a}_0$ with that for the exact back-forecast $[a_0]$ in the MA(1) model, where the expression for $[a_0]$ is given preceding the equation (A7.3.9) in Appendix A7.3, and verify that the two expressions are identical.

(b) By first writing this model in the backward form $w_t = (1 - \theta F)e_t$ and recursively computing the e's, show that the value of a_0 obtained in (a) is the same as that obtained by the back-forecasting method.

7.3. Using the value of $\hat{a}_0$ calculated in Exercise 7.2:

(a) Show that the unconditional sum of squares $S(-0.5)$ is 143.4.

(b) Show that for the (0, 1, 1) model, for large n,

$$S(\theta) = S_*(\theta|0) - \frac{\hat{a}_0^2}{1 - \theta^2}$$

7.4. For the process $w_t = \mu_w + (1 - \theta B)a_t$ show that for long series the variance–covariance matrix of the maximum likelihood estimates $\hat{\mu}_w, \hat{\theta}$ is approximately

$$n^{-1}\begin{bmatrix} (1-\theta)^2\sigma_a^2 & 0 \\ 0 & 1 - \theta^2 \end{bmatrix}$$

7.5. **(a)** Problems were experienced in obtaining a satisfactory fit to a series, the last 16 values of which were recorded as follows:

129, 135, 130, 130, 127, 126, 131, 152,
123, 124, 131, 132, 129, 127, 126, 124

Plot the series and suggest where the difficulty might lie.

(b) In fitting a model of the form $(1 - \phi_1 B - \phi_2 B^2) z_t = (1 - \theta B) a_t$ to a set of data, convergence was slow and the coefficient estimates in successive iterations oscillated wildly. Final estimates having large standard errors were obtained as follows: $\hat{\phi}_1 = 1.19$, $\hat{\phi}_2 = -0.34$, $\hat{\theta} = 0.52$. Can you suggest an explanation for the unstable behavior of the model? Why should preliminary identification have eliminated the problem?

(c) In fitting the model $\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2) a_t$ convergence was not obtained. The last iteration yielded the values $\hat{\theta}_1 = 1.81$, $\hat{\theta}_2 = 0.52$. Can you explain the difficulty?

7.6. For the ARIMA(1, 1, 1) model $(1 - \phi B) w_t = (1 - \theta B) a_t$, where $w_t = \nabla z_t$:

(a) Write down the linearized form of the model.

(b) Set out how you would start off the calculation of the conditional nonlinear least-squares algorithm with start values $\phi = 0.5$ and $\theta = 0.4$ for a series whose first nine values are shown below.

t	z_t	t	z_t
0	149	5	150
1	145	6	147
2	152	7	142
3	144	8	146
4	150		

7.7. (a) Show that the second-order autoregressive model $\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + a_t$ may be written in orthogonal form as

$$\tilde{z}_t = \frac{\phi_1}{1 - \phi_2} \tilde{z}_{t-1} + \phi_2 \left(\tilde{z}_{t-2} - \frac{\phi_1}{1 - \phi_2} \tilde{z}_{t-1} \right) + a_t$$

suggesting that the approximate estimates

$$r_1 \quad \text{of} \quad \frac{\phi_1}{1 - \phi_2} \quad \text{and} \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2} \quad \text{of} \quad \phi_2$$

are uncorrelated for long series.

(b) Starting from the variance–covariance matrix of $\hat{\phi}_1$ and $\hat{\phi}_2$ or otherwise, show that the variance–covariance matrix of r_1 and $\hat{\phi}_2$ for long series is given approximately by

$$n^{-1} \begin{bmatrix} (1 - \phi_2^2)(1 - \rho_1^2) & 0 \\ 0 & 1 - \phi_2^2 \end{bmatrix}$$

7.8. The preliminary model identification performed in Chapter 6 suggested that either an ARIMA(1, 1, 0) or an ARIMA(0, 2, 2) model might be appropriate for the chemical process temperature readings in Series C. The series is available for download from http://pages.stat.wisc.edu/ reinsel/bjr-data/.

(a) Estimate the parameters of the ARIMA(1, 1, 0) for this series using R.

(b) Estimate the parameters of the ARIMA(0, 2, 2) model and compare the results with those in part (a).

7.9. Repeat the analysis in Exercise 7.8 by fitting (a) an AR(1) and (b) an ARMA(0, 1, 1) model to the chemical process viscosity readings in Series D.

7.10. Daily air quality measurements in New York, from May to September 1973, are available in a file called 'airquality' in the R datasets package. The file provides data on four air quality variables: mean ozone levels at Roosevelt Island, solar radiation at Central Park, maximum daily temperature at La Guardia Airport, and average wind speeds at La Guardia Airport.

(a) Identify suitable models for the daily temperature and wind speed series.

(b) Estimate the parameters of selected models and comment.

7.11. Consider the solar radiation series that is part of the New York airquality data file described in Problem 7.10. This series has a few missing values.

(a) Impute suitable estimates of the missing values. (Note: A formal procedure for estimating missing values is described in Chapter 13, but is not needed here).

(b) Identify a model for the resulting series.

(c) Estimate the parameters of selected model and comment.

7.12. Refer to the annual river flow measurements in the time series 'Nile' analyzed in Exercise 6.7. Estimate the parameters of the model or models identified for this time series and comment.

8

MODEL DIAGNOSTIC CHECKING

The model having been identified and the parameters estimated, *diagnostic checks* are then applied to the fitted model. One useful method of checking a model is to *overfit*, that is, to estimate the parameters in a model somewhat more general than that which we believe to be true. This method assumes that we can guess the direction in which the model is likely to be inadequate. Therefore, it is necessary to supplement this approach by less specific checks applied to the residuals from the fitted model. These allow the data themselves to suggest modifications to the model. In this chapter, we describe two such checks that employ (1) the autocorrelation function of the residuals and (2) the cumulative periodogram of the residuals. Some alternative diagnostic procedures are also discussed. Numerical examples are included to demonstrate the results.

8.1 CHECKING THE STOCHASTIC MODEL

8.1.1 General Philosophy

Suppose that using a particular time series, the model has been identified and the parameters estimated using the methods described in Chapters 6 and 7. The question remains of deciding whether this model is adequate. If there is evidence of serious inadequacy, we need to know how the model should be modified in the next iterative cycle. What we are doing is described only partially by the words ''testing goodness of fit.'' We need to discover *in what way* a model is inadequate, so as to suggest appropriate modification. To illustrate, by reference to familiar procedures outside time series analysis, the scrutiny of residuals for the analysis of variance, described by Anscombe (1961) and Anscombe and Tukey (1963), and the

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

criticism of factorial experiments, leading to normal plotting and other methods, described by Daniel (1959), would be called *diagnostic checks*.

All models are approximations and no model form can ever represent the truth absolutely. Given sufficient data, statistical tests can discredit models that could nevertheless be entirely adequate for the purpose at hand. Alternatively, tests can fail to indicate serious departures from assumptions because of small sample sizes or because these tests are insensitive to the types of discrepancies that occur. The best policy is to devise the most sensitive statistical procedures possible but be prepared to employ models that exhibit slight lack of fit. If diagnostic checks, which have been thoughtfully devised, are applied to a model fitted to a reasonably large body of data and fail to show serious discrepancies, then we should feel comfortable using that model.

8.1.2 Overfitting

One technique that can be used for diagnostic checking is *overfitting*. Having identified what is believed to be a correct model, we actually fit a more elaborate one. This puts the identified model in jeopardy because the more elaborate model contains additional parameters covering feared directions of discrepancy. Careful thought should be given to the question of how the model should be augmented. In particular, in accordance with the discussion on model redundancy in Section 7.3.5, it would not make sense to add factors *simultaneously* to both sides of the ARMA model. Moreover, if the analysis fails to show that the additions are needed, we have, of course, not proved that our model is correct. A model is only capable of being ''proved'' in the biblical sense of being put to the test. As was recommended by Saint Paul in his first epistle to the Thessalonians, what we can do is to ''Prove all things; hold fast to that which is good.''

Example of Overfitting. As an example, we consider again some IBM stock price data. For this analysis, data were employed that are listed as Series B′ in the Collection of Time Series in Part Five of this book. This series consists of IBM stock prices for the period[1] June 29, 1959–June 30, 1960. The (0, 1, 1) model

$$\nabla z_t = (1 - \theta B)a_t$$

with $\hat{\lambda}_0 = 1 - \hat{\theta} = 0.90$, was identified and fitted to the 255 available observations.

The (0, 1, 1) model can equally well be expressed in the form

$$\nabla z_t = \lambda_0 a_{t-1} + \nabla a_t$$

The extended model that was considered in the overfitting procedure was the (0, 3, 3) process

$$\nabla^3 z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$$

or using (4.3.21), in the form

$$\nabla^3 z_t = (\lambda_0 \nabla^2 + \lambda_1 \nabla + \lambda_2)a_{t-1} + \nabla^3 a_t$$

[1]The IBM stock data previously considered, referred to as Series B, cover a different period, May 17, 1961–November 2, 1962.

While this model may seem overly elaborate, the immediate motivation for extending the model in this particular way was to test a suggestion made by Brown (1962) that the series should be forecasted by an adaptive *quadratic* forecast function. Now, it was shown in Chapter 5 that an IMA(0, q, q) process has for its optimal forecasting function an adaptive polynomial of degree $q - 1$. Thus, for the extended (0, 3, 3) model above, the optimal lead l forecast function is the quadratic polynomial in l:

$$\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} l + b_2^{(t)} l^2$$

where the coefficients $b_0^{(t)}$, $b_1^{(t)}$, and $b_2^{(t)}$ are adjusted as each new piece of data becomes available.

By comparison, the model we have identified is an IMA(0, 1, 1) process, which yields a forecast function

$$\hat{z}_t(l) = b_0^{(t)} \tag{8.1.1}$$

This is a ''polynomial in l'' of degree zero. Hence, the model implies that the forecast $\hat{z}_t(l)$ is independent of l, that is, the forecast at any particular time t is the same for one step ahead, two steps ahead, and so on. In other words, the series contains information only on the future *level* of the series, and nothing about slope or curvature. At first sight, this is somewhat surprising because, using hindsight, quite definite linear and curvilinear trends appear to be present in the series. Therefore, it is worthwhile to check whether nonzero values of λ_1 and λ_2, which would produce predictable trends, actually occur. Sum-of-squares grids for $S(\lambda_1, \lambda_2 | \lambda_0)$ similar to those shown in Figure 7.2 were produced for $\lambda_0 = 0.7$, 0.9, and 1.1, which showed a minimum close to $\hat{\lambda}_0 = 0.9$, $\hat{\lambda}_1 = 0$, and $\hat{\lambda}_2 = 0$. It was clear that values of $\lambda_1 > 0$ and $\lambda_2 > 0$ lead to higher sum of squares, and do not support augmenting the identified IMA(0, 1, 1) model in these directions. This implies, in particular, that a quadratic forecast function would give worse instead of better forecasts than those obtained from (8.1.1), as was indeed shown to be the case in Section A5.3.3.

Computations in R. Estimation of the parameters in the more elaborate IMA(0, 3, 3) models for the IBM series using R also shows that the model can be simplified. The relevant commands along with a partial model output are provided below:

```
>library(astsa}
>ibm2=read.table("ibm2.txt",header=TRUE}
>ibm.ts=ts(ibm2)
>sarima(ibm.ts,0,3,3)

 Coefficients:
          ma1     ma2      ma3
      -2.0215  1.0686  -0.0469
s.e.   0.0705  0.1370   0.0692  sigma^2 estimated as 25.5

> polyroot(c(1,-2.0215,1.0686,-0.0469))
1.013484+0.005832i 1.013484-0.005832i 20.757680+0.000000i

>sarima(ibm.ts,0,1,1)

 Coefficients:
```

```
          ma1  constant
      -0.0848    0.3028
s.e.   0.0634    0.2878         sigma^2 estimated as 25.1
```

We note that the parameter estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ in the IMA(0, 3, 3) model are highly significant. However, the large estimates are introduced as compensation for overdifferencing by setting $d = 3$ in this model. This is confirmed by finding the roots of the moving average polynomial using the command polyroot() in R. The results, which are included above, show that two of the roots are very close to one. Hence, cancellation is possible, reducing the IMA(0, 3, 3) model to a IMA(0, 1, 1) model. The IMA(0, 1, 1) model also provides a slightly better fit to the data as can be seen from the smaller value of $\hat{\sigma}^2$ in the R output for this model.

8.2 DIAGNOSTIC CHECKS APPLIED TO RESIDUALS

The method of overfitting, by extending the model in a particular direction, assumes that we know what kind of discrepancies are to be feared. Procedures less dependent upon such knowledge are based on the analysis of *residuals*. It cannot be too strongly emphasized that *visual inspection of a plot of the residuals themselves* is an indispensable first step in the checking process.

8.2.1 Autocorrelation Check

Suppose that a model $\phi(B)\tilde{w}_t = \theta(B)a_t$ has been fitted to the observed time series with ML estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ obtained for the parameters. The quantities

$$\hat{a}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)\tilde{w}_t \tag{8.2.1}$$

are then referred to as the *residuals*. The residuals are computed recursively from $\hat{\theta}(B)\hat{a}_t = \hat{\phi}(B)\tilde{w}_t$ as

$$\hat{a}_t = \tilde{w}_t - \sum_{j=1}^{p} \hat{\phi}_j \tilde{w}_{t-j} + \sum_{j=1}^{q} \hat{\theta}_j \hat{a}_{t-j} \qquad t = 1, 2, \ldots, n$$

using either zero initial values (conditional method) or back-forecasted initial values (exact method) for the initial $\hat{a}_t$'s and $\tilde{w}_t$'s. Now, it is possible to show that, if the model is adequate,

$$\hat{a}_t = a_t + O\left(\frac{1}{\sqrt{n}}\right)$$

As the series length increases, the $\hat{a}_t$'s become close to the white noise a_t's. Therefore, one might expect that study of the $\hat{a}_t$'s could indicate the existence and nature of model inadequacy. In particular, recognizable patterns in the estimated autocorrelation function of the $\hat{a}_t$'s could point to appropriate modifications in the model. This point is discussed further in Section 8.3.

Now, suppose that the form of the model was correct and that we *knew* the true parameter values $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Then, using (2.1.13) and a result of Anderson (1942), the estimated

autocorrelations $r_k(a)$, of the a_t's, would be uncorrelated and distributed approximately normally about zero with variance n^{-1}, and hence with a standard error of $n^{-1/2}$. We could use these facts to assess approximately the statistical significance of apparent departures of these autocorrelations from zero.

Now, in practice, we do not know the *true* parameter values. We have only the estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$, from which, using (8.2.1), we can calculate not the a_t's but the $\hat{a}_t$'s. The autocorrelations $r_k(\hat{a})$ of the $\hat{a}_t$'s can yield valuable evidence concerning lack of fit and the possible nature of model inadequacy. However, it was pointed out by Durbin (1970) that it might be dangerous to assess the statistical significance of apparent discrepancies of these autocorrelations $r_k(\hat{a})$ from their theoretical zero values on the basis of a standard error $n^{-1/2}$, appropriate to the $r_k(a)$'s. Durbin was able to show, for example, that for the AR(1) process with parameter ϕ, the variance of $r_1(\hat{a})$ is $\phi^2 n^{-1}$, which can be substantially *smaller* than n^{-1}. The large-sample variances and covariances for all the autocorrelations of the $\hat{a}_t$'s from any ARMA process were subsequently derived by Box and Pierce (1970). They showed that while in all cases, a reduction in variance can occur for low lags, and that at these low lags the $r_k(\hat{a})$'s can be highly correlated, these effects usually disappear rather quickly at high lags. Thus, the use of $n^{-1/2}$ as the standard error for $r_k(\hat{a})$ would underestimate the statistical significance of apparent departures from zero of the autocorrelations at low lags but could usually be employed for moderate or high lags.

For illustration, the large-sample one- and two-standard-error limits of the residual autocorrelations $r_k(\hat{a})$'s, for two AR(1) processes and two AR(2) processes, are shown in Figure 8.1. These also supply the corresponding approximate standard errors for moving average processes with the same parameters as indicated in the figure. It is evident that, except at moderately high lags, $n^{-1/2}$ provides an upper bound for the standard errors of the $r_k(\hat{a})$'s rather than the standard errors themselves. If for low lags we use the standard

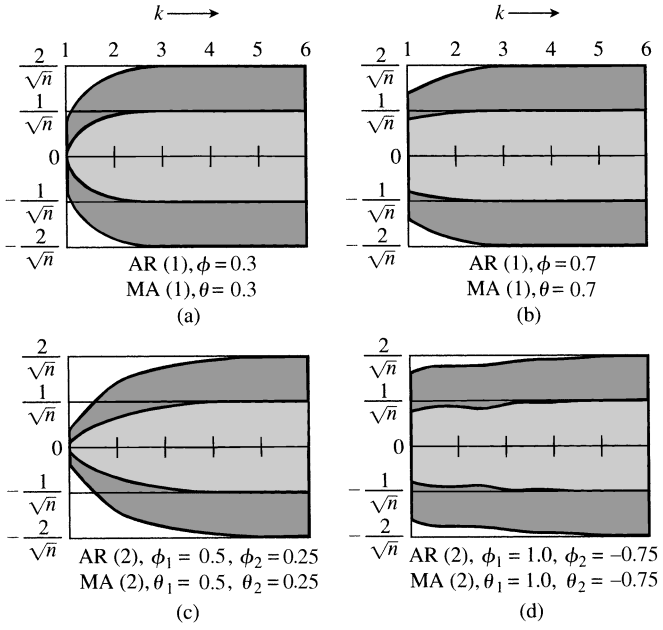

FIGURE 8.1 Standard-error limits for residual autocorrelations $r_k(\hat{a})$.

error $n^{-1/2}$ for the $r_k(\hat{a})$'s, we may seriously *underestimate* the significance of apparent discrepancies.

8.2.2 Portmanteau Lack-of-Fit Test

In addition to considering the $r_k(\hat{a})$'s individually, an indication is often needed of whether, say, the first 10–20 autocorrelations of the $\hat{a}_t$'s *taken as a whole* indicate inadequacy of the model. Suppose that we have the first K autocorrelations[2] $r_k(\hat{a})$ $(k = 1, 2, \ldots, K)$ from any ARIMA(p, d, q) model, then it is possible to show (Box and Pierce, 1970) that if the fitted model is appropriate,

$$Q = n \sum_{k=1}^{K} r_k^2(\hat{a}) \tag{8.2.2}$$

is approximately distributed as $\chi^2(K - p - q)$, where $n = N - d$ is the number of w's used to fit the model. On the other hand, if the model is inappropriate, the average values of Q will be inflated. Therefore, an approximate ''portmanteau'' test of the hypothesis of model adequacy, designed to take account of the difficulties discussed above, may be made by referring an observed value of Q to the percentage points of this χ^2 distribution.

However, Ljung and Box (1978) later showed that, for sample sizes common in practice, the chi-squared distribution may not provide an adequate approximation to the distribution of the statistic Q under the null hypothesis, with the values of Q tending to be somewhat smaller than what is expected under the chi-squared distribution. Empirical evidence to support this was also presented by Davies et al. (1977). Ljung and Box (1978) proposed a modified form of the statistic,

$$\tilde{Q} = n(n + 2) \sum_{k=1}^{K} (n - k)^{-1} r_k^2(\hat{a}) \tag{8.2.3}$$

such that the modified statistic has, approximately, the mean $E[\tilde{Q}] \approx K - p - q$ of the $\chi^2(K - p - q)$ distribution. The motivation for (8.2.3) is that a more accurate value for the variance of $r_k(a)$ from a white noise series is $(n - k)/n(n + 2)$, rather than $1/n$ used in (8.2.2). This modified form of the portmanteau test statistic has been recommended for use as having a null distribution that is much closer to the $\chi^2(K - p - q)$ distribution for typical sample sizes n. Because of its computationally convenient form, this statistics has been implemented in many software packages and has become widely used in applied work. We emphasize, however, that this statistic should not be used as a substitute for careful examination of the residuals and their individual autocorrelation coefficients, and for other diagnostic checks on the fitted model.

Remark. Diagnostic checks based on the residuals and their autocorrelation coefficients are conveniently performed using R. Having fitted a model `m1` to the observed series, the command `tsdiag(m1$residuals, gof.lag=20)` provides a plot of the standardized residuals, a plot of the first 20 residual autocorrelation coefficients, and a plot of the p-values for the

[2]It is assumed here that K is taken sufficiently large so that the weights ψ_j in the model, written in the form $\tilde{w}_t = \phi^{-1}(B)\theta(B)a_t = \psi(B)a_t$ will be negligibly small after $j = K$.

portmanteau statistic $\tilde{Q}$ for increasing values of K. However, while these diagnostics are useful, it appears that the command tsdiag(), at present, determines p-values for $\tilde{Q}$ using a chi-square distribution with K rather than $K - p - q$ degrees of freedom. An alternative is to use diagnostic tools in the R package astsa, where this problem does not appear. An illustration of the use of this package is provided below.

An Empirical Example. In Chapter 7, we examined two potential models for a time series of chemical temperature readings referred to as Series C. The two models were (1) the IMA(0, 2, 2) model $\nabla^2 z_t = (1 - 0.13B - 0.12B^2)a_t$ and (2) the ARIMA(1, 1, 0) model $(1 - 0.82B)\nabla z_t = a_t$. It was decided that the second model gave a preferable representation of the series. Model diagnostics for the IMA(0, 2, 2) model generated using R are provided in Figure 8.2. These include graphs of the standardized residuals, the residual autocorrelation coefficients $r(\hat{a}_k)$, for lags $k = 1, \ldots, 25$, a normal Q–Q plot of the standardized residuals, and a plot of the p-values for the portmanteau statistic $\tilde{Q}$ in (8.2.3) determined for increasing values of K. The graph of the standardized residuals reveals some large residuals around $t = 60$, but apart from that there are no issues. The Q–Q plot confirms the presence of three large residuals but indicates that the normal approximation is adequate otherwise.

Approximate two-standard-error upper bounds on the residual autocorrelation coefficients are included in the graph of the autocorrelation function. Since there are $n = 224$ observations after differencing the series, the approximate upper bound for the standard

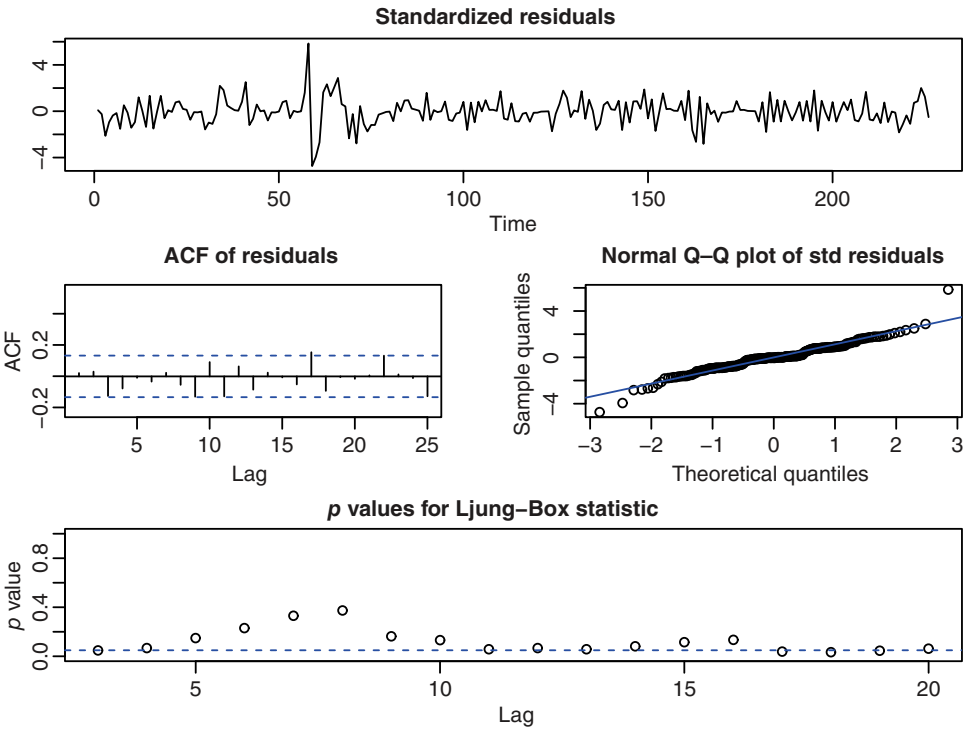

FIGURE 8.2 Model diagnostics for the ARIMA(0, 2, 2) model fitted to the temperature readings in Series C.

error of a single autocorrelation is $1/\sqrt{224} \approx 0.07$. While most of the individual autocorrelations fall within the two-standard-error bounds, several values including $r_3(\hat{a})$, $r_9(\hat{a})$, $r_{11}(\hat{a})$, $r_{17}(\hat{a})$, $r_{22}(\hat{a})$, and $r_{25}(\hat{a})$ are close to these bounds. Of course, occasional large deviations occur even in random series, but taking these results as a whole, there is a suspicion of some lack of fit. This is confirmed by examining the p-values of the portmanteau statistic shown in the bottom graph of Figure 8.2. We note that most of the p-values are at or near the 5% level indicating some lack of fit. This is especially the case for the larger values of K, where the chi-squared distribution is expected to provide a valid approximation.

Model diagnostics for the ARIMA(1, 1, 0) model $(1 - 0.82B)\nabla z_t = a_t$ fitted to the same time series are displayed in Figure 8.3. The graph of the residual autocorrelation function shows fewer large values for this model. This is also reflected in the p-values of the portmanteau statistic shown at the bottom of the graph. These diagnostic checks show a clear improvement over the IMA(0, 2, 2) model examined in Figure 8.2. The graph of the standardized residuals and the normal Q–Q plot reveal that outliers are still present, however. Methods for outlier detection and adjustments will be discussed in Section 13.2, where the ARIMA(1, 1, 0) model for Series C is refitted allowing the outliers at $t = 58$, 59, and 60. Allowing these outliers in the parameter estimation changes the estimate $\hat{\phi}$ only slightly from 0.82 to 0.85. However, a larger change occurs in the estimate of the residual variance, which is reduced by about 26% when the outliers are accounted for in the model.

Before proceeding, we note that Figures 8.2 and 8.3 can be reproduced in R using the following commands:

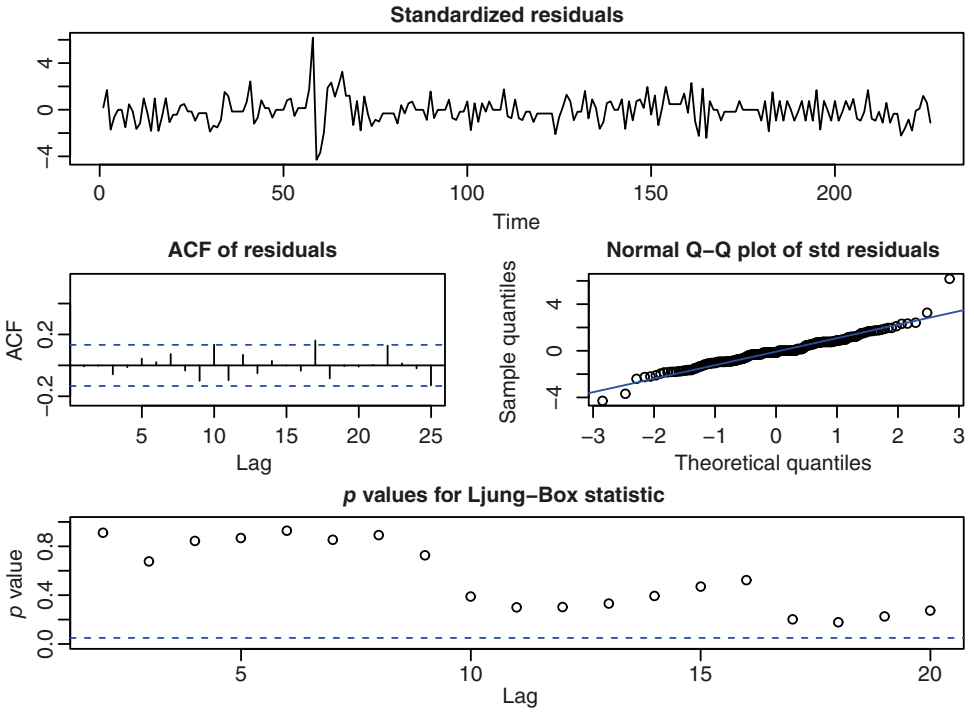

FIGURE 8.3 Model diagnostics for the ARIMA(1, 1, 0) model fitted to the temperature readings in Series C.

```
>library(astsa)
>seriesC=read.table("seriesC.txt",header=T)
>sarima(seriesC,0,2,2,no.constant=TRUE) % Figure 8.2
>sarima(seriesC,1,1,0,no.constant=TRUE) % Figure 8.3
```

Portmanteau Tests for Series A–F. Table 8.1 summarizes the values of the criterion $\tilde{Q}$ in (8.2.3) based on $K = 25$ residual autocorrelations for the models fitted to Series A–F in Table 7.11. However, in regards to the choice of K, a somewhat smaller value would be recommended for use in practice, especially for shorter series such as Series E and F, since the asymptotic theory involved in the distribution of the statistic $\tilde{Q}$ relies on K growing (but only slowly, such that $K/n \to 0$) as the series length n increases. In addition, as noted by Ljung (1986), smaller values of K also have advantages in terms of increased power. This is particularly true for nonseasonal series, where the lack of fit is expected to be most evident in residual autocorrelations at the first few lags.

Inspection of Table 8.1 shows that only two suspiciously large values of $\tilde{Q}$ occur. One is the value $\tilde{Q} = 36.2$ obtained after fitting the IMA(0, 2, 2) model to Series C, which we have discussed already. The other is the value $\tilde{Q} = 38.8$ obtained after fitting an IMA(0, 1, 1) model to Series B. This suggests some model inadequacy since the 5 and 2.5% points for χ^2 with 24 degrees of freedom are 36.4 and 39.3, respectively. The nature of possible model inadequacy for Series B will be examined further in Section 8.2.3.

Other Portmanteau Statistics to Test Model Adequacy. Instead of a portmanteau statistic based on residual autocorrelations, as in (8.2.3), one could alternatively consider a test for model adequacy based on residual *partial* autocorrelations. If the model fitted is adequate, the associated error process a_t is white noise and one should expect the residual partial autocorrelation at any lag k, which we denote as $\hat{\phi}_{kk}(\hat{a})$, not to be significantly different from zero. Therefore, a test for model adequacy can be based on the statistic

$$Q^* = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}\hat{\phi}_{kk}^2(\hat{a}) \tag{8.2.4}$$

TABLE 8.1 Summary of Results of Portmanteau Test Applied to Residuals of Various Models Fitted to Series A–F

Series	$n = N - d$	Fitted Model	$\tilde{Q}$	Degrees of Freedom
A	197	$z_t - 0.92z_{t-1} = 1.45 + a_t - 0.58a_{t-1}$	28.4	23
	196	$\nabla z_t = a_t - 0.70a_{t-1}$	31.9	24
B	368	$\nabla z_t = a_t + 0.09a_{t-1}$	38.8	24
C	225	$\nabla z_t - 0.82\nabla z_{t-1} = a_t$	31.3	24
	224	$\nabla^2 z_t = a_t - 0.13a_{t-1} - 0.12a_{t-2}$	36.2	23
D	310	$z_t - 0.87z_{t-1} = 1.17 + a_t$	11.5	24
	309	$\nabla z_t = a_t - 0.06a_{t-1}$	18.8	24
E	100	$z_t - 1.42z_{t-1} + 0.73z_{t-2} = 14.35 + a_t$	26.8	23
	100	$z_t - 1.57z_{t-1} + 1.02z_{t-2} - 0.21z_{t-3} = 11.31 + a_t$	20.0	22
F	70	$z_t + 0.34z_{t-1} - 0.19z_{t-2} = 58.87 + a_t$	14.7	23

Under the hypothesis of model adequacy, Monti (1994) argued that the statistic Q^* in (8.2.4) is asymptotically distributed as $\chi^2(K-p-q)$, analogous to the asymptotic distribution of the statistic $\tilde{Q}$ in (8.2.3). Hence, a test of model adequacy can be based on referring the value of Q^* to the upper critical value determined from this distribution. The test based on Q^* has been found to be typically at least as powerful as $\tilde{Q}$ in detecting departures from model adequacy, and it seems to be particularly sensitive when the alternative model includes a higher order moving average term. In practice, since residual partial autocorrelations are routinely available, we could consider using both the statistic $\tilde{Q}$ in (8.2.3) and Q^* in (8.2.4) simultaneously in standard model checking procedures.

Another portmanteau goodness-of-fit test statistic based on a general measure of multivariate dependence was proposed by Peña and Rodríguez (2002). Denote the correlation matrix up to order (lag) K of the residuals $\hat{a}_t$ from the fitted ARIMA(p,d,q) model by

$$\hat{\mathbf{P}}_K(\hat{a}) = \begin{bmatrix} 1 & r_1(\hat{a}) & r_2(\hat{a}) & \dots & r_K(\hat{a}) \\ r_1(\hat{a}) & 1 & r_1(\hat{a}) & \dots & r_{K-1}(\hat{a}) \\ r_2(\hat{a}) & r_1(\hat{a}) & 1 & \dots & r_{K-2}(\hat{a}) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ r_K(\hat{a}) & r_{K-1}(\hat{a}) & r_{K-2}(\hat{a}) & \dots & 1 \end{bmatrix}$$

The proposed statistic is based on the determinant of this correlation matrix, a general measure of dependence in multivariate analysis, and is given by

$$\hat{D}_K = n(1 - |\hat{\mathbf{P}}_K(\hat{a})|^{1/K}) \tag{8.2.5}$$

An alternate interpretation for the statistic is obtained from the following relation given by Peña and Rodríquez (2002)

$$|\hat{\mathbf{P}}_K(\hat{a})|^{1/K} = \prod_{k=1}^{K} [1 - \hat{\phi}_{kk}^2(\hat{a})]^{(K+1-k)/K}$$

where the $\hat{\phi}_{kk}(\hat{a})$ are the residual partial autocorrelations as in (8.2.4). This expression shows that $|\hat{\mathbf{P}}_K(\hat{a})|^{1/K}$ is also a weighted function of the first K partial autocorrelations of the residuals. However, in comparison to the statistics (8.2.3) and (8.2.4), relatively more weight is given to the lower lag residual correlations in the statistic (8.2.5). The asymptotic distribution of $\hat{D}_K$ is shown to be a linear combination of K-independent $\chi^2(1)$ random variates, which can be approximated by a gamma distribution (see Peña and Rodríguez, 2002). The authors also proposed and recommended a modification of the statistic $\hat{D}_K$, here denoted as $\tilde{D}_K$, in which the residual autocorrelations $r_k(\hat{a})$ used to form $\hat{\mathbf{P}}_K(\hat{a})$ are replaced by the modified values $\sqrt{(n+2)/(n-k)}r_k(\hat{a})$, similar to the modifications used in the $\tilde{Q}$ and Q^* statistics. Simulation evidence indicates that the statistic $\tilde{D}_K$ may provide considerable increase in power over the statistics $\tilde{Q}$ and Q^* in many cases, due to its greater sensitivity to the lower lag residual correlations. Application of this procedure to detection of several types of nonlinearity, by using sample autocorrelations of squared residuals $\hat{a}_t^2$, was also explored in Peña and Rodríguez (2002). (For discussion of nonlinearities, see Sections 10.2 and 10.3).

Peña and Rodríguez (2006) proposed a modification of their earlier test that has the same asymptotic distribution as $\hat{D}_K$ but better performance in finite sam-

ples. The modified test statistics has the form $D_K^* = -n\sum_{k=1}^{K} w_k \ln[1 - \hat{\phi}_{kk}^2(\hat{a})]$, where $w_k = (K + 1 - k)/(K + 1)$. The statistic is thus proportional to a weighted average of the squared partial autocorrelation coefficients with larger weights given to low-order coefficients and smaller weights to high-order coefficients. The authors considered two approximations to the asymptotic distribution of this statistic, and demonstrated using simulation that the test performs well. Several other authors have extended the work of Peña and Rodríguez (2002) and proposed portmanteau statistics that are asymptotically similar to their statistics; for a discussion and references, see Fisher and Gallagher (2012). See also Li (2004) for a more detailed discussion of diagnostic testing.

8.2.3 Model Inadequacy Arising from Changes in Parameter Values

Another form of model inadequacy occurs when the *form* of the model remains the same but the parameters change over a prolonged period of time. In fact, it appears that this can explain the possible inadequacy of the (0, 1, 1) model fitted to the IBM data.

Table 8.2 shows the results obtained by fitting (0, 1, 1) models separately to the first and second halves of Series B as well as to the complete series. Denoting the estimates of $\lambda = 1 - \theta$ obtained from the two halves by $\hat{\lambda}_1$ and $\hat{\lambda}_2$, we find that the standard error of $\hat{\lambda}_1 - \hat{\lambda}_2$ is $\sqrt{(0.070)^2 + (0.074)^2} = 0.102$. Since the difference $\hat{\lambda}_1 - \hat{\lambda}_2 = 0.26$ is 2.6 times its standard error, it is likely that a real change in λ has occurred. Inspection of the $\tilde{Q}$ values suggests that the (0, 1, 1) model, with parameters appropriately modified for different time periods, might explain the series more exactly. The estimation results for the residual variances $\hat{\sigma}_a^2$ also strongly indicate that a real *change in variability* has occurred between the two halves of the series.

This is confirmed by Figure 8.4 that shows the standardized residuals and other model diagnostics for the IMA(0, 1, 1) model fitted to Series B. An increase in the standardized residuals around time $t = 236$ indicates a change in the characteristics of the series around that time. In fact, fitting the IMA(0, 1, 1) model separately to the first 235 observations and to the remaining 134 observations yields the estimates $\hat{\theta}_1 = -0.26, \hat{\sigma}_{a_1}^2 = 24.55$, and $\hat{\theta}_2 = -0.02, \hat{\sigma}_{a_2}^2 = 99.49$, respectively. Hence, a substantial increase in variability during the latter portion of the series is clearly indicated. Additional approaches to explain and account for inadequacy in the overall IMA(0, 1, 1) model for Series B, which include allowance for conditional heteroscedasticity in the noise, nonlinearity, and mixture transition distributions, have been discussed by Tong (1990) and Le et al. (1996), among others. Some of these modeling approaches will be surveyed in general in Chapter 10.

TABLE 8.2 Comparison of IMA(0, 1, 1) Models Fitted to First and Second Halves of Series B

	n	$\hat{\theta}$	$\hat{\lambda} = 1 - \hat{\theta}$	$\hat{\sigma}(\hat{\lambda}) = [\frac{\hat{\lambda}(2-\hat{\lambda})}{n}]^{1/2}$	Residual Variance $\hat{\sigma}_a^2$	$\tilde{Q}$	Degrees of Freedom
First half	184	−0.29	1.29	±0.070	26.3	24.6	24
Second half	183	−0.03	1.03	±0.074	77.3	37.1	24
Complete	368	−0.09	1.09	±0.052	52.2	38.8	24

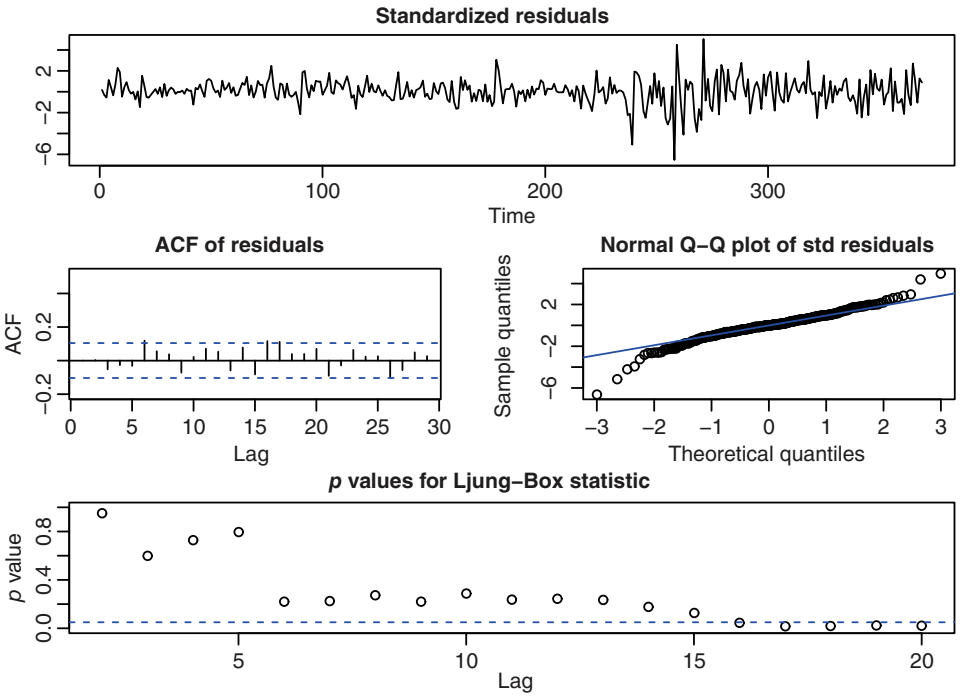

FIGURE 8.4 Model diagnostics for the IMA(0, 1, 1) model fitted to the IBM daily closing stock prices in Series B.

8.2.4 Score Tests for Model Checking

An alternative to the direct use of overfitting in model checking is provided by the Lagrange multiplier or score test procedure, which is also closely related to the portmanteau test procedure. The general score test procedure was presented by Silvey (1959), and its use in diagnostic checking for ARIMA models was discussed initially by Godfrey (1979) and Poskitt and Tremayne (1980). A computational advantage of the score test procedure is that it requires maximum likelihood estimation of parameters only under the null model under test, but it yields tests asymptotically equivalent to the corresponding likelihood ratio tests obtained by directly overfitting the model. Furthermore, the score test statistic is easily computed in the form of the sample size n times a coefficient of determination from a particular ''auxiliary'' regression.

Hence, we assume that an ARMA(p, q) model has been fitted by the maximum likelihood method to the observations $\tilde{w}_t$, and we want to assess the adequacy of the model by testing this null model against the alternative of an ARMA$(p + r, q)$ model or of an ARMA$(p, q + r)$ model. That is, for the ARMA$(p + r, q)$ alternative, we test H_0: $\phi_{p+1} = \cdots = \phi_{p+r} = 0$, while for the ARMA$(p, q + r)$ alternative, we test H_0: $\theta_{q+1} = \cdots = \theta_{q+r} = 0$. The score test procedure is based on the first partial derivatives, or scores, of the log-likelihood function with respect to the model parameters of the alternative model, but evaluated at the ML estimates obtained under the null model. The log-likelihood function is essentially given by $l = -(n/2)\ln(\sigma_a^2) - (\frac{1}{2}\sigma_a^{-2})\sum_{t=1}^{n} a_t^2$. So, the partial derivatives of l with respect

to the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ are

$$\frac{\partial l}{\partial \phi_j} = -\frac{1}{\sigma_a^2} \sum_{t=1}^{n} \frac{\partial a_t}{\partial \phi_j} a_t$$
$$\frac{\partial l}{\partial \theta_j} = -\frac{1}{\sigma_a^2} \sum_{t=1}^{n} \frac{\partial a_t}{\partial \theta_j} a_t$$

As in (7.2.9) and (7.2.10), we have

$$-\frac{\partial a_t}{\partial \phi_j} = u_{t-j} \qquad -\frac{\partial a_t}{\partial \theta_j} = v_{t-j}$$

where $u_t = \theta^{-1}(B)\tilde{w}_t = \phi^{-1}(B)a_t$, and $v_t = -\theta^{-1}(B)a_t$. Given residuals $\hat{a}_t$, obtained from ML fitting of the null model, as

$$\hat{a}_t = \tilde{w}_t - \sum_{j=1}^{p} \hat{\phi}_j \tilde{w}_{t-j} + \sum_{j=1}^{q} \hat{\theta}_j \hat{a}_{t-j} \qquad t = 1, 2, \ldots, n$$

the u_t's and v_t's evaluated under the ML estimates of the null model can be calculated recursively, starting with initial values set equal to zero, for example, as

$$u_t = \tilde{w}_t + \hat{\theta}_1 u_{t-1} + \cdots + \hat{\theta}_q u_{t-q}$$
$$v_t = -\hat{a}_t + \hat{\theta}_1 v_{t-1} + \cdots + \hat{\theta}_q v_{t-q}$$

The score vector of first partial derivatives with respect to all the model parameters $\boldsymbol{\beta}$ can be expressed as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_a^2} \mathbf{X}'\mathbf{a} \tag{8.2.6}$$

where $\mathbf{a} = (a_1, \ldots, a_n)'$ and $\mathbf{X}$ denotes the $n \times (p + q + r)$ matrix whose tth row consists of $(u_{t-1}, \ldots, u_{t-p-r}, v_{t-1}, \ldots, v_{t-q})$ in the case of the ARMA$(p + r, q)$ alternative model and $(u_{t-1}, \ldots, u_{t-p}, v_{t-1}, \ldots, v_{t-q-r})$ in the case of the ARMA$(p, q + r)$ alternative model. Then, similar to (7.2.17), since the large-sample information matrix for $\boldsymbol{\beta}$ can be consistently estimated by $\hat{\sigma}_a^{-2}\mathbf{X}'\mathbf{X}$, where $\hat{\sigma}_a^2 = n^{-1}\sum_{t=1}^{n} \hat{a}_t^2 = n^{-1}\hat{\mathbf{a}}'\hat{\mathbf{a}}$, it follows that the score test statistic for testing that the additional r parameters are equal to zero is

$$\Lambda = \frac{\hat{\mathbf{a}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{a}}}{\hat{\sigma}_a^2} \tag{8.2.7}$$

Godfrey (1979) noted that the computation of the test statistic in (8.2.7) can be given the interpretation as being equal to n times the coefficient of determination in an auxiliary regression equation. That is, if the alternative model is ARMA$(p + r, q)$, we consider the auxiliary regression equation

$$\hat{a}_t = \alpha_1 u_{t-1} + \cdots + \alpha_{p+r} u_{t-p-r} + \beta_1 v_{t-1} + \cdots + \beta_q v_{t-q} + \varepsilon_t$$

while if the alternative model is ARMA$(p, q + r)$, we consider the regression equation

$$\hat{a}_t = \alpha_1 u_{t-1} + \cdots + \alpha_p u_{t-p} + \beta_1 v_{t-1} + \cdots + \beta_{q+r} v_{t-q-r} + \varepsilon_t$$

Let $\hat{\varepsilon}_t$ denote the residuals from the ordinary least-squares estimation of this regression equation. Then from (8.2.7), it is seen that Λ can be expressed, essentially, as

$$\Lambda = \frac{n(\sum_{t=1}^{n} \hat{a}_t^2 - \sum_{t=1}^{n} \hat{\varepsilon}_t^2)}{\sum_{t=1}^{n} \hat{a}_t^2} = n\left(1 - \frac{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}{\sum_{t=1}^{n} \hat{a}_t^2}\right)$$

which is n times the coefficient of determination of the regression of the $\hat{a}_t$'s on the u_{t-j}'s and the v_{t-j}'s. Under the null hypothesis that the fitted ARMA(p, q) model is correct, the statistic Λ has an asymptotic χ^2 distribution with r degrees of freedom, and the null model is rejected as inadequate for large values of Λ.

As argued by Godfrey (1979) and others, rejection of the null model by the score test procedure should not be taken as evidence to adopt the specific alternative model involved, but simply as evidence against the adequacy of the fitted model. Similarly, the score test is expected to have reasonable power even when the alternative model is not correctly specified. Poskitt and Tremayne (1980) showed, for example, that the score test against an ARMA$(p + r, q)$ model alternative is asymptotically identical to a test against an ARMA$(p, q + r)$ alternative. Hence, the score test procedure may not be sensitive to the particular model specified under the alternative, but its performance will, of course, depend on the choice of the number r of additional parameters specified.

We also note an alternative form for the score statistic Λ. By the ML estimation procedure, it follows that the first partial derivatives, $\partial l/\partial \phi_j$, $j = 1, \ldots, p$, and $\partial l/\partial \theta_j$, $j = 1, \ldots, q$, will be identically equal to zero when evaluated at the ML estimates. Hence, the score vector, $\partial l/\partial \boldsymbol{\beta}$, will contain only r nonzero elements when evaluated at the ML estimates from the null model, these being the partial derivatives with respect to the additional r parameters of the alternative model. Thus, the score statistic in (8.2.7) can also be viewed as a quadratic form in these r nonzero values, whose matrix in the quadratic form is a consistent estimate of the inverse of the covariance matrix of these r score values when evaluated at the ML estimates obtained under the null model. Since these r score values are asymptotically normal with zero means under the null model, the validity of the asymptotic $\chi^2(r)$ distribution under the null hypothesis is easily seen.

Newbold (1980) noted that a score test against the alternative of r additional parameters is closely related to an appropriate test statistic based on the first r residual autocorrelations $r_k(\hat{a})$ from the fitted model. The test statistic is essentially a quadratic form in these first r residual autocorrelations, but of a more complex form than the portmanteau statistic in (8.2.2). As a direct illustration, suppose that the fitted or null model is a pure AR(p) model, and the alternative is an ARMA(p, r) model. Then, it follows from above that the variables v_{t-j} are identical to $-\hat{a}_{t-j}$, since $\theta(B) \equiv 1$ under the null model. Hence, the nonzero elements of the score vector in (8.2.6) are equal to $-n$ times the first r residual autocorrelations, $r_1(\hat{a}), \ldots, r_r(\hat{a})$ from the fitted model, and the score test is thus directly seen to be a quadratic form in these first r residual autocorrelations.

8.2.5 Cumulative Periodogram Check

In some situations, particularly in the fitting of seasonal time series, which are discussed in Chapter 9, it may be feared that we have not adequately taken into account the *periodic* characteristics of the series. Therefore, we are on the lookout for periodicities in the residuals. The autocorrelation function will not be a sensitive indicator of such departures from randomness because periodic effects will typically dilute themselves among several

autocorrelations. The periodogram, on the other hand, is specifically designed for the detection of periodic patterns in a background of white noise.

The periodogram of a time series a_t, $t = 1, 2, \ldots, n$, as defined in Section 2.2.1, is

$$I(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} a_t \cos(2\pi f_i t)\right)^2 + \left(\sum_{t=1}^{n} a_t \sin(2\pi f_i t)\right)^2\right] \tag{8.2.8}$$

where $f_i = i/n$ is the frequency. Thus, it is a device for correlating the a_t's with sine and cosine waves of different frequencies. A pattern with given frequency f_i in the residuals is reinforced when correlated with a sine or cosine wave at that same frequency, and so produces a large value of $I(f_i)$.

Cumulative Periodogram. Bartlett (1955) and other authors have shown that the *cumulative periodogram* provides an effective means for the detection of periodic nonrandomness.

The power spectrum $p(f)$ for white noise has a constant value $2\sigma_a^2$ over the frequency domain 0–0.5 cycle. Consequently, the cumulative spectrum for white noise

$$P(f) = \int_0^f p(g)\, dg \tag{8.2.9}$$

plotted against f is a straight-line running from (0, 0) to $(0.5, \sigma_a^2)$, that is, $P(f)/\sigma_a^2$ is a straight-line running from (0, 0) to (0.5, 1).

The periodogram $I(f)$ provides an estimate of the power spectrum at frequency f. In fact, for white noise, $E[I(f)] = 2\sigma_a^2$, and hence the estimate is unbiased. It follows that $(1/n)\sum_{i=1}^{j} I(f_i)$ provides an unbiased estimate of the integrated spectrum $P(f_j)$, and

$$C(f_j) = \frac{\sum_{i=1}^{j} I(f_i)}{ns^2} \tag{8.2.10}$$

an estimate of $P(f_j)/\sigma_a^2$, where s^2 is an estimate of σ_a^2. We will refer to $C(f_j)$ as the *normalized cumulative periodogram.*

Now, if the model was adequate and the parameters known *exactly*, the a_t's could be computed from the data and would yield a white noise series. For a white noise series, the plot of $C(f_j)$ against f_j would be scattered about a straight-line joining the points (0, 0) and (0.5, 1). On the other hand, model inadequacies would produce nonrandom a_t's, whose cumulative periodogram could show systematic deviations from this line. In particular, periodicities in the a_t's would tend to produce a series of neighboring values of $I(f_j)$ that were large. These large ordinates would reinforce each other in $C(f_j)$ and form a bump on the expected straight line.

In practice, we do not know the exact values of the parameters, but only their estimated values. Hence, we do not have the a_t's, but only the estimated residuals $\hat{a}_t$'s. However, for large samples, the periodogram for the $\hat{a}_t$'s will have similar properties to that for the a_t's. Thus, careful inspection of the periodogram of the $\hat{a}_t$'s can provide a useful additional diagnostic check, particularly for indicating periodicities taken account of inadequately.

Example: Series C. We have seen that Series C is well fitted by the (1, 1, 0) model:

$$(1 - 0.82B)\nabla z_t = a_t$$

and somewhat less well by the IMA(0, 2, 2) model:

$$\nabla^2 z_t = (1 - 0.13B - 0.12B^2)a_t$$

which is rather similar to it. We illustrate the cumulative periodogram test by showing what happens when we analyze the residual a's after fitting to the series an inadequate IMA(0, 1, 1) model:

$$\nabla z_t = (1 - \theta B)a_t$$

where the least squares estimate of θ is found to be -0.65. The normalized cumulative periodogram plot of the residuals from this model is shown in Figure 8.5(a). We see immediately that there are marked departures from linearity in the cumulative periodogram. These departures are very pronounced at low frequencies, as might be expected, for example, if the degree of differencing is insufficient. Figure 8.5(b) shows the corresponding plot for the best-fitting IMA(0, 2, 2) model. The points of the cumulative periodogram now cluster more closely about the expected line, although, as we have seen in Table 8.1 and Figure 8.2, other evidence points to the inadequacy of this model.

It is wise to indicate on the diagram the period as well as the frequency. This makes for easy identification of the bumps that occur when residuals contain periodicities. For example, in monthly sales data, bumps near periods 12, 24, 36, and so on might indicate that seasonal effects were accounted for inadequately.

The probability relationship between the cumulative periodogram and the integrated spectrum is precisely the same as that between the empirical cumulative frequency function and the cumulative distribution function. For this reason we can assess deviations of the periodogram from that expected if the $\hat{a}_t$'s were white noise, by use of the Kolmogorov–Smirnov test. Using this test, we can place limit lines about the theoretical line. The limit lines are such that if the $\hat{a}_t$ series were white noise, the cumulative periodogram would deviate from the straight line sufficiently to cross these limits only with the stated probability. Now, because the $\hat{a}_t$'s are fitted values and not the true $\hat{a}_t$'s, we know that even when the model is correct, they will not precisely follow a white noise process. Thus, as a test for model inadequacy, application of the Kolmogorov–Smirnov limits will indicate only approximate probabilities. However, it is worthwhile to show these limits on the cumulative periodogram to provide a rough guide as to what deviations to regard with skepticism and what to take more note of.

The limit lines are such that for a truly random or white noise series, they would be crossed a proportion ε of the time. They are drawn at distances $\pm K_\varepsilon/\sqrt{q}$ above and below the theoretical line, where $q = (n - 2)/2$ for n even and $(n - 1)/2$ for n odd. Approximate values for K_ε are given in Table 8.3.

TABLE 8.3 Coefficients for Calculating Approximate Probability Limits for Cumulative Periodogram Test

ε	0.01	0.05	0.10	0.25
K_ε	1.63	1.36	1.22	1.02

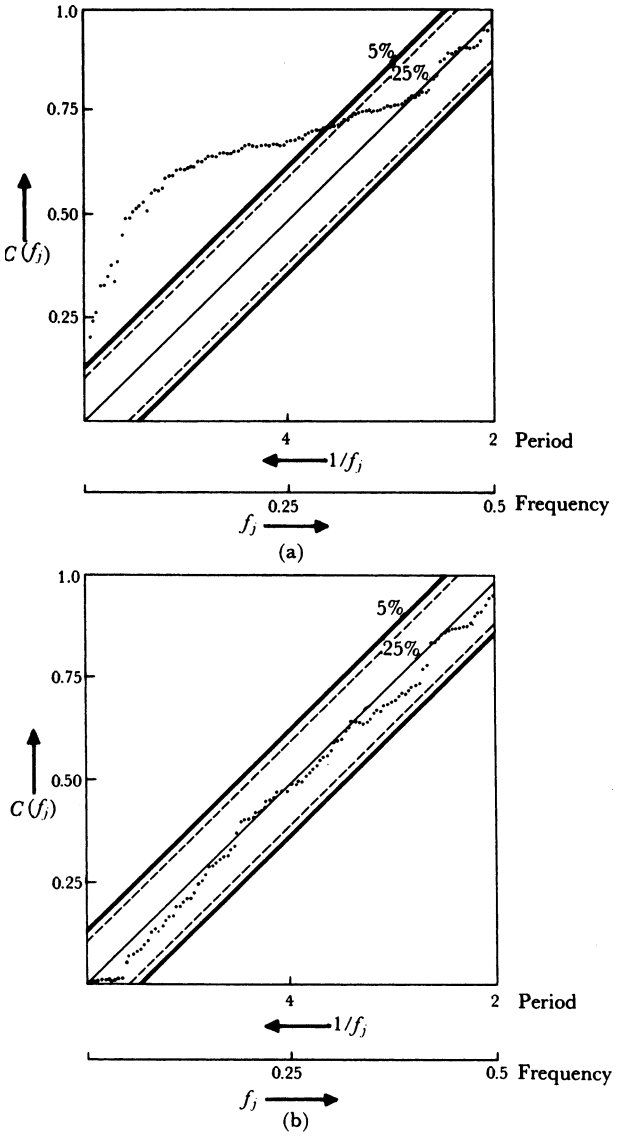

FIGURE 8.5 Series C: cumulative periodograms of residuals from best-fitting models (a) of order (0, 1, 1) and (b) of order (0, 2, 2).

For Series C, $q = (224 - 2)/2 = 111$, and the 5% limit lines inserted on Figure 8.5 deviate from the theoretical line by amounts $\pm 1.36/\sqrt{111} = \pm 0.13$. Similarly, the 25% limit lines deviate by $\pm 1.02/\sqrt{111} = \pm 0.10$.

Conclusions. Each of the model checking procedures described above has essential advantages and disadvantages. Checks based on the study of the estimated autocorrelation function and the cumulative periodogram, although they can point out *unsuspected* peculiarities of the series, may not be particularly sensitive. Tests for specific departures by

overfitting are more sensitive but may fail to warn of trouble other than that specifically anticipated. Portmanteau tests based on the residual autocorrelation and partial autocorrelations, while not always sensitive, provide convenient summary measures that are easy to use. As a result, they are now available in many software packages.

8.3 USE OF RESIDUALS TO MODIFY THE MODEL

8.3.1 Nature of the Correlations in the Residuals When an Incorrect Model Is Used

When the autocorrelation function of the residuals from a fitted model indicates that the model is inadequate, it is necessary to consider in what way the model should be modified. In Section 8.3.2, we show how the autocorrelations of the residuals can be used to suggest such modifications. As an introduction, we consider the effect of fitting an incorrect model on the autocorrelation function of the residuals.

Suppose that the correct model is

$$\phi(B)i\tilde{w}_t = \theta(B)a_t$$

but that an incorrect model

$$\phi_0(B)i\tilde{w}_t = \theta_0(B)b_t$$

is used. Then the residuals b_t, in the incorrect model, will be correlated and since

$$b_t = \theta_0^{-1}(B)\theta(B)\phi_0(B)\phi^{-1}(B)a_t \tag{8.3.1}$$

the autocovariance generating function of the b_t's will be

$$\sigma_a^2[\theta_0^{-1}(B)\theta_0^{-1}(F)\theta(B)\theta(F)\phi_0(B)\phi_0(F)\phi^{-1}(B)\phi^{-1}(F)] \tag{8.3.2}$$

For example, suppose that in an IMA(0, 1, 1) process, instead of the correct value θ, we use some other value θ_0. Then the residuals b_t would follow the mixed process of order (1, 0, 1):

$$(1 - \theta_0 B)b_t = (1 - \theta B)a_t$$

and using (3.4.8), we have

$$\rho_1 = \frac{(1 - \theta\theta_0)(\theta_0 - \theta)}{1 + \theta^2 - 2\theta\theta_0}$$
$$\rho_j = \rho_1\theta_0{}^{j-1} \qquad j = 2, 3, \ldots$$

For example, suppose that in the IMA(0, 1, 1) process,

$$\nabla z_t = (1 - \theta B)a_t$$

we took $\theta_0 = 0.8$ when the correct value was $\theta = 0$. Then

$$\theta_0 = 0.8 \qquad \theta = 0.0$$
$$\rho_1 = 0.8 \qquad \rho_j = 0.8^j$$

Thus, the b_t's would be highly autocorrelated and, since $(1 - 0.8B)b_t = \nabla z_t = a_t$, b_t would follow the autoregressive process

$$(1 - 0.8B)b_t = a_t$$

8.3.2 Use of Residuals to Modify the Model

Suppose that the residuals b_t from the model

$$\phi_0(B)\nabla^{d_0} z_t = \theta_0(B)b_t \tag{8.3.3}$$

appear to be nonrandom, that is, to deviate from white noise behavior. Using the autocorrelation function of b_t, the methods of Chapter 6 may now be applied to identify a model:

$$\phi_1(B)\nabla^{d_1} b_t = \theta_1(B)a_t \tag{8.3.4}$$

for the b_t series.On eliminating b_t between (8.3.3) and (8.3.4), we arrive at a new model:

$$\phi_0(B)\phi_1(B)\nabla^{d_0}\nabla^{d_1} z_t = \theta_0(B)\theta_1(B)a_t \tag{8.3.5}$$

which can now be fitted and diagnostically checked.

For example, suppose that a series had been wrongly identified as an IMA(0, 1, 1) process and fitted to give the model:

$$\nabla z_t = (1 + 0.6B)b_t \tag{8.3.6}$$

Also, suppose that a model

$$\nabla b_t = (1 + 0.8B)a_t \tag{8.3.7}$$

was identified for this residual series. Then on eliminating b_t between (8.3.6) and (8.3.7), we would obtain

$$\begin{aligned}\nabla^2 z_t &= (1 + 0.6B)\nabla b_t \\ &= (1 + 0.6B)(1 - 0.8B)a_t \\ &= (1 - 0.2B - 0.48B^2)a_t\end{aligned}$$

which would suggest that an IMA(0, 2, 2) process should now be entertained.

EXERCISES

8.1. The following are the first 30 residuals obtained when a tentative model was fitted to a time series:

t	Residuals					
1–6	0.78	0.91	0.45	−0.78	−1.90	−2.10
7–12	−0.54	−1.05	0.68	−3.77	−1.40	−1.77
13–18	1.18	0.02	1.29	−1.30	−6.20	−1.89
19–24	0.95	1.49	1.08	0.80	2.02	1.25
25–30	0.52	2.31	1.64	0.78	1.99	1.36

Plot the values and state any reservations you have concerning the adequacy of the model.

8.2. The residuals from a model $\nabla z_t = (1 - 0.6B)a_t$ fitted to a series of $N = 82$ observations yielded the following residual autocorrelations:

k	$r_k(\hat{a})$	k	$r_k(\hat{a})$
1	0.39	6	−0.13
2	0.20	7	−0.05
3	0.09	8	0.06
4	0.04	9	0.11
5	0.09	10	0.02

(a) Plot the residual ACF and determine whether there are any abnormal values relative to white noise behavior.

(b) Calculate the chi-square statistic $\tilde{Q}$ for lags up to $K = 10$ and check whether the residual autocorrelation function as a whole is indicative of model inadequacy.

(c) What modified model would you now tentatively entertain, fit, and check?

8.3. A long series containing $N = 326$ observations was split into two halves and a (1, 1, 0) model $(1 - \phi B)\nabla z_t = a_t$ identified, fitted, and checked for each half. If the estimates of the parameter ϕ for the two halves are $\hat{\phi}^{(1)} = 0.5$ and $\hat{\phi}^{(2)} = 0.7$, is there any evidence that the parameter ϕ has changed?

8.4. **(a)** Show that the variance of the sample mean $\overline{z}$ of n observations from a stationary AR(1) process $(1 - \phi B)\tilde{z}_t = a_t$ is given by

$$\text{var}[\overline{z}] \simeq \frac{\sigma_a^2}{n(1-\phi)^2}$$

(b) The yields from consecutive batches of a chemical process obtained under fairly uniform conditions of process control were shown to follow a stationary AR(1) process $(1 + 0.5B)\tilde{z}_t = a_t$. A technical innovation is made at a given point in time leading to 85 data points with mean $\overline{z}_1 = 41.0$ and residual variance $s_{a1}^2 = 0.1012$ before the innovation is made and 60 data points with $\overline{z}_2 = 43.5$ and $s_{a2}^2 = 0.0895$

after the innovation. Is there any evidence that the innovation has improved (increased) the yield?

8.5. Suppose that a (0, 1, 1) model $\nabla z_t = (1 - \theta B)e_t$, corresponding to the use of an exponentially weighted moving average forecast, with θ arbitrarily chosen to be equal to 0.5, was used to forecast a series that was, in fact, well represented by the (0, 1, 2) model $\nabla z_t = (1 - 0.9B + 0.2B^2)a_t$.

(a) Calculate the autocorrelation function of the lead 1 forecast errors e_t obtained from the (0, 1, 1) model.

(b) Show how this ACF could be used to identify a model for the e_t series, leading to the identification of a (0, 1, 2) model for the z_t series.

8.6. Two time series models, AR(2) and AR(3), were fitted to the yearly time series of sunspot numbers for the period 1770–1869 in Chapter 7. The sunspot data are available for the slightly longer time period 1700–1988 as series 'sunspot.year' in the datasets package in R; type help(sunspot.year) for details. Perform diagnostic checking to determine the adequacy of the AR(2) and AR(3) models for this longer time period. Are there alternative models that you would consider for this series? Would you recommend that a data transformation be used in this case?

8.7. Monthly sales, $\{Y_t\}$, of a company over a period of 150 months are provided as part of Series M in Part 5 of this book. This series is also available as series $BJsales$ along with a related series $BJsales.lead$ in the datasets package in R.

(a) Plot the data and comment.

(b) Perform a statistical analysis to determine a suitable model for this series. Estimate the parameters using the maximum likelihood method.

(c) Repeat the analysis for the series of leading indicator $BJsales.lead$ that is part of the same dataset.

(d) Perform diagnostic checking to determine if there is any lack of fit in the models selected for the two series?

8.8 Global mean surface temperature deviations (from the 1951–1980 average level) are available for the period 1880–2009 as series 'gtemp2' in the astsa package in R.

(a) Plot the data and comment. Are there any unusual features worth noting?

(b) Perform a statistical analysis to determine a suitable model for this series. Estimate the parameters using the maximum likelihood method.

(c) Is there evidences of any lack of fit in the models selected for this series?

(d) Can you suggest an alternative way to analyze this time series? How might an analysis of model generated forecasts impact your choice of model?

8.9 Refer to the daily air quality measurements for New York, May to September 1973, analyzed in Problem 7.10 of Chapter 7. Perform diagnostic checks to determine the adequacy of the models fitted to average daily temperature and wind speed series.

8.10 Repeat the analysis in Problem 8.9 by performing diagnostic checks on the model, or models, considered for the solar radiation series in Problem 7.11.

9

ANALYSIS OF SEASONAL TIME SERIES

In Chapters 3–8, we have considered the properties of a class of linear stochastic models, which are of value in representing stationary and nonstationary time series, and we have seen how these models may be used for forecasting. We then considered the practical problems of identification, fitting, and diagnostic checking that arise when relating these models to actual data. In this chapter, we apply these methods to analyzing and forecasting seasonal time series. A key focus is on seasonal multiplicative time series models that account for time series dependence across seasons as well as between adjacent values in the series. These models are extensions of the ARIMA models discussed in earlier chapters. The methodology is illustrated using a time series commonly referred to as the airline data in the time series literature. We also describe an alternate structural component model approach to representing stochastic seasonal and trend behavior that includes the possibility of the components being deterministic. The chapter concludes with a brief discussion of regression models with autocorrelated errors. These models could include deterministic sine or cosine terms to describe the seasonal behavior of the series.

9.1 PARSIMONIOUS MODELS FOR SEASONAL TIME SERIES

Figure 9.1 shows monthly totals of international airline passengers for the 12-year period from January 1949 to December 1960. This series was discussed by Brown (1962) and is listed as Series G in Part Five of this book. The series is also included as series ''AirPassengers'' in the R datasets package and is conveniently downloaded from there. The series shows a marked seasonal pattern since travel is at its highest in the late summer months, while a secondary peak occurs in the spring. Many other series, particularly sales data, show similar seasonal characteristics.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

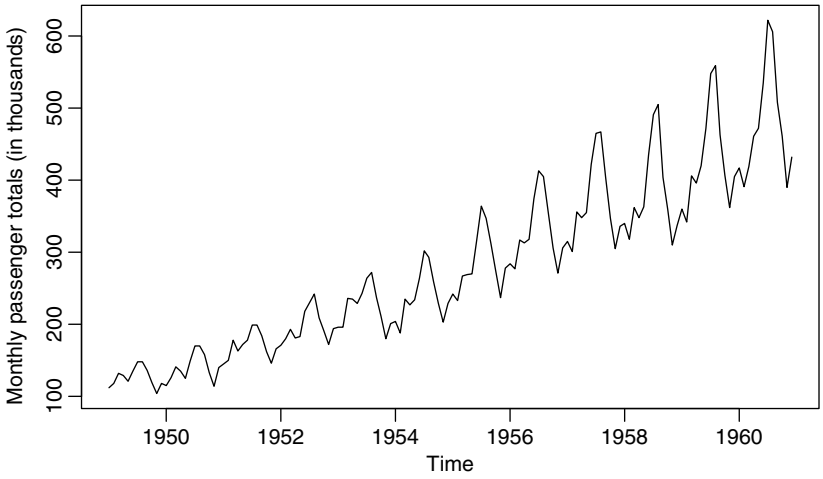

FIGURE 9.1 Totals of international airline passengers in thousands (Series G).

In general, we say that a series exhibits periodic behavior with period s, when similarities in the series occur after s basic time intervals. In the above example, the basic time interval is 1 month and the period is $s = 12$ months. However, examples occur when s can take on other values. For example, $s = 4$ for quarterly data showing seasonal effects within years. It sometimes happens that there is more than one period. Thus, because bills tend to be paid monthly, we would expect weekly business done by a bank to show a periodicity of about 4 within months, while monthly business shows a periodicity of 12.

9.1.1 Fitting Versus Forecasting

A common method of analyzing seasonal time series in the past was to decompose the series arbitrarily into three components: a *trend*, a *seasonal component*, and a *random component*. The trend might be fitted by a polynomial and the seasonal component by a Fourier series. A forecast was then made by projecting these fitted functions. However, such methods could give misleading results if applied indiscriminately. For example, we have seen that the behavior of IBM stock prices in Series B is closely approximated by the random walk model $\nabla z_t = a_t$, that is,

$$z_t = z_0 + \sum_{j=0}^{t-1} a_{t-j} \tag{9.1.1}$$

This implies that $\hat{z}_t(l) = z_t$. In other words, the best forecast of future values of the stock is very nearly today's price. While it is true that short segments of Series B look as if they might be fitted by quadratic curves, this simply reflects the fact that a sum of random deviates can sometimes have this appearance. There is no basis for the use of a quadratic forecast function, which would produce very poor forecasts for this particular series. Similarly, while deterministic trend and seasonal components can provide a good fit to the data, they are often too rigid when it comes to forecasting. In this section, we introduce a seasonal

time series model that requires very few parameters and avoids the assumption of a trend and seasonal component that remains fixed over time.

9.1.2 Seasonal Models Involving Adaptive Sines and Cosines

The general linear model

$$\tilde{z}_t = \sum_{j=1}^{\infty} \pi_j \tilde{z}_{t-j} + a_t = \sum_{j=1}^{\infty} \psi_j a_{t-j} + a_t \tag{9.1.2}$$

with suitable values for the coefficients π_j and ψ_j can be used to describe many seasonal time series. The problem is to choose a suitable *parsimonious parameterization* for such models. We have seen that for nonseasonal series, it is usually possible to obtain a useful and parsimonious representation in the form

$$\varphi(B)\tilde{z}_t = \theta(B)a_t \tag{9.1.3}$$

Moreover, the generalized autoregressive operator $\varphi(B)$ determines the eventual forecast function, which is the solution of the difference equation

$$\varphi(B)\hat{z}_t(l) = 0$$

where B is understood to operate on l. In representing seasonal behavior, we want the forecast function to trace out a periodic pattern. A first thought might be that $\varphi(B)$ should produce a forecast function consisting of a mixture of sines and cosines, and possibly mixed with polynomial terms, to allow changes in the level of the series and changes in the seasonal pattern. Such a forecast function could arise naturally within the structure of the general model (9.1.3). For example, with monthly data, a forecast function that is a sine wave with a 12-month period, adaptive in phase and amplitude, will satisfy the difference equation

$$(1 - \sqrt{3}B + B^2)\hat{z}_t(l) = 0$$

where B is understood to operate on l. However, periodic behavior may not be *economically* represented by mixtures of sines and cosines. Many sine–cosine components would, for example, be needed to represent sales data affected by Christmas, Easter, and other seasonal buying. To take an extreme case, sales of fireworks in Britain are largely confined to the weeks immediately before November 5, when the abortive attempt of Guy Fawkes to blow up the Houses of Parliament is celebrated. An attempt to represent the ''single spike'' of fireworks sales data directly by sines and cosines might be unprofitable. It is clear that a more careful consideration of the problem is needed.

Now, in our previous analysis, we have not necessarily estimated *all* the components of $\varphi(B)$. Where differencing d times was needed to induce stationarity, we have written $\varphi(B) = \phi(B)(1 - B)^d$, which is equivalent to setting d roots of the equation $\varphi(B) = 0$ equal to unity. When such a representation proved adequate, we could proceed with the simpler analysis of $w_t = \nabla^d z_t$. Thus, we have used $\nabla = 1 - B$ as a simplifying operator. In other problems, different types of simplifying operators might be appropriate. For example, the consumption of fuel oil for heat is highly dependent on ambient temperature, which, because the Earth rotates around the sun, is known to follow approximately a sine wave with

period of 12 months. In analyzing sales of fuel oil, it might then make sense to introduce $1 - \sqrt{3}B + B^2$ as a simplifying operator, constituting one of the contributing components of the generalized autoregressive operator $\varphi(B)$. If such a representation proved useful, we could then proceed with the simpler analysis of $w_t = (1 - \sqrt{3}B + B^2)z_t$. This operator is of the homogeneous nonstationary variety, having zeros $e^{\pm(i2\pi/12)}$ on the unit circle.

9.1.3 General Multiplicative Seasonal Model

Simplifying Operator $\mathbf{1-}B^s$**.** The fundamental fact about seasonal time series with period s is that observations that are s intervals apart are similar. Therefore, one can expect that the operation $B^s z_t = z_{t-s}$ will play a particularly important role in the analysis of seasonal series. Furthermore, since nonstationarity is to be expected in the series $z_t, z_{t-s}, z_{t-2s}, \ldots$, the simplifying operation

$$\nabla_s z_t = (1 - B^s)z_t = z_t - z_{t-s}$$

should be useful. This nonstationary operator $1 - B^s$ has s zeros $e^{i(2\pi k/s)}(k = 0, 1, \ldots, s - 1)$ evenly spaced on the unit circle. Moreover, the eventual forecast function satisfies $(1 - B^s)\hat{z}_t(l) = 0$ and so may (but need not) be represented by a full complement of sines and cosines:

$$\hat{z}_t(l) = b_0^{(t)} + \sum_{j=1}^{[s/2]} \left[b_{1j}^{(t)} \cos\left(\frac{2\pi jl}{s}\right) + b_{2j}^{(t)} \sin\left(\frac{2\pi jl}{s}\right) \right]$$

where the b's are adaptive coefficients, and where $[s/2] = \frac{1}{2}s$ if s is even and $[s/2] = \frac{1}{2}(s - 1)$ if s is odd.

Multiplicative Model. When a series exhibits seasonal behavior with known periodicity s, it is useful to display the data in the form of a table containing s columns, such as Table 9.1, which shows the logarithms of the airline data. For seasonal data, special care is needed in selecting an appropriate transformation. In this example, data analysis supports the use of the logarithm (see Section 9.3.5).

The arrangement of Table 9.1 emphasizes the fact that, in periodic data, there are not one but two time intervals of importance. For this example, these intervals correspond to months and years. Specifically, we expect relationships to occur (a) between the observations for successive months in a particular year and (b) between the observations for the same month in successive years. The situation is somewhat like that in a two-way analysis of variance model, where similarities can be expected between observations in the same column and between observations in the same row.

For the airline data, the seasonal effect implies that an observation for a particular month, say April, is related to the observations for previous Aprils. Suppose that the t-th observation z_t is for the month of April. We might be able to link this observation z_t to observations in previous Aprils by a model of the form

$$\Phi(B^s)\nabla_s^D z_t = \Theta(B^s)\alpha_t \tag{9.1.4}$$

TABLE 9.1 Natural Logarithms of Monthly Passenger Totals (Measured in Thousands) in International Air Travel (Series G)

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1949	4.718	4.771	4.883	4.860	4.796	4.905	4.997	4.997	4.913	4.779	4.644	4.771
1950	4.745	4.836	4.949	4.905	4.828	5.004	5.136	5.136	5.063	4.890	4.736	4.942
1951	4.977	5.011	5.182	5.094	5.147	5.182	5.293	5.293	5.215	5.088	4.984	5.112
1952	5.142	5.193	5.263	5.199	5.209	5.384	5.438	5.489	5.342	5.252	5.147	5.268
1953	5.278	5.278	5.464	5.460	5.434	5.493	5.576	5.606	5.468	5.352	5.193	5.303
1954	5.318	5.236	5.460	5.245	5.455	5.576	5.710	5.680	5.557	5.434	5.313	5.434
1955	5.489	5.451	5.587	5.595	5.598	5.753	5.897	5.849	5.743	5.613	5 648	5.628
1956	5.649	5.624	5.759	5.746	5.762	5.924	6.023	6.004	5.872	5.724	5.602	5.724
1957	5.753	5.707	5.875	5.852	5.872	6.045	6.142	6.146	6.001	5.849	5.720	5.817
1958	5.829	5.762	5.892	5.852	5.894	6.075	6.196	6.225	6.001	5.883	5.737	5.820
1959	5.886	5.835	6.006	5.981	6.040	6.157	6.306	6.326	6.138	6.009	5.892	6.004
1960	6.033	5.969	6.038	6.133	6.157	6.282	6.433	6.407	6.230	6.133	5.966	6.068

where $s = 12$, $\nabla_s = 1 - B^s$, and $\Phi(B^s)$, $\Theta(B^s)$ are polynomials in B^s of degrees P and Q, respectively, and satisfying stationarity and invertibility conditions. Similarly, a model

$$\Phi(B^s)\nabla_s^D z_{t-1} = \Theta(B^s)\alpha_{t-1} \tag{9.1.5}$$

might be used to link the current behavior for March with previous March observations, and so on, for each of the 12 months. Moreover, it is usually reasonable to assume that the parameters $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ contained in these monthly models would be approximately the same for each month.

Now the error components, $\alpha_t, \alpha_{t-1}, \ldots$, in these models would not in general be uncorrelated. For example, the total of airline passengers in April 1960, while related to previous April totals, would also be related to totals in March 1960, February 1960, January 1960, and so on. Thus, we would expect that α_t in (9.1.4) would be related to α_{t-1} in (9.1.5) and to α_{t-2}, and so on. Therefore, to account for such relationships, we introduce a second model

$$\phi(B)\nabla^d \alpha_t = \theta(B)a_t \tag{9.1.6}$$

where now a_t is a white noise process and $\phi(B)$ and $\theta(B)$ are polynomials in B of degrees p and q, respectively, and satisfying stationarity and invertibility conditions, and $\nabla = \nabla_1 = 1 - B$.

Substituting (9.1.6) in (9.1.4), we obtain a general multiplicative model

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{9.1.7}$$

where, for this particular example, $s = 12$. Also, the subscripts p, P, q, and Q have been added to indicate the orders of the various operators. The resulting multiplicative process will be said to be of order $(p, d, q) \times (P, D, Q)_s$. A similar argument can be used to obtain models with three or more periodic components to take care of multiple seasonalities.

In the next two sections, we examine some basic forms of the seasonal model introduced above and demonstrate their potential for forecasting. We also consider the problems of identification, estimation, and diagnostic checking that arise in relating such models to data. No new principles are needed to do this, merely an application of the procedures and ideas already discussed in Chapters 6–8. This is illustrated in the next section where a seasonal ARIMA model of order $(0, 1, 1) \times (0, 1, 1)_{12}$ is used to represent the airline data.

9.2 REPRESENTATION OF THE AIRLINE DATA BY A MULTIPLICATIVE $(0, 1, 1) \times (0, 1, 1)_{12}$ MODEL

9.2.1 Multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ Model

We have seen that a simple and widely applicable stochastic model for the analysis of nonstationary time series, which contains no seasonal component, is the IMA(0, 1, 1) process. Suppose, following the argument presented above, that we have a seasonal time series and employ the model

$$\nabla_{12} z_t = (1 - \Theta B^{12})\alpha_t$$

for linking z's 1-year apart. Suppose further that we employ a similar model

$$\nabla \alpha_t = (1 - \theta B) a_t$$

for linking α's 1-month apart, where in general θ and Θ will have different values. Then, on combining these expressions, we obtain the seasonal multiplicative model

$$\nabla \nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12}) a_t \tag{9.2.1}$$

of order $(0, 1, 1) \times (0, 1, 1)_{12}$. The model written explicitly is

$$z_t - z_{t-1} - z_{t-12} + z_{t-13} = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta \Theta a_{t-13} \tag{9.2.2}$$

The invertibility region for this model, required by the condition that the roots of $(1 - \theta B)(1 - \Theta B^{12}) = 0$ lie outside the unit circle, is defined by the inequalities $-1 < \theta < 1$ and $-1 < \Theta < 1$. Note that the moving average operator $(1 - \theta B)(1 - \Theta B^{12}) = 1 - \theta B - \Theta B^{12} + \theta \Theta B^{13}$, on the right-hand side of (9.2.1), is of order $q + sQ = 1 + 12(1) = 13$.

We will show below that the logged airline data are well represented by a model of this form, where to a sufficient approximation, $\hat{\theta} = 0.4$, $\hat{\Theta} = 0.6$, and $\hat{\sigma}_a^2 = 1.34 \times 10^{-3}$. However, as a preliminary, we first consider how this model and with these parameter values inserted can be used to forecast future values of the series.

9.2.2 Forecasting

In Chapter 4, we saw that there are three basically different ways of considering the general model, each giving rise to a different way of viewing the forecast in Chapter 5. We consider now these three approaches for the forecasting of the seasonal model introduced above.

Difference Equation Approach. Forecasts are best *computed* directly from the difference equation itself. Thus, since

$$z_{t+l} = z_{t+l-1} + z_{t+l-12} - z_{t+l-13} + a_{t+l} - \theta a_{t+l-1} - \Theta a_{t+l-12} + \theta \Theta a_{t+l-13} \tag{9.2.3}$$

after setting $\theta = 0.4$, $\Theta = 0.6$, the minimum mean square error forecast at lead time l and origin t is given immediately by

$$\hat{z}_t(l) = [z_{t+l-1} + z_{t+l-12} - z_{t+l-13} + a_{t+l} - 0.4 a_{t+l-1} - 0.6 a_{t+l-12} + 0.24 a_{t+l-13}] \tag{9.2.4}$$

where

$$[z_{t+l}] = E[z_{t+l} | z_t, z_{t-1}, \cdots; \theta, \Theta]$$

is the conditional expectation of z_{t+l} taken at origin t. In this expression, the parameters are assumed to be known, and knowledge of the series $z_t, z_{t-1}, \ldots$ is assumed to extend into the remote past.

Practical application depends upon the following facts:

1. Invertible models fitted to actual data usually yield forecasts that depend appreciably only on recent values of the series.

2. The forecasts are insensitive to small changes in parameter values such as are introduced by estimation errors.

Now

$$[z_{t+j}] = \begin{cases} z_{t+j} & j \le 0 \\ \hat{z}_t(j) & j > 0 \end{cases} \tag{9.2.5}$$

$$[a_{t+j}] = \begin{cases} a_{t+j} & j \le 0 \\ 0 & j > 0 \end{cases} \tag{9.2.6}$$

Thus, to obtain the forecasts, we simply replace unknown z's by forecasts and unknown a's by zeros. The known a's are, of course, the one-step-ahead forecast errors already computed, that is, $a_t = z_t - \hat{z}_{t-1}(1)$.

For example, to obtain the 3-months-ahead forecast, we have

$$z_{t+3} = z_{t+2} + z_{t-9} - z_{t-10} + a_{t+3} - 0.4a_{t+2} - 0.6a_{t-9} + 0.24a_{t-10}$$

Taking conditional expectations at the origin t gives

$$\hat{z}_t(3) = \hat{z}_t(2) + z_{t-9} - z_{t-10} - 0.6a_{t-9} + 0.24a_{t-10}$$

Substituting $a_{t-9} = z_{t-9} - \hat{z}_{t-10}(1)$ and $a_{t-10} = z_{t-10} - \hat{z}_{t-11}(1)$ on the right-hand side also yields

$$\hat{z}_t(3) = \hat{z}_t(2) + 0.4z_{t-9} - 0.76z_{t-10} + 0.6\hat{z}_{t-10}(1) - 0.24\hat{z}_{t-11}(1) \tag{9.2.7}$$

which expresses the forecast in terms of previous z's and previous forecasts of z's.

Figure 9.2 shows the forecasts for lead times up to 36 months, all made at the arbitrarily selected origin, July 1957. We see that the simple model, containing only two parameters, faithfully reproduces the seasonal pattern and supplies excellent forecasts. It is to be remembered, of course, that like all predictions obtained from the general linear stochastic model, the forecast function is adaptive. When changes occur in the seasonal pattern, these will be appropriately projected into the forecast. It will be noticed that when the 1-month-ahead forecast is too high, there is a tendency for all future forecasts from the point to be high. This is to be expected because, as has been noted in Appendix A5.1, forecast errors from the same origin, but for different lead times, are highly correlated. Of course, a forecast for a long lead time, such as 36 months, may necessarily contain a fairly large error. However, in practice, an initially remote forecast will be updated continually, and as the lead shortens, greater accuracy will be possible.

The preceding forecasting procedure is robust to moderate changes in the parameter values. Thus, if we used $\theta = 0.5$ and $\Theta = 0.5$, instead of $\theta = 0.4$ and $\Theta = 0.6$, the forecasts would not be greatly affected. This is true even for forecasts made several steps ahead (e.g., 12 months). The approximate effect on the one-step-ahead forecasts of modifying the values of the parameters can be seen by studying the sum-of-squares surface. Thus, we know that the approximate confidence region for the k parameters $\boldsymbol{\beta}$ is bounded, in general, by the contour $S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}})[1 + \chi^2_\varepsilon(k)/n]$, which includes the true parameter point with probability $1 - \varepsilon$. Therefore, we know that, had the *true* parameter values been employed,

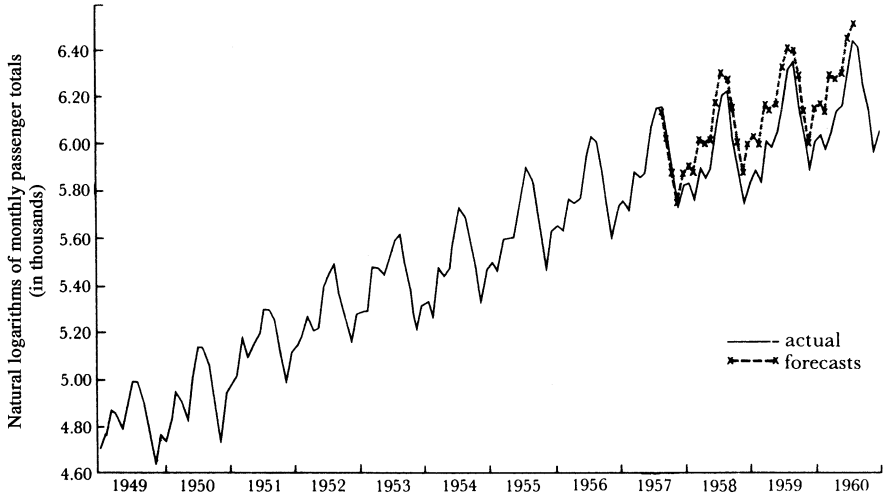

FIGURE 9.2 Airline data with forecasts for 1, 2, 3, …, 36 months ahead, all made from an arbitrary selected origin, July 1957.

with this same probability the mean square of the one-step-ahead forecast errors could not have been increased by a factor greater than $1 + \chi^2_\varepsilon(k)/n$.

Forecast Function, Its Updating, and the Forecast Error Variance. In practice, the difference equation procedure is by far the simplest and most convenient way for actually *computing* forecasts and updating them. However, the difference equation itself does not reveal very much about the *nature* of the forecasts and their updating. To cast light on these aspects, we now consider the forecasts from other points of view.

Forecast Function. Using (5.1.12) yields $z_{t+l} = \hat{z}_t(l) + e_t(l)$, where

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} \tag{9.2.8}$$

Now, the moving average operator on the right-hand side of (9.2.1) is of order 13. Hence, for $l > 13$, the forecasts satisfy the difference equation

$$(1 - B)(1 - B^{12})\hat{z}_t(l) = 0 \qquad l > 13 \tag{9.2.9}$$

where, in this equation, B operates on the lead time l.

We now write $l = (r, m) = 12r + m, r = 0, 1, 2, \ldots$ and $m = 1, 2, \ldots, 12$, to represent a lead time of r years and m months, so that, for example, $l = 15 = (1, 3)$. Then, the forecast function, which is the solution of (9.2.9), with starting conditions given by the first 13 forecasts, is of the form

$$\hat{z}_t(l) = \hat{z}_t(r, m) = b^{(t)}_{0,m} + r b^{(t)}_1 \qquad l > 0 \tag{9.2.10}$$

This forecast function contains 13 adjustable coefficients $b^{(t)}_{0,1}, b^{(t)}_{0,2}, \ldots, b^{(t)}_{0,12}, b^{(t)}_1$. These represent 12 monthly contributions and 1 yearly contribution and are determined by the

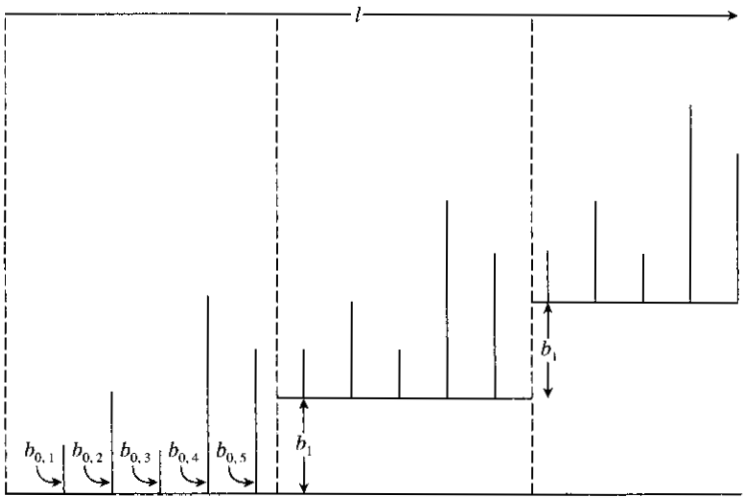

FIGURE 9.3 Seasonal forecast function generated by the model $\nabla\nabla_{12}z_t = (1 - \theta B)(1 - \Theta B^s)a_t$, with $s = 5$.

first 13 forecasts. The nature of this function is more clearly understood from Figure 9.3, which shows a forecast function of this kind, but with period $s = 5$, so that there are six adjustable coefficients $b_{0,1}^{(t)}, b_{0,2}^{(t)}, \ldots, b_{0,5}^{(t)}, b_1^{(t)}$.

Equivalently, since $\hat{z}_t(l)$ satisfies (9.2.9) and the roots of $(1 - B)(1 - B^{12}) = 0$ are $1, 1, -1, e^{\pm(i2\pi k/12)}, k = 1, \ldots, 5$, on the unit circle, the forecast function, as in (5.3.3), can be represented as

$$\hat{z}_t(l) = \sum_{j=1}^{5} \left[b_{1j}^{(t)} \cos\left(\frac{2\pi jl}{12}\right) + b_{2j}^{(t)} \sin\left(\frac{2\pi jl}{12}\right) \right] + b_{16}^{(t)}(-1)^l + b_0^{(t)} + b_1^{*(t)} l$$

This shows that $\hat{z}_t(l)$ consists of a mixture of sinusoids at the seasonal frequencies $2\pi j/12, j = 1, \ldots, 6$, plus a linear trend with slope $b_1^{*(t)}$. The coefficients $b_{1j}^{(t)}, b_{2j}^{(t)}, b_0^{(t)}$, and $b_1^{*(t)}$ in the expression above are all adaptive with regard to the forecast origin t, being determined by the first 13 forecasts. In comparison to (9.2.10), it is clear, for example, that $b_1^{(t)} = 12b_1^{*(t)}$, and represents the *annual* rate of change in the forecasts $\hat{z}_t(l)$, whereas $b_1^{*(t)}$ is the *monthly* rate of change.

The ψ Weights. To determine updating formulas and to obtain the variance of the forecast error $e_t(l)$ in (9.2.8), we need the ψ weights in the form $z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$ of the model. We can write the moving average operator in (9.2.1) in the form

$$(1 - \theta B)(1 - \Theta B^{12}) = (\nabla + \lambda B)(\nabla_{12} + \Lambda B^{12})$$

where $\lambda = 1 - \theta, \Lambda = 1 - \Theta, \nabla_{12} = 1 - B^{12}$. Hence, the model may be written as

$$\nabla\nabla_{12}z_t = (\nabla + \lambda B)(\nabla_{12} + \Lambda B^{12})a_t$$

By equating coefficients in $\nabla\nabla_{12}\psi(B) = (\nabla + \lambda B)(\nabla_{12} + \Lambda B^{12})$, it can be seen that the ψ weights satisfy $\psi_0 = 1, \psi_1 - \psi_0 = \lambda - 1, \psi_{12} - \psi_{11} - \psi_0 = \Lambda - 1, \psi_{13} - \psi_{12} - \psi_1 + \psi_0 = (\lambda - 1)(\Lambda - 1)$, and $\psi_j - \psi_{j-1} - \psi_{j-12} + \psi_{j-13} = 0$ otherwise. Thus, the ψ weights for this process are

$$\begin{aligned} \psi_1 = \psi_2 = \cdots = \psi_{11} &= \lambda & \psi_{12} &= \lambda + \Lambda \\ \psi_{13} = \psi_{14} = \cdots = \psi_{23} &= \lambda(1 + \Lambda) & \psi_{24} &= \lambda(1 + \Lambda) + \Lambda \\ \psi_{25} = \psi_{26} = \cdots = \psi_{35} &= \lambda(1 + 2\Lambda) & \psi_{36} &= \lambda(1 + 2\Lambda) + \Lambda \end{aligned}$$

and so on. Writing ψ_j as $\psi_{r,m} = \psi_{12r+m}$, where $r = 0, 1, 2, \ldots$ and $m = 1, 2, \ldots, 12$, refer, respectively, to years and months, we obtain

$$\psi_{r,m} = \lambda(1 + r\Lambda) + \delta\Lambda \tag{9.2.11}$$

where

$$\delta = \begin{cases} 1 \text{ when } m = 12 \\ 0 \text{ when } m \neq 12 \end{cases}$$

Updating. The general updating formula (5.2.5) is

$$\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + \psi_l a_{t+1}$$

Thus, if $m \neq s = 12$,

$$b_{0,m}^{(t+1)} + r b_1^{(t+1)} = b_{0,m+1}^{(t)} + r b_1^{(t)} + (\lambda + r\lambda\Lambda) a_{t+1}$$

and on equating coefficients of r, the updating formulas are

$$\begin{aligned} b_{0,m}^{(t+1)} &= b_{0,m+1}^{(t)} + \lambda a_{t+1} \\ b_1^{(t+1)} &= b_1^{(t)} + \lambda\Lambda a_{t+1} \end{aligned} \tag{9.2.12}$$

Alternatively, if $m = s = 12$,

$$b_{0,12}^{(t+1)} + r b_1^{(t+1)} = b_{0,1}^{(t)} + (r+1) b_1^{(t)} + (\lambda + \Lambda + r\lambda\Lambda) a_{t+1}$$

and in this case,

$$\begin{aligned} b_{0,12}^{(t+1)} &= b_{0,1}^{(t)} + b_1^{(t)} + (\lambda + \Lambda) a_{t+1} \\ b_1^{(t+1)} &= b_1^{(t)} + \lambda\Lambda a_{t+1} \end{aligned} \tag{9.2.13}$$

In studying these relations, it should be remembered that $b_{0,m}^{(t+1)}$ will be the updated version of $b_{0,m+1}^{(t)}$. Thus, if the origin t was January of a particular year, $b_{0,2}^{(t)}$ would be the coefficient for March. After a month had elapsed, we should move the forecast origin to February and the updated version for the March coefficient would now be $b_{0,1}^{(t+1)}$.

Forecast Error Variance. Knowledge of the ψ weights enables us to calculate the variance of the forecast errors at any lead time l, using the result (5.1.16), namely

$$V(l) = (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 \tag{9.2.14}$$

Thus, setting $\lambda = 0.6, \Lambda = 0.4, \sigma_a^2 = 1.34 \times 10^{-3}$ in (9.2.11) and (9.2.14), the estimated standard deviations $\hat{\sigma}(l)$ of the forecast errors of the log airline data are readily calculated for different lead times.

Forecasts as a Weighted Average of Previous Observations. If we write the model in the form

$$z_t = \sum_{j=1}^{\infty} \pi_j z_{t-j} + a_t$$

the one-step-ahead forecast is

$$\hat{z}_t(1) = \sum_{j=1}^{\infty} \pi_j z_{t+1-j}$$

The π weights may be obtained by equating coefficients in

$$(1 - B)(1 - B^{12}) = (1 - \theta B)(1 - \Theta B^{12})(1 - \pi_1 B - \pi_2 B^2 - \cdots)$$

Thus,

$$\begin{aligned} \pi_j &= \theta^{j-1}(1 - \theta) \qquad j = 1, 2, \ldots, 11 \\ \pi_{12} &= \theta^{11}(1 - \theta) + (1 - \Theta) \\ \pi_{13} &= \theta^{12}(1 - \theta) - (1 - \theta)(1 - \Theta) \\ \pi_j &- \theta\pi_{j-1} - \Theta\pi_{j-12} + \theta\Theta\pi_{j-13} \qquad j \geq 14 \end{aligned} \tag{9.2.15}$$

These weights are plotted in Figure 9.4 for the parameter values $\theta = 0.4$ and $\Theta = 0.6$.

The reason that the weight function takes the particular form shown in the figure may be understood as follows: the process (9.2.1) may be written as

$$a_{t+1} = \left(1 - \frac{\lambda B}{1 - \theta B}\right)\left(1 - \frac{\Lambda B^{12}}{1 - \Theta B^{12}}\right) z_{t+1} \tag{9.2.16}$$

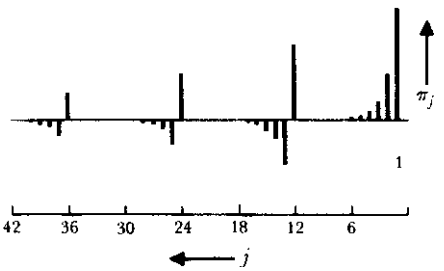

FIGURE 9.4 The π weights for $(0, 1, 1) \times (0, 1, 1)_{12}$ process fitted to the airline data ($\theta = 0.4, \Theta = 0.6$).

We now use the notation $\text{EWMA}_\lambda(z_t)$ to mean an exponentially weighted moving average, with parameter $\lambda = 1 - \theta$ of values $z_t, z_{t-1}, z_{t-2}, \ldots$, so that

$$\text{EWMA}_\lambda(z_t) = \frac{\lambda}{1 - \theta B} z_t = \lambda z_t + \lambda\theta z_{t-1} + \lambda\theta^2 z_{t-2} + \cdots$$

Similarly, we use $\text{EWMA}_\Lambda(z_t)$ to mean an exponentially weighted moving average, with parameter $\Lambda = 1 - \Theta$, of values $z_t, z_{t-12}, z_{t-24}, \ldots$, so that

$$\text{EWMA}_\Lambda(z_t) = \frac{\Lambda}{1 - \Theta B^{12}} z_t = \Lambda z_t + \Lambda\Theta z_{t-12} + \Lambda\Theta^2 z_{t-24} + \cdots$$

Substituting $\hat{z}_t(1) = z_{t+1} - a_{t+1}$, in (9.2.16), we obtain

$$\hat{z}_t(1) = \text{EWMA}_\lambda(z_t) + \text{EWMA}_\Lambda(z_{t-11} - \text{EWMA}_\lambda(z_{t-12})) \tag{9.2.17}$$

Thus, the forecast is an EWMA taken over previous months, modified by a second EWMA of discrepancies found between similar monthly EWMAs and actual performance in previous years. As a particular case, if $\theta = 0$ ($\lambda = 1$), (9.2.17) would reduce to

$$\begin{aligned}\hat{z}_t(1) &= z_t + \text{EWMA}_\Lambda(z_{t-11} - z_{t-12}) \\ &= z_t + \Lambda[(z_{t-11} - z_{t-12}) + \Theta(z_{t-23} - z_{t-24}) + \cdots]\end{aligned}$$

which shows that first differences are forecast as the seasonal EWMA of first differences for similar months from previous years.

For example, suppose that we were attempting to predict December sales for a department store. These sales would include a heavy component from Christmas buying. The first term on the right-hand side of (9.2.17) would be an EWMA taken over previous months up to November. However, we know this will be an underestimate, so we correct it by taking a second EWMA over previous years of the *discrepancies* between actual December sales and the corresponding monthly EWMAs taken over previous months in those years.

The forecasts for lead times $l > 1$ can be generated from the π weights by substituting forecasts of shorter lead time for unknown values, as displayed in the general expression (5.3.6) of Section 5.3.3. Alternatively, explicit values for the weights applied directly to $z_t, z_{t-1}, z_{t-2}, \ldots$ may be computed, for example, from (5.3.9) or from (A5.2.3).

Calculation of Forecasts in R. Forecasts of future values of a time series that follows a multiplicative seasonal model can be calculated using R. A convenient option available in R is the command sarima.for() in the astsa package. For a series z_t that follows a multiplicative model with period s, the command is sarima.for(z,n.ahead,p,d,q,P,D,Q,s), where n.ahead is the lead time. Thus, to generate forecasts up to 24 steps ahead for the logged airline series using the model $\nabla\nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12}) a_t$, the commands are

```
> library(astsa)
> ap=ts(seriesG,start=c(1949,1),frequency=12)
> log.AP=log(ap)
> m1=sarima.for(log.AP,24,0,1,1,0,1,1,12)
> m1   % retrieves output from a file
```

The output includes the forecasts (''pred'') and the prediction errors (''se'') of the forecasts. A graph of the forecasts with ± 2 prediction error limits attached is provided as part of the

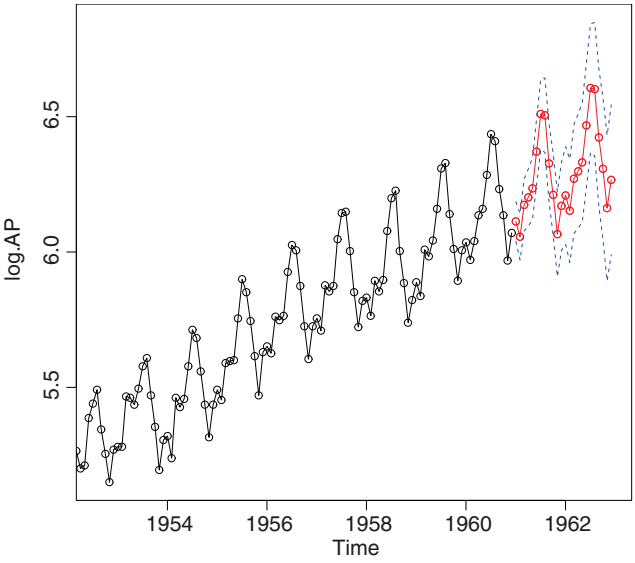

FIGURE 9.5 Forecasts along with ±2 prediction error limits for the logarithm of the airline data generated from the model $\nabla\nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12})a_t$.

output. Figure 9.5 shows the forecasts generated for the logged airline data using these commands.

9.2.3 Model Identification

The identification of the nonseasonal IMA(0, 1, 1) process depends upon the fact that, after taking first differences, the autocorrelations for all lags beyond the first are zero. For the multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ process (9.2.1), the only nonzero autocorrelations of $\nabla\nabla_{12} z_t$ are those at lags 1, 11, 12, and 13. In fact, from (9.2.2) the model is viewed as

$$w_t = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta\Theta a_{t-13}$$

which is an MA model of order 13 for $w_t = \nabla\nabla_{12} z_t$. The autocovariances of w_t are thus given by

$$\begin{aligned}
\gamma_0 &= [1 + \theta^2 + \Theta^2 + (\theta\Theta)^2]\sigma_a^2 = (1 + \theta^2)(1 + \Theta^2)\sigma_a^2 \\
\gamma_1 &= [-\theta - \Theta(\theta\Theta)]\sigma_a^2 = -\theta(1 + \Theta^2)\sigma_a^2 \\
\gamma_{11} &= \theta\Theta\sigma_a^2 \\
\gamma_{12} &= [-\Theta - \theta(\theta\Theta)]\sigma_a^2 = -\Theta(1 + \theta^2)\sigma_a^2 \\
\gamma_{13} &= \theta\Theta\sigma_a^2
\end{aligned} \tag{9.2.18}$$

In particular, these expressions imply that

$$\rho_1 = \frac{-\theta}{1 + \theta^2} \qquad \text{and} \qquad \rho_{12} = \frac{-\Theta}{1 + \Theta^2}$$

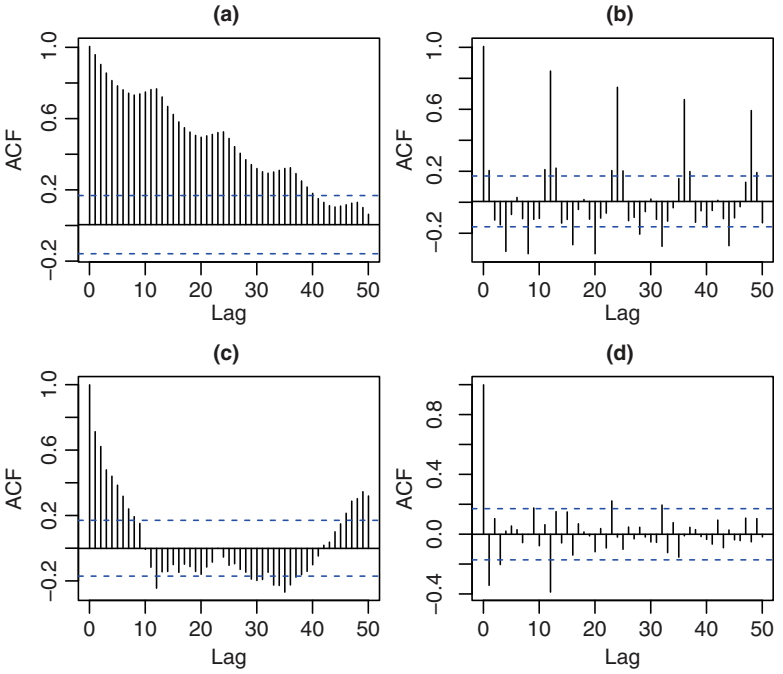

FIGURE 9.6 Estimated autocorrelation function of logged airline data: (a) undifferenced series, (b) first differenced series, (c) seasonally differenced series, and (d) series with regular and seasonal differencing.

so that the value ρ_1 is unaffected by the presence of the seasonal MA factor $(1 - \Theta B^{12})$ in the model (9.2.1), while the value of ρ_{12} is unaffected by the nonseasonal or regular MA factor $(1 - \theta B)$.

Figure 9.6 shows the estimated autocorrelations of the airline data for (a) the logged series, z_t, (b) the logged series differenced with respect to months only, ∇z_t, (c) the logged series differenced with respect to years only, $\nabla_{12} z_t$, and (d) the logged series differenced with respect to months and years, $\nabla\nabla_{12} z_t$. The autocorrelations for z_t are large and fail to die out at higher lags. While simple differencing reduces the correlations in general, a very heavy periodic component remains. This is evidenced particularly by very large correlations at lags 12, 24, 36, and 48. Simple differencing with respect to period 12 results in correlations which are first persistently positive and then persistently negative. By contrast, the differencing $\nabla\nabla_{12}$ markedly reduces correlations throughout.

The autocorrelations of $\nabla\nabla_{12} z_t$ exhibit spikes at lags 1 and 12, compatible with the theoretical autocovariances in (9.2.18) for model (9.2.1). As an alternative, however, the autocorrelations for $\nabla_{12} z_t$ might be viewed as dying out at a slow exponential rate beginning from lag one. Hence, there is also the possibility that $\nabla_{12} z_t$ may follow a nonseasonal ARMA(1, 1) model with ϕ relatively close to one, rather than a nonstationary IMA(0, 1, 1) model as in (9.2.1). However, in practice, the distinction between these two models may not be substantial and the latter model will not be explored further here. The choice between the nonstationary and stationary AR(1) factor could, in fact, be tested using unit root procedures similar to those described in Section 10.1 of the next chapter.

The autocorrelation functions shown in Figure 9.6 was generated in R using the following commands:

```
> library(astsa)
> log.AP=log(ts(seriesG))
> par(mfrow=c(2,2))
> acf(log.AP,50,main='(a)')
> acf(diff(log.AP),50,main='(b)')
> acf(diff(log.AP,12),50,main='(c)')
> acf(diff(diff(log.AP,12)),50,main='(d)')
```

On the assumption that the model is of the form (9.2.1), the variances for the estimated higher lag autocorrelations are approximated by Bartlett's formula (2.1.15), which in this case becomes

$$\text{var}[r_k] \simeq \frac{1 + 2(\rho_1^2 + \rho_{11}^2 + \rho_{12}^2 + \rho_{13}^2)}{n} \qquad k > 13 \tag{9.2.19}$$

Substituting estimated correlations for the ρ's and setting $n = 144 - 13 = 131$ in (9.2.19), where $n = 131$ is the number of differences $\nabla\nabla_{12}z_t$, we obtain a standard error $\hat{\sigma}(r) \simeq 0.11$. The dashed lines shown in Figure 9.6 are approximate two-standard-error limits computed under the assumption that there is no autocorrelation in the series so that var$[r_k]$ = 1/n.

Preliminary Estimates. As with the nonseasonal model, by equating appropriate observed sample correlations to their expected values, approximate values can be obtained for the parameters θ and Θ. On substituting the sample estimates $r_1 = -0.34$ and $r_{12} = -0.39$ in the expressions

$$\rho_1 = \frac{-\theta}{1 + \theta^2} \qquad \rho_{12} = \frac{-\Theta}{1 + \Theta^2}$$

we obtain rough estimates $\hat{\theta} \simeq 0.39$ and $\hat{\Theta} \simeq 0.48$. A table summarizing the behavior of the autocorrelation function for some specimen seasonal models, useful in identification and in obtaining preliminary estimates of the parameters, is given in Appendix A9.1.

9.2.4 Parameter Estimation

Contours of the sum-of-squares function $S(\theta, \Theta)$ for the model (9.2.1) fitted to the airline data are shown in Figure 9.7, together with the appropriate 95% confidence region. The least-squares estimates (LE) are seen to be very nearly $\hat{\theta} = 0.4$ and $\hat{\Theta} = 0.6$. The grid of values for $S(\theta, \Theta)$ was computed using the technique described in Chapter 7. It was shown there that given n observations $\mathbf{w}$ from a linear process defined by

$$\phi(B)w_t = \theta(B)a_t$$

the quadratic form $\mathbf{w}'\mathbf{M}_n\mathbf{w}$, which appears in the exponent of the likelihood, can always be expressed in terms of a sum of squares of the conditional expectation of a's and a quadratic function of the conditional expectation of the $p + q$ initial values

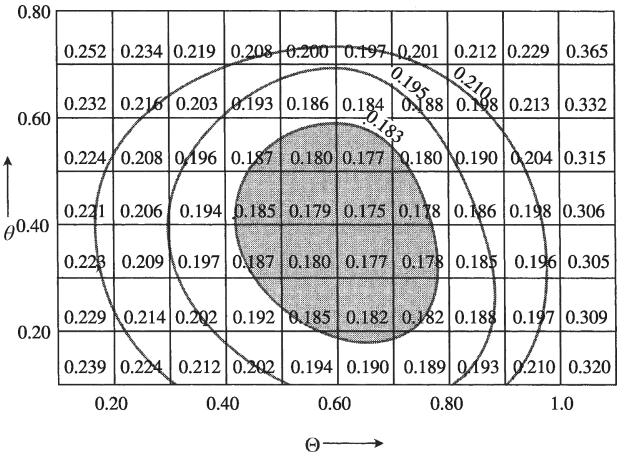

FIGURE 9.7 Contours of $S(\theta, \Theta)$ with shaded 95% confidence region for the model $\nabla\nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12}) a_t$ fitted to the airline data.

$\mathbf{e}_* = (w_{1-p}, \ldots, w_0, a_{1-q}, \ldots, a_0)'$, that is,

$$\mathbf{w}'\mathbf{M}_n\mathbf{w} = S(\boldsymbol{\phi}, \theta) = \sum_{t=-\infty}^{n} [a_t]^2 = \sum_{t=1}^{n} [a_t]^2 + [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*]$$

where $[a_t] = [a_t|\mathbf{w}, \boldsymbol{\phi}, \theta]$, $[\mathbf{e}_*] = [\mathbf{e}_*|\mathbf{w}, \boldsymbol{\phi}, \theta]$, and $\text{cov}[\mathbf{e}_*] = \sigma_a^2\boldsymbol{\Omega}$. Furthermore, $S(\boldsymbol{\phi}, \theta)$ plays a central role in the estimation of the parameters $\boldsymbol{\phi}$ and θ from both a sampling theory and a likelihood or Bayesian point of view.

The computation for seasonal models follows precisely the course described in Section 7.1.5 for nonseasonal models. The airline series has $N = 144$ observations. This reduces to n=131 observations after the differencing $w_t = \nabla\nabla_{12} z_t$. The $[a_t]$ in $S(\theta, \Theta)$ can be calculated recursively using an approximate approach that iterates between the forward and backward versions of the $(0, 1, 1) \times (0, 1, 1)_{12}$ model. Alternatively, an exact method discussed in Appendix A7.3 and also used in Section 7.1.5 can be employed. For the present model, this involves first computing the conditional estimates of the a_t, using zero initial values $a_{-12}^0 = a_{-11}^0 = \cdots = a_0^0 = 0$, through a recursive calculation as

$$a_t^0 = w_t + \theta a_{t-1}^0 + \Theta a_{t-12}^0 - \theta\Theta a_{t-13}^0 \qquad t = 1, \ldots, n \tag{9.2.20}$$

Then a backward recursion is used to obtain a series u_t as

$$u_t = a_t^0 + \theta u_{t+1} + \Theta u_{t+12} - \theta\Theta u_{t+13} \qquad t = n, \ldots, 1$$

using zero initial values $u_{n+1} = \cdots = u_{n+13} = 0$. Finally, the exact estimate for the vector of initial values $\mathbf{a}_*' = (a_{-12}, \ldots, a_o)$ is obtained by solving the equations $\mathbf{D}[\mathbf{a}_*] = \mathbf{F}'\mathbf{u}$, as described in (A7.3.12) of Appendix A7.3. Letting $\mathbf{h} = \mathbf{F}'\mathbf{u} = (h_{-12}, h_{-11}, \ldots, h_0)'$, the values h_{-j} are computed as

$$h_{-j} = -(\theta u_{-j+1} + \Theta u_{-j+12} - \theta\Theta u_{-j+13})$$

with $u_{-j} = 0, j \geq 0$. Once the initial values are estimated, the remaining $[a_t]$ values for $t = 1, 2, \ldots, n$ are calculated recursively as in (9.2.20), and hence the exact sum of squares $S(\theta, \Theta) = \sum_{t=-12}^{131} [a_t]^2$ is obtained.

Iterative Calculation of Least-Squares Estimates $\hat{\theta}$, $\hat{\Theta}$. While it is essential to plot sums-of-squares surfaces in a new situation, or whenever difficulties arise, an iterative linearization technique may be used in straightforward situations to supply the least-squares estimates and their approximate standard errors. The procedure has been set out in Section 7.2.1, and no new difficulties arise in estimating the parameters of seasonal models.

For the present example, we can write approximately

$$a_{t,0} = (\theta - \theta_0)x_{t,1} + (\Theta - \Theta_0)x_{t,2} + a_t$$

where

$$x_{t,1} = -\frac{\partial a_t}{\partial \theta}\bigg|_{\theta_0,\Theta_0} \qquad x_{t,2} = -\frac{\partial a_t}{\partial \Theta}\bigg|_{\theta_0,\Theta_0}$$

and where θ_0 and Θ_0 are guessed values and $a_{t,0} = [a_t|\theta_0, \Theta_0]$. As explained and illustrated in Section 7.2.2, the derivatives are most easily computed numerically. Alternatively, the derivatives could be obtained to any degree of accuracy by recursive calculation.

Proceeding this way and using as starting values, the preliminary estimates $\hat{\theta} = 0.39, \hat{\Theta} = 0.48$ obtained above, parameter estimates correct to two decimals are available in three iterations. The estimated variance of the residuals is $\hat{\sigma}_a^2 = 1.34 \times 10^{-3}$. From the inverse of the matrix of sums of squares and products of the x's on the last iteration, the standard errors of the estimates may now be calculated. The least-squares estimates followed by their standard errors are then

$$\hat{\theta} = 0.40 \pm 0.08$$
$$\hat{\Theta} = 0.61 \pm 0.07$$

agreeing closely with the values obtained from the sum-of-squares plot.

Large-Sample Variances and Covariances for the Estimates. As in Section 7.2.6, large-sample formulas for the variances and covariances of the parameter estimates may be obtained. In this case, from the model equation $w_t = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta\Theta a_{t-13}$, the derivatives $x_{t,1} = -\partial a_t/\partial\theta$ are seen to satisfy

$$x_{t,1} - \theta x_{t-1,1} - \Theta x_{t-12,1} + \theta\Theta x_{t-13,1} + a_{t-1} - \Theta a_{t-13} = 0$$

hence $(1 - \theta B)(1 - \Theta B^{12})x_{t,1} = -(1 - \Theta B^{12})a_{t-1}$, or simply $(1 - \theta B)x_{t,1} = -a_{t-1}$. Thus, using a similar derivation for $x_{t,2} = -\partial a_t/\partial\Theta$, we obtain that

$$x_{t,1} \simeq -(1 - \theta B)^{-1} a_{t-1} = -\sum_{j=0}^{\infty} \theta^j B^j a_{t-1}$$

$$x_{t,2} \simeq -(1 - \Theta B^{12})^{-1} a_{t-12} = -\sum_{i=0}^{\infty} \Theta^i B^{12i} a_{t-12}$$

Therefore, for large samples, the information matrix is

$$\mathbf{I}(\theta, \Theta) = n \begin{bmatrix} (1-\theta^2)^{-1} & \theta^{11}(1-\theta^{12}\Theta)^{-1} \\ \theta^{11}(1-\theta^{12}\Theta)^{-1} & (1-\Theta^2)^{-1} \end{bmatrix}$$

Provided that $|\theta|$ is not close to unity, the off-diagonal term is negligible, and approximate values for the variances and covariances of $\hat{\theta}$ and $\hat{\Theta}$ are

$$V(\hat{\theta}) \simeq n^{-1}(1-\theta^2) \qquad V(\hat{\Theta}) \simeq n^{-1}(1-\Theta^2)$$
$$\operatorname{cov}[\hat{\theta}, \hat{\Theta}] \simeq 0 \tag{9.2.21}$$

In the present example, substituting the values $\hat{\theta} = 0.40$, $\hat{\Theta} = 0.61$, and $n = 131$, we obtain

$$V(\hat{\theta}) \simeq 0.0064 \qquad V(\hat{\Theta}) \simeq 0.0048$$

and

$$\sigma(\hat{\theta}) \simeq 0.08 \qquad \sigma(\hat{\Theta}) \simeq 0.07$$

which, to this accuracy, are identical with the values obtained directly from the iteration. It is also interesting to note that the parameter estimates $\hat{\theta}$ and $\hat{\Theta}$, associated with months and years, respectively, are virtually uncorrelated.

Parameter Estimation in R. The parameters of the model

$$\nabla\nabla_{12} z_t = w_t = (1-\theta B)(1-\Theta B^{12})a_t$$

can be estimated in R using the command sarima(log.AP,p,d,q,P,D,Q,S=12) in the astsa package as demonstrated below. The resulting estimates of the two parameters θ and Θ are 0.40 and 0.56, respectively, with corresponding standard errors of 0.09 and 0.07. The full likelihood function, including the determinant, is used for parameter estimation, which accounts for the difference between the parameter estimates derived above and those obtained in R. Also, in viewing the output, it should be noted that R defines the moving average operators with positive signs, in contrast to the negative signs used in this text.

```
  > library(astsa)
  > log.AP=log(ts(seriesG))
  > m1.AP=sarima(log.AP, 0,1,1,0,1,1,S=12)
  > m1.AP % Retrieves output from file

OUTPUT:
Call:
stats:arima(x=xdata,order=c(p,d,q),seasonal= list(order=c(P,D,Q),
period=S),optim.control=list(trace=trc,REPORT=1,reltol=tol))

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

sigma^2 estimated as 0.001348: log likelihood=244.7,  aic=-483.4
```

9.2.5 Diagnostic Checking

Before proceeding further, we check the adequacy of fit of the model by examining the residuals from the fitted model.

Autocorrelation Checks. The standardized residuals calculated from the fitted model and the estimated autocorrelations of the residuals are shown in Figure 9.8. The figure is generated as part of the output from the estimation command ''sarima'' in R. The residual autocorrelations do not present evidence of any lack of fit, since none of the values fall outside the approximate two-standard-error limits of 0.18. This conclusion is also supported by the p values of the portmanteau statistics $\tilde{Q} = n(n+2)\sum_{k=1}^{K} r_k^2(\hat{a})/(n-k)$ which are shown for different values of K in the last part of the graph.

Periodogram Check. The cumulative periodogram (see Section 8.2.5) for the residuals is shown in Figure 9.9. The Kolmogorov–Smirnov 5 and 25% probability limits, which as we have seen in Section 8.2.5 supply a very rough guide to the significance of apparent deviations, fail in this instance to indicate any significant departure from the assumed model.

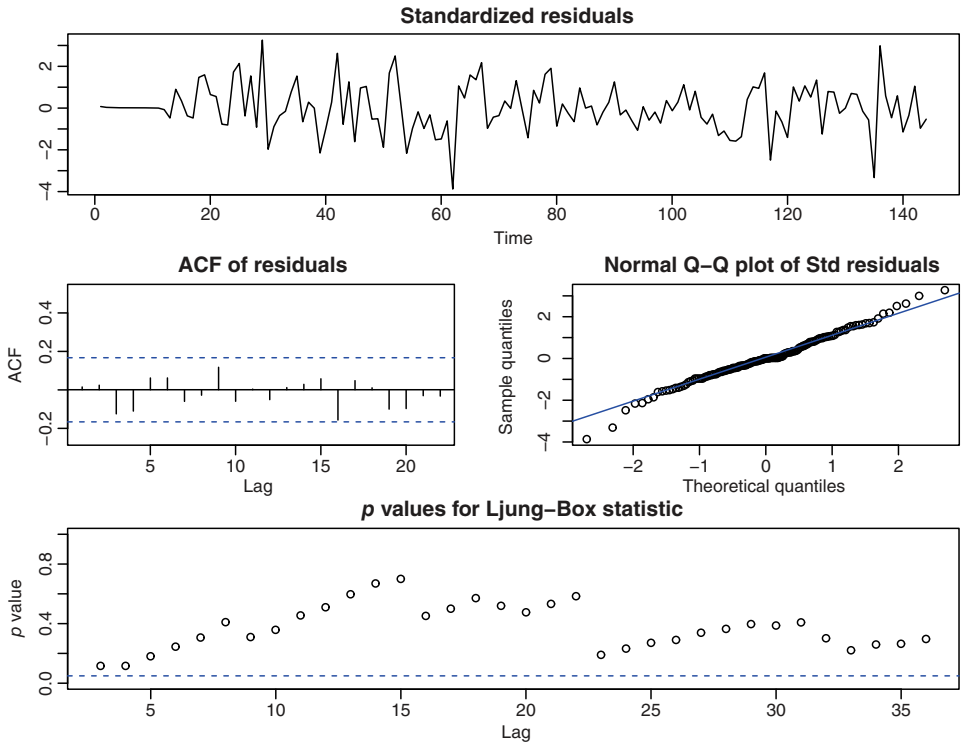

FIGURE 9.8 Diagnostic checks on the residuals from the fitted model.

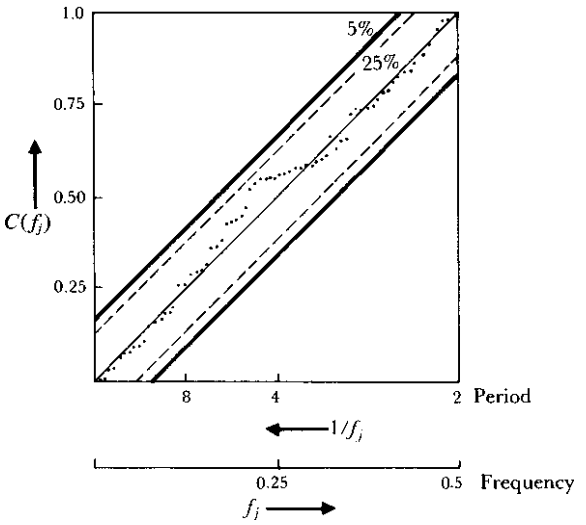

FIGURE 9.9 Cumulative periodogram check on residuals from the model $\nabla\nabla_{12}z_t = (1 - 0.40B)(1 - 0.61B^{12})a_t$, fitted to the airline data.

9.3 SOME ASPECTS OF MORE GENERAL SEASONAL ARIMA MODELS

9.3.1 Multiplicative and Nonmultiplicative Models

In previous sections, we discussed methods of dealing with seasonal time series, and in particular, we examined an example of a multiplicative model. We have seen how this model can provide a useful representation with remarkably few parameters. It now remains to study other seasonal models of this kind, and insofar as new considerations arise, the associated processes of identification, estimation, diagnostic checking, and forecasting.

Suppose, in general, that we have a seasonal effect associated with period s. Then, the general class of multiplicative models may be typified in the manner shown in Figure 9.10. In the multiplicative model, it is assumed that the ''between periods'' development of the series is represented by some model

$$\Phi_P(B^s)\nabla_s^D z_{r,m} = \Theta_Q(B^s)\alpha_{r,m}$$

while ''within periods'' the α's are related by

$$\phi_p(B)\nabla^d \alpha_{r,m} = \theta_q(B)a_{r,m}$$

Obviously, we could change the order in which we considered the two types of models and in either case obtain the general multiplicative model

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_{r,m} = \theta_q(B)\Theta_Q(B^s)a_{r,m} \tag{9.3.1}$$

where $a_{r,m}$ is a white noise process with zero mean. In practice, the usefulness of models such as (9.3.1) depends on how far it is possible to parameterize actual time series parsimoniously in these terms. In fact, experience has shown that this is possible for a variety of seasonal time series coming from widely different sources. While the multiplicative model (9.2.1)

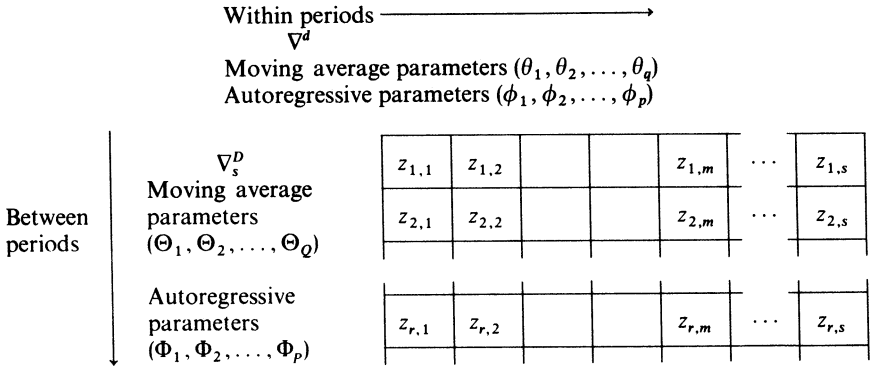

FIGURE 9.10 Two-way table for multiplicative seasonal model.

has been found to fit many time series, other models of the form (9.3.1) have also been found to be useful in practise.

It is not possible to obtain a completely adequate fit with multiplicative models for all series. One modification that is sometimes useful allows the mixed moving average operator to be nonmultiplicative. By this is meant that we replace the operator $\theta_q(B)\Theta_Q(B^s)$ on the right-hand side of (9.3.1) by a more general moving average operator $\theta^*_{q^*}(B)$. Alternatively, or in addition, it may be necessary to replace the autoregressive operator $\phi_p(B)\Phi_P(B^s)$ on the left by a more general autoregressive operator $\phi^*_{p^*}(B)$. Some examples of nonmultiplicative models are given in Appendix A9.1. These are numbered 4, 4a, 5, and 5a.

In those cases where a nonmultiplicative model is found necessary, experience suggests that the best-fitting multiplicative model can provide a good starting point from which to construct a better nonmultiplicative model. The situation is reminiscent of the problems encountered in analyzing two-way analysis of variance tables, where additivity of row and column constants may or may not be an adequate assumption, but may provide a good point of departure.

Our general strategy for relating multiplicative or nonmultiplicative models to data is that which we have already discussed and illustrated in some detail in Section 9.2. Using the autocorrelation function for guidance:

1. The series is differenced with respect to ∇ and/or ∇_s, so as to produce stationarity.
2. By inspection of the autocorrelation function of the suitably differenced series, a tentative model is selected.
3. From the values of appropriate autocorrelations of the differenced series, preliminary estimates of the parameters are obtained. These can be used as starting values in the search for the least-squares or maximum likelihood estimates.
4. After fitting, the diagnostic checking process applied to the residuals either may lead to the acceptance of the tentative model or, alternatively, may suggest ways in which it can be improved, leading to refitting and repetition of the diagnostic checks.

As a few practical guidelines for model specification, we note that for seasonal series the order of seasonal differencing D needed would almost never be greater than one, and especially for monthly series with $s = 12$, the orders P and Q of the seasonal AR and MA

operators $\Phi(B^s)$ and $\Theta(B^s)$ would rarely need to be greater than 1. This is particularly so when the series length of available data is not sufficient to warrant a more complicated form of model with $P > 1$ or $Q > 1$.

9.3.2 Model Identification

A useful aid in model identification is the list in Appendix A9.1 that gives the autocovariance structure of $w_t = \nabla^d \nabla_s^D z_t$ for a number of simple seasonal models. This list makes no claim to be comprehensive. However, it does include some frequently encountered models, and the reader should have no difficulty in discovering the characteristics of others that may seem useful. It should be emphasized that rather simple models, such as models 1 and 2 in the appendix, have provided adequate representations for many seasonal series.

Since the multiplicative seasonal ARMA models for the differences $w_t = \nabla\nabla_s z_t$ may be viewed as special forms of ARMA models with orders $p + sP$ and $q + sQ$, their autocovariances can be derived from the principles of Chapter 3, as was done in the previous section for the MA model $w_t = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta\Theta a_{t-13}$. For further illustration, consider the model

$$(1 - \phi B)w_t = (1 - \Theta B^s)a_t$$

which is a special form of ARMA model with AR order 1 and MA order s. First, since the ψ weights for this model for w_t satisfy $\psi_j - \phi\psi_{j-1} = 0, j = 1, \ldots, s-1$, we have $\psi_j = \phi^j, j = 1, \ldots, s-1$, as well as $\psi_s = \phi^s - \Theta$ and $\psi_j = \phi\psi_{j-1}, j > s$. It is then easy to see that the autocovariances for w_t will satisfy

$$\begin{aligned} \gamma_0 &= \phi\gamma_1 + \sigma_a^2(1 - \Theta\psi_s) \\ \gamma_j &= \phi\gamma_{j-1} - \sigma_a^2\Theta\psi_{s-j} \qquad j = 1, \ldots, s \\ \gamma_j &= \phi\gamma_{j-1} \qquad j > s \end{aligned} \tag{9.3.2}$$

Solving the first two equations for γ_0 and γ_1, we obtain

$$\gamma_0 = \sigma_a^2 \frac{1 - \Theta(\phi^s - \Theta) - \phi^s\Theta}{1 - \phi^2} = \sigma_a^2 \frac{1 + \Theta^2 - 2\phi^s\Theta}{1 - \phi^2}$$

$$\gamma_1 = \sigma_a^2 \frac{\phi[1 - \Theta(\phi^s - \Theta)] - \phi^{s-1}\Theta}{1 - \phi^2} = \sigma_a^2 \frac{\phi(1 + \Theta^2 - \phi^s\Theta) - \phi^{s-1}\Theta}{1 - \phi^2}$$

with $\gamma_j = \phi\gamma_{j-1} - \sigma_a^2\Theta\phi^{s-j} = \phi^j\gamma_0 - \sigma_a^2\Theta\phi^{s-j}(1 - \phi^{2j})/(1 - \phi^2), j = 1, \ldots, s$ and $\gamma_j = \phi\gamma_{j-1} = \phi^{j-s}\gamma_s, j > s$. Hence, in particular, for monthly data with $s = 12$ and $|\phi|$ not too close to one, the autocorrelation function ρ_j for this process will behave, for low lags, similarly to that of a regular AR(1) process, $\rho_j \simeq \phi^j$ for small j, while the value of ρ_{12} will be close to $-\Theta/(1 + \Theta^2)$.

A fact of considerable utility in deriving autocovariances of a multiplicative process is that for such a process, the autocovariance generating function (3.1.11) is the product or the generating functions of the components. Thus, in (9.3.1) if the component models for $\nabla^d z_t$ and $\nabla_s^D \alpha_t$,

$$\phi_p(B)\nabla^d z_t = \theta_q(B)\alpha_t \qquad \Phi_P(B^s)\nabla_s^D \alpha_t = \Theta_Q(B)a_t$$

have autocovariance generating function $\gamma(B)$ and $\Gamma(B^s)$, the autocovariance generating function for $w_t = \nabla^d \nabla_s^D z_t$ in (9.3.1) is

$$\gamma(B)\Gamma(B^s)$$

Another point to be remembered is that it may be useful to parameterize more general models in terms of their departures from related multiplicative forms in a manner now illustrated.

The three-parameter nonmultiplicative operator

$$1 - \theta_1 B - \theta_{12} B^{12} - \theta_{13} B^{13} \tag{9.3.3}$$

employed in models 4 and 5 in the appendix may be written as

$$(1 - \theta_1 B)(1 - \theta_{12} B^{12}) - k B^{13}$$

where

$$k = \theta_1 \theta_{12} - (-\theta_{13})$$

An estimate of k that was large compared with its standard error would indicate the need for a nonmultiplicative model in which the value of θ_{13} is not tied to the values of θ_1 and θ_{12}. On the other hand, if k is small, then on writing $\theta_1 = \theta$, $\theta_{12} = \Theta$, the model approximates the multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ model.

9.3.3 Parameter Estimation

No new problems arise in the estimation of the parameters of general seasonal models. The unconditional sum of squares is computed quite generally by the methods set out fully in Section 7.1.5 and illustrated further in Section 9.2.4. As always, contour plotting can illuminate difficult situations. In well-behaved situations, iterative least-squares with numerical determination of derivatives yield rapid convergence to the least-squares estimates, together with approximate variances and covariances of the estimates. Recursive procedures can be derived in each case, which allow direct calculation of derivatives, if desired.

Large-Sample Variances and Covariances of the Estimates. The large-sample information matrix $\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta})$ is given by evaluating $E[\mathbf{X}'\mathbf{X}]$, where, as in Section 7.2.6, $\mathbf{X}$ is the $n \times (p + q + P + Q)$ matrix of derivatives with reversed signs. Thus, for the general multiplicative model

$$a_t = \theta^{-1}(B)\Theta^{-1}(B^s)\phi(B)\Phi(B^s)w_t$$

where $w_t = \nabla^d \nabla_s^D z_t$, the required derivatives are

$$\frac{\partial a_t}{\partial \theta_i} = \theta^{-1}(B)B^i a_t \qquad \frac{\partial a_t}{\partial \Theta_i} = \Theta^{-1}(B^s)B^{si} a_t$$

$$\frac{\partial a_t}{\partial \phi_j} = -\phi^{-1}(B)B^j a_t \qquad \frac{\partial a_t}{\partial \Phi_j} = -\Phi^{-1}(B^s)B^{sj} a_t$$

Approximate variances and covariances of the estimates are obtained as before, by inverting the matrix $\mathbf{I}(\boldsymbol{\phi}, \theta, \boldsymbol{\Phi}, \boldsymbol{\Theta})$.

9.3.4 Eventual Forecast Functions for Various Seasonal Models

We now consider the characteristics of the eventual forecast functions for a number of seasonal models. For a seasonal model with single periodicity s, the eventual forecast function at origin t for lead time l is the solution of the difference equation

$$\phi(B)\Phi(B^s)\nabla^d \nabla_s^D \hat{z}_t(l) = 0$$

Table 9.2 shows this solution for various choices of the difference equation; also shown is the number of initial values on which the behavior of the forecast function depends.

In Figure 9.11, the behavior of each forecast function is illustrated for $s = 4$. It will be convenient to regard the lead time $l = rs + m$ as referring to a forecast r years and m quarters ahead. In the diagram, an appropriate number of initial values (required to start the forecast off and indicated by bold dots) has been set arbitrarily and the course of the forecast

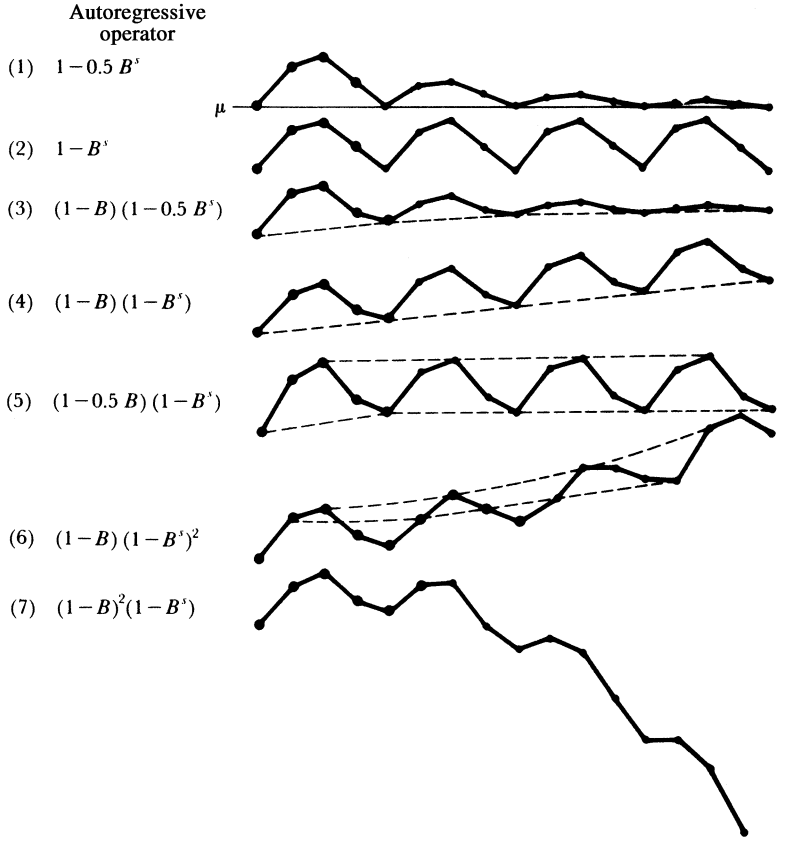

FIGURE 9.11 Behavior of the seasonal forecast function for various choices of the general seasonal autoregressive operator.

TABLE 9.2 Eventual Forecast Functions for Various Generalized Autoregressive Operators

Generalized Autoregressive Operator	Eventual Forecast Function $\hat{z}(r, m)^a$	Number of Initial Values on which Forecast Function Depends
(1) $1 - \Phi B^s$	$\mu + (b_{0,m} - \mu)\Phi^r$	s
(2) $1 - B^s$	$b_{0,m}$	s
(3) $(1 - B)(1 - \Phi B^s)$	$b_0 + (b_{0,m} - b_0)\Phi^r + b_1 \left\{ \frac{1 - \Phi^r}{1 - \Phi} \right\}$	$s + 1$
(4) $(1 - B)(1 - B^s)$	$b_{0,m} + b_1 r$	$s + 1$
(5) $(1 - \phi B)(1 - B^s)$	$b_{0,m} + b_1 \phi^{m-1} \left\{ \frac{1 - \phi^{sr}}{1 - \phi^s} \right\}$	$s + 1$
(6) $(1 - B)(1 - B^s)^2$	$b_{0,m} + b_{1,m} r + \frac{1}{2} b_2 r(r - 1)$	$2s + 1$
(7) $(1 - B)^2(1 - B^s)$	$b_{0,m} + [b_1 + (m - 1)b_2]r + \frac{1}{2} b_2 sr(r - 1)$	$s + 2$

[a]Coefficients b are all adaptive and depend upon forecast origin t.

function traced to the end of the fourth period. When the difference equation involves an autoregressive parameter, its value has been set equal to 0.5.

The constants $b_{0,m}$, b_1, and so on, appearing in the solutions in Table 9.2, should strictly be indicated by $b_{0,m}^{(t)}$, $b_1^{(t)}$, and so on, since each one depends on the origin t of the forecast, and these constants are adaptively modified each time the origin changes. The superscript t has been omitted temporarily to simplify notation.

The operator labeled (1) in Table 9.2 is stationary, with the model containing a fixed mean μ. It is autoregressive in the seasonal pattern, and the forecast function decays with each period, approaching closer and closer to the mean.

Operator (2) in Table 9.2 is nonstationary in the seasonal component. The forecasts for a particular quarter are linked from year to year by a polynomial of degree 0. Thus, the basic forecast of the seasonal component is exactly reproduced in forecasts of future years.

Operator (3) in Table 9.2 is nonstationary with respect to the basic time interval but stationary in the seasonal component. Operator (3) in Figure 9.11 shows the general level of the forecast approaching asymptotically the new level

$$b_0 + \frac{b_1}{1 - \Phi}$$

where, at the same time, the superimposed predictable component of the stationary seasonal effect dies out exponentially.

In Table 9.2, operator (4) is the limiting case of the operator (3) as Φ approaches unity. The operator is nonstationary with respect to both the basic time interval and the periodic component. The basic initial forecast pattern is reproduced, as is the incremental yearly increase. This is the type of forecast function given by the multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ process fitted to the airline data.

Operator (5) is nonstationary in the seasonal pattern but stationary with respect to the basic time interval. The pattern approaches exponentially an asymptotic basic pattern

$$\hat{z}_t(\infty, m) = b_{0,m} + \frac{b_1 \phi^{m-1}}{1 - \phi^s}$$

Operator (6) is nonstationary in both the basic time interval and the seasonal component. An overall quadratic trend occurs over years, and a particular kind of modification occurs in the seasonal pattern. Individual quarters not only have their own level $b_{0,m}$ but also their own rate of change of level $b_{1,m}$. Therefore, when this kind of forecast function is appropriate, we can have a situation where, for example, as the lead time is increased, the difference in summer over spring sales can be forecast to increase from one year to the next, while at the same time, the difference in autumn over summer sales can be forecast to decrease.

In Table 9.2, operator (7) is again nonstationary in both the basic time interval and in the seasonal component, and there is again a quadratic tendency over years with the incremental changes in the forecasts from one quarter to the next changing linearly. However, in this case, they are restricted to have a common *rate* of change.

9.3.5 Choice of Transformation

It is particularly true for seasonal models that the weighted averages of previous data values, which comprise the forecasts, may extend far back into the series. Care is therefore needed in choosing a transformation in terms or which a parsimonious linear model will closely apply over a sufficient stretch of the series. Simple graphical analysis can often suggest such a transformation. Thus, an appropriate transformation may be suggested by determining in what metric the amplitude of the seasonal component is roughly independent of the level of the series. To illustrate how a data-based transformation may be chosen more exactly, denote the *untransformed* airline data by x, and let us assume that some power transformation [$z = x^\lambda$ for $\lambda \neq 0$, $z = \ln(x)$ for $\lambda = 0$] may be needed to make the model (9.2.1) appropriate. Then, as suggested in Section 4.1.3, the approach of Box and Cox (1964) may be followed, and the maximum likelihood value obtained by fitting the model to $x^{(\lambda)} = (x^\lambda - 1)/\lambda \dot{x}^{\lambda-1}$ for various values of λ, and choosing the value of λ that results in the smallest residual sum of squares s_λ. In this expression, $\dot{x}$ is the geometric mean of the series x, and it is easily shown that $x^{(0)} = \dot{x} \ln(x)$. For the airline data, we find

λ	S_λ	λ	S_λ	λ	S_λ
−0.4	13,825.5	−0.1	11,627.2	0.2	11,784.3
−0.3	12,794.6	0.0	11,458.1	0.3	12,180.0
−0.2	12,046.0	0.1	11,554.3	0.4	12,633.2

The maximum likelihood value is thus close to $\lambda = 0$, confirming the appropriateness of the logarithmic transformation for the airline series.

9.4 STRUCTURAL COMPONENT MODELS AND DETERMINISTIC SEASONAL COMPONENTS

A traditional method to represent a seasonal time series has been to decompose the series into trend, seasonal, and noise components, as $z_t = T_t + S_t + N_t$, where the trend T_t and seasonal component S_t are represented as deterministic functions of time using polynomial and sinusoidal functions, respectively. However, as noted in Section 9.1.1, the deterministic nature of the trend and seasonal components limits the applicability of these models.

Subsequently, models that permit random variation in the trend and seasonal components, referred to as *structural component* models, have become increasingly popular for time series modeling (e.g., Harvey, 1989; Harvey and Todd, 1983; Gersch and Kitagawa, 1983; Kitagawa and Gersch, 1984; Hillmer and Tiao, 1982; and Durbin and Koopman, 2012). We discuss these models briefly in the following sections.

9.4.1 Structural Component Time Series Models

In general, a univariate structural component time series model is one in which an observed series z_t is formulated as the sum of unobservable component or "signal" time series. Although the components are unobservable and cannot be uniquely specified, they will usually have direct meaningful interpretation, such as representing the seasonal behavior or the long-term trend of an economic time series or a physical signal that is corrupted by measurement noise in the engineering setting. Thus, the models attempt to describe the main features of the series as well as provide a basis for forecasting, signal extraction, seasonal adjustments, and other applications. For a monthly time series, the trend T_t might be assumed to follow a simple random walk model or some extension such as the ARIMA(0, 1, 1) model $(1 - B)T_t = (1 - \theta B)a_t$, or the AIRMA(0, 2, 2) model $(1 - B)^2 T_t = (1 - \theta_1 B - \theta_2 B^2)a_t$, while the seasonal component might be specified as a ''seasonal random walk'' $(1 - B^{12})S_t = b_t$, where a_t and b_t are independent white noise processes.

An appeal of this structural modeling approach, especially for seasonal adjustments and signal extraction, is that Kalman filtering and smoothing methods based on state-space formulations of the model, as discussed in Section 5.5, can be employed. The exact likelihood function can be constructed based on the state-space model form, as described in Section 7.4, and used for parameter estimation. The Kalman filtering and smoothing procedures can then be used to obtain estimates of the unobservable component series such as the trend $\{T_t\}$ and seasonal $\{S_t\}$ components, which are now included as elements within the state vector Y_t in the general state-space model (5.5.4) and (5.5.5).

Basic Structural Model. As a specific illustration, consider the *basic structural model* (BSM) for seasonal time series with period s as formulated by Harvey (1989). The model is defined by $z_t = T_t + S_t + \varepsilon_t$, where T_t follows the ''local linear trend model'' defined by

$$T_t = T_{t-1} + \beta_{t-1} + \eta_t \quad \beta_t = \beta_{t-1} + \xi_t \tag{9.4.1}$$

and S_t follows the ''dummy variable seasonal component model'' defined by

$$(1 + B + B^2 + \ldots + B^{s-1})S_t = \omega_t \tag{9.4.2}$$

where η_t, ξ_t, ω_t, and ε_t are mutually uncorrelated white noise processes with zero means and variances σ_η^2, σ_ξ^2, σ_ω^2, and σ_ε^2, respectively.

This local linear trend model is a stochastic generalization of the deterministic linear trend $T_t = \alpha + \beta t$, where α and β are constants. In (9.4.1), the effect of the random disturbance η_t is to allow the level of the trend to shift up and down, while ξ_t allows the slope to change. As special limiting cases, if $\sigma_\xi^2 = 0$, then $\beta_t = \beta_{t-1}$ and so β_t is a fixed constant β for all t and the trend follows the random walk with drift $(1 - B)T_t = \beta + \eta_t$. If $\sigma_\eta^2 = 0$ in addition, then (9.4.1) collapses to the deterministic model $T_t = T_{t-1} + \beta$ or $T_t = \alpha + \beta t$. The seasonal component model (9.4.2) requires the seasonal effects S_t to sum to zero over

s consecutive values of a seasonal period, subject to a random disturbance with mean zero which allows the seasonal effects to change gradually over time. Again, a special limiting case of deterministic seasonal components with a fixed seasonal pattern about an average of zero, $S_t = S_{t-s}$ with $S_t + S_{t-1} + \ldots + S_{t-s+1} = 0$, occurs when $\sigma_\omega^2 = 0$. Thus, one attraction of a model such as (9.4.1) and (9.4.2) is that it generalizes a regression-type in which the trend is represented by a fixed straight line and the seasonality by fixed seasonal effects using indicator variables, by allowing the trend and seasonality to vary over time, and still yields the deterministic components as special limiting cases.

We illustrate the state-space representation of the model (9.4.1) and (9.4.2) for the case of quarterly time series with $s = 4$. For this, we define the state vector as

$$Y_t = (T_t, \beta_t, S_t, S_{t-1}, S_{t-2})'$$

and let $\mathbf{a}_t = (\eta_t, \xi_t, \omega_t)'$. Then we have the transition equation

$$Y_t = \begin{bmatrix} T_t \\ \beta_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} T_{t-1} \\ \beta_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \eta_t \\ \xi_t \\ \omega_t \end{bmatrix} \tag{9.4.3}$$

or $\boldsymbol{Y}_t = \boldsymbol{\Phi} Y_{t-1} + \Psi \mathbf{a}_t$, together with the observation equation $z_t = T_t + S_t + \varepsilon_t \equiv [1 \quad 0 \quad 1 \quad 0 \quad 0] Y_t + \varepsilon_t = \mathbf{H} Y_t + \varepsilon_t$. Hence, the variance component parameters of the structural model can be estimated by maximum likelihood methods using the state-space representation and innovations form of the likelihood function, as discussed in Sections 5.5 and 7.4. Once these estimates are obtained, the desired optimal smoothed estimates $\hat{T}_{t|n} = E[T_t | z_1, \ldots, z_n]$ and $\hat{S}_{t|n} = E[S_t | z_1, \ldots, z_n]$ of the trend and seasonal components based on the observed series $z_1, \ldots, z_n$ can readily be obtained by applying the Kalman filtering and smoothing techniques to the state-space representation.

Relation to ARIMA Model. It should be noted from general results of Appendix A4.3 that structural models such as the BSM have an equivalent ARIMA model representation, which is sometimes referred to as its *reduced form* in this context. For instance, the process T_t defined by the local linear trend model (9.4.1) satisfies

$$(1 - B)^2 T_t = (1 - B)\beta_{t-1} + (1 - B)\eta_t = \xi_{t-1} + (1 - B)\eta_t$$

It follows from Appendix A4.3.1 that $\xi_{t-1} + (1 - B)\eta_t$ can be represented as an MA(1) process $(1 - \theta B)a_t$, so that $(1 - B)^2 T_t = (1 - \theta B)a_t$ and T_t has the ARIMA(0, 2, 1) model as a reduced form. For another illustration, consider $z_t = T_t + S_t + N_t$, where it is assumed that

$$(1 - B)T_t = (1 - \theta_T B)a_t \quad (1 - B^{12})S_t = (1 - \Theta_s B^{12})b_t$$

and $N_t = c_t$ is white noise. Then, we have

$$\begin{aligned} &(1 - B)(1 - B^{12})z_t \\ &\quad = (1 - B^{12})(1 - \theta_T B)a_t + (1 - B)(1 - \Theta_s B^{12})b_t + (1 - B)(1 - B^{12})c_t \end{aligned}$$

and according to the developments in Appendix A4.3, the right-hand-side expression above can be represented as the MA model $(1 - \theta_1 B - \theta_{12} B^{12} - \theta_{13} B^{13})\epsilon_t$, where ϵ_t is white noise, since the right-hand side will have nonzero autocovariances only at the lags 0, 1, 11, 12, and 13. Under additional structure, the MA operator could have the multiplicative form, but in general we see that the foregoing structural model, $z_t = T_t + S_t + N_t$, has an equivalent ARIMA model representation as

$$(1 - B)(1 - B^{12})z_t = (1 - \theta_1 B - \theta_{12} B^{12} - \theta_{13} B^{13})\epsilon_t$$

Example: Airline Data. Harvey (1989, Sec. 4.5) reported results of maximum likelihood estimation of the BSM defined by (9.4.1) and (9.4.2) for the logged monthly airline passenger data, using the data period from 1949 to 1958. The ML estimates were such that $\hat{\sigma}_\xi^2 = 0$ and $\hat{\sigma}_\omega^2$ was very small relative to $\hat{\sigma}_\eta^2$ and $\hat{\sigma}_\epsilon^2$. The zero estimate $\hat{\sigma}_\xi^2 = 0$ implies that the model (9.4.1) for the trend T_t reduces to the random walk with *constant* drift, $(1 - B)T_t = \beta + \eta_t$, while the seasonal component model is $(1 + B + \ldots + B^{11})S_t = \omega_t$. Differencing the series z_t thus implies that

$$\begin{aligned} w_t = (1 - B)(1 - B^{12})z_t &= (1 - B)(1 - B^{12})T_t + (1 - B)(1 - B^{12})S_t \\ &\quad + (1 - B)(1 - B^{12})\epsilon_t \\ &= (1 - B^{12})\eta_t + (1 - B)^2\omega_t + (1 - B)(1 - B^{12})\epsilon_t \end{aligned}$$

It readily follows that the autocovariances of the differenced series $w_t = \nabla\nabla_{12} z_t$ for this model are

$$\begin{aligned} \gamma_0 &= 2\sigma_\eta^2 + 6\sigma_\omega^2 + 4\sigma_\epsilon^2 \\ \gamma_1 &= -4\sigma_\omega^2 - 2\sigma_\epsilon^2 \\ \gamma_2 &= \sigma_\omega^2 \\ \gamma_{11} &= \sigma_\epsilon^2 = \gamma_{13} \\ \gamma_{12} &= -\sigma_\eta^2 - 2\sigma_\epsilon^2 \end{aligned} \tag{9.4.4}$$

and $\gamma_j = 0$ otherwise. In particular, these give the autocorrelations

$$\rho_1 = -\frac{\sigma_\epsilon^2 + 2\sigma_\omega^2}{2\sigma_\epsilon^2 + \sigma_\eta^2 + 3\sigma_\omega^2}$$

$$\rho_{12} = \frac{2\sigma_\epsilon^2 + \sigma_\eta^2}{2(2\sigma_\eta^2 + \sigma_\eta^2 + 3\sigma_\omega^2)}$$

and $\rho_{11} = \rho_{13} = \sigma_\epsilon^2/[2(2\sigma_\epsilon^2 + \sigma_\eta^2 + 3\sigma_\omega^2)]$.

The autocorrelations calculated using estimates of the variance components given in Table 4.5.3 of Harvey (1989) are shown in Table 9.3 for the logged airline data. Also shown in Table 9.3 are the autocorrelations for the differenced series $w_t = \nabla\nabla_{12} z_t$ in the seasonal $(0, 1, 1) \times (0, 1, 1)_{12}$ model. These were calculated from (9.2.18) using the parameter estimates $\hat{\theta} = 0.396$, $\hat{\Theta} = 0.614$, and $\hat{\sigma}_a^2 = 1.34 \times 10^{-3}$ reported in Section 9.2.4. Table 9.3 shows a close agreement between the two sets of autocorrelations. Hence, for the logged airline data, both modeling approaches provide very similar representations of the basic trend and seasonality in the series.

TABLE 9.3 Comparison of the Autocorrelations of $w_t = \nabla\nabla_{12} z_t$ for the Basic Structural Model and the Seasonal ARIMA Model $(0,1,1) \times (0,1,1)_{12}$ for Logged Airline Data

Model	ρ_1	ρ_2	ρ_{11}	ρ_{12}	ρ_{13}
Basic structural model	−0.26	0.00	0.12	−0.49	0.12
ARIMA $(0,1,1) \times (0,1,1)_{12}$	−0.34	0.00	0.15	−0.45	0.15

9.4.2 Deterministic Seasonal and Trend Components and Common Factors

Now in some applications, particularly in the physical sciences, a seasonal or trend component could be nearly deterministic. For example, suppose the seasonal component can be approximated as

$$S_t = \beta_0 + \sum_{j=1}^{6} \left[\beta_{1j} \cos\left(\frac{2\pi jt}{12}\right) + \beta_{2j} \sin\left(\frac{2\pi jt}{12}\right) \right]$$

where the β coefficients are constants. We note that this can be viewed as a special case of the previous examples, since S_t satisfies $(1 + B + B^2 + \ldots + B^{11})S_t = 12\beta_0$ or $(1 - B^{12})S_t = 0$. Now, ignoring the trend component for the present and assuming that $z_t = S_t + N_t$, where $(1 - B^{12})S_t = 0$ and $N_t = (1 - \theta_N B)a_t$, say, we find that z_t follows the seasonal ARIMA model

$$(1 - B^{12})z_t = (1 - \theta_N B)(1 - B^{12})a_t$$

However, we now notice the presence of a *common factor* of $1 - B^{12}$ in both the generalized AR operator and the MA operator of this model; equivalently, we might say that $\Theta = 1$ for the seasonal MA operator $\Theta(B^{12}) = (1 - \Theta B^{12})$. This is caused by and, in fact, is indicative of the presence of the deterministic seasonal component S_t in the original form of the model.

In general, the presence of deterministic seasonal or trend components in the structure of a time series z_t is characterized by common factors of $(1 - B^s)$ or $(1 - B)$ in the generalized AR operator and the MA operator of the model. We can state the result more formally as follows. Suppose that z_t follows the model $\varphi(B)z_t = \theta_0 + \theta(B)a_t$, and the operators $\varphi(B)$ and $\theta(B)$ contain a common factor $G(B)$, so that $\varphi(B) = G(B)\varphi_1(B)$ and $\theta(B) = G(B)\theta_1(B)$. Hence, the model is

$$G(B)\varphi_1(B)z_t = \theta_0 + G(B)\theta_1(B)a_t \tag{9.4.5}$$

Let $G(B) = 1 - g_1 B - \cdots - g_r B^r$ and suppose that this polynomial has roots $G_1^{-1}, \ldots, G_r^{-1}$ which are distinct. Then, the common factor $G(B)$ can be canceled from both sides of the above model, but a term of the form $\sum_{i=1}^{r} c_i G_i^t$ needs to be added. Thus, the model (9.4.5) can be expressed in the equivalent form as

$$\varphi_1(B)z_t = c_{0t} + \sum_{i=1}^{r} c_i G_i^t + \theta_1(B)a_t \tag{9.4.6}$$

where the c_i are constants, and c_{0t} is a term that satisfies $G(B)c_{0t} = \theta_0$. Modifications of the result for the case where some of the roots G_i^{-1} are repeated are straightforward.

Thus, it is seen that an equivalent representation for the above model is

$$\varphi_1(B)z_t = x_t + \theta_1(B)a_t$$

where x_t is a deterministic function of t that satisfies $G(B)x_t = \theta_0$. Note that roots in $G(B)$ corresponding to ''stationary factors,'' such that $|G_i| < 1$, will make a contribution to the component x_t that is only transient and so negligible, and hence these terms may be ignored. Thus, only those factors whose roots correspond to nonstationary ''differencing'' and other ''simplifying'' operators, such as $(1 - B)$ and $(1 - B^s)$, with roots $|G_i| = 1$ need to be included in the deterministic component x_t. These common factors will, of course, give rise to deterministic functions in x_t that are of the form of polynomials, sine and cosine functions, and products of these, depending on the roots of the common factor $G(B)$.

Examples. For a few simple examples, the model $(1 - B)z_t = \theta_0 + (1 - B)\theta_1(B)a_t$ has an equivalent form $z_t = c_1 + \theta_0 t + \theta_1(B)a_t$, which occurs upon cancellation of the common factor $(1 - B)$, while the model $(1 - \sqrt{3}B + B^2)z_t = \theta_0 + (1 - \sqrt{3}B + B^2)\theta_1(B)a_t$ has an equivalent model form as $z_t = c_0 + c_1 \cos(2\pi t/12) + c_2 \sin(2\pi t/12) + \theta_1(B)a_t$, where $(1 - \sqrt{3} + 1)c_0 = \theta_0$.

Detection of a deterministic component such as x_t above in a time series z_t may occur after an ARIMA model is estimated and common or near-common factors are identified. Hence, the ARIMA time series methodology, in a sense, can indicate when a time series may contain deterministic seasonal or trend components. The presence of a deterministic component is characterized by a factor in the MA operator with roots on, or very near to the unit circle, which correspond to a differencing factor that has been applied to the original series in the formulation of the ARIMA model. When this situation occurs, the series is sometimes said to be ''over-differenced''. Formal tests for the presence of a unit root in the MA operator implying the presence of a deterministic component, have been developed by Saikkonen and Luukkonen (1993), Leybourne and McCabe (1994), and Tam and Reinsel (1997, 1998), among others. These tests can also be viewed as tests for unit roots in the generalized AR operator $\varphi(B)$ in the sense that if one performs the differencing and then concludes that the MA operator does not have a unit root, then the unit root in the AR operator is supported.

Deterministic components implied by the cancellation of factors could be estimated directly by a combination of regression models and ARIMA time series methods, as will be discussed in Section 9.5. An additional consequence of the presence of deterministic factors for forecasting is that at least some of the coefficients $b_j^{(t)}$ in the general forecast function $\hat{z}_t(l)$ for z_{t+l} in (5.3.3) will not be adaptive but will be deterministic (fixed) constants. Results such as those described above concerning the relationship between common factors in the generalized AR and the MA operators of ARIMA models and the presence of deterministic polynomial and sinusoidal components have been discussed by Abraham and Box (1978), Harvey (1981), and Bell (1987).

9.4.3 Estimation of Unobserved Components in Structural Models

A common problem of interest for the structural model is the estimation of the unobservable series S_t from values of the observed series z_t. We suppose that S_t and z_t are stationary processes with zero means and autocovariance functions $\gamma_s(l) = E[S_t S_{t+1}]$ and $\gamma_z(l) = E[z_t z_{t+1}]$, and cross-covariance function $\gamma_{sz}(l) = E[S_t z_{t+1}]$. Then, specifically, suppose

we observe the values $z_t, t \le \tau$, and want to determine the linear filter

$$\hat{S}_t = \sum_{u=0}^{\infty} v_u^{(\tau)} z_{\tau-u} \equiv v^{(\tau)}(B) z_\tau \tag{9.4.7}$$

of $\{z_t\}$ such that the value $\hat{S}_t$ is close to S_t in the mean square error sense, that is, $E[(S_t - \hat{S}_t)^2]$ is a minimum among all possible linear filters. A typical model for which this problem arises is the ''signal extraction'' model, in which there is a signal S_t of interest, but what is observed is a noise-corrupted version of the signal so that

$$z_t = S_t + N_t$$

where N_t is a noise component. The problem then is to estimate values of the signal series S_t given values on the observed series z_t. Often, the filtering and smoothing algorithms for the state-space model, as discussed in Section 5.5.3, can be applied to this situation. However, while these algorithms are computationally attractive in practice, explicit expressions for the coefficients $v_u^{(\tau)}$ in (9.4.7) cannot usually be obtained directly from the state-space algorithms. These expressions can be derived more readily in the ''classical'' approach, which assumes that an infinite extent of observations is available for filtering or smoothing. This section provides a brief overview of some classical filtering and smoothing results that can be used to study the coefficients in (9.4.7). Typically, from a practical point of view, the classical results provide a good approximation to exact filtering and smoothing results that are based on a finite sample of observations $z_1, \ldots, z_n$.

Smoothing and Filtering for Time Series. We suppose that $\{z_t\}$ has the infinite MA representation

$$z_t = \psi(B) a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

where the a_t are white noise with variance σ_a^2. Also, let $g_{zs}(B) = \sum_{j=-\infty}^{\infty} \gamma_{zs}(j) B^j$ be the cross-covariance generating function between z_t and S_t. Then, it can be derived (e.g., Whittle, 1963, Chapters 5 and 6; Priestley, 1981, Chapter 10) that the optimal linear filter for the estimate $\hat{S}_t = \sum_{u=0}^{\infty} v_u^{(\tau)} z_{\tau-u} = v^{(\tau)}(B) z_\tau$, where $v^{(\tau)}(B) = \sum_{u=0}^{\infty} v_u^{(\tau)} B^u$, is given by

$$v^{(\tau)}(B) = \frac{1}{\sigma_a^2 \psi(B)} \left[\frac{B^{\tau-t} g_{zs}(B)}{\psi(B^{-1})} \right]_+ \tag{9.4.8}$$

Here, for a general operator $v(B) = \sum_{j=-\infty}^{\infty} v_j B^j$, the notation $[v(B)]_+$ is used to denote $\sum_{j=0}^{\infty} v_j B^j$.

To derive the result (9.4.8) for the optimal linear filter, note that, since $z_t = \psi(B) a_t$, the linear filter can be expressed as

$$\hat{S}_t = v^{(\tau)}(B) z_\tau = v^{(\tau)}(B) \psi(B) a_\tau = h^{(\tau)}(B) a_\tau$$

where $h^{(\tau)}(B) = v^{(\tau)}(B)\psi(B) = \sum_{j=0}^{\infty} h_j^{(\tau)} B^j$. Then, we can determine the coefficients $h_j^{(\tau)}$ to minimize the mean squared error $E[(S_t - \hat{S}_t)^2] = E[(S_t - \sum_{j=0}^{\infty} h_j^{(\tau)} a_{\tau-j})^2]$. Since the $\{a_t\}$ are mutually uncorrelated, by standard linear least-squares arguments the values of

the coefficients that minimize this mean squared error are

$$h_j^{(\tau)} = \frac{\text{cov}[a_{\tau-j}, S_t]}{\text{var}[a_{\tau-j}]} = \frac{\gamma_{as}(j+t-\tau)}{\sigma_a^2} \qquad j \geq 0$$

Hence, the optimal linear filter is

$$h^{(\tau)}(B) = \frac{1}{\sigma_a^2} \sum_{j=0}^{\infty} \gamma_{as}(j+t-\tau) B^j = \frac{1}{\sigma_a^2} [B^{\tau-t} g_{as}(B)]_+ \tag{9.4.9}$$

where $g_{as}(B)$ denotes the cross-covariance generating function between a_t and S_t. Also, note that $\gamma_{zs}(j) = \text{cov}[\sum_{i=0}^{\infty} \psi_i a_{t-i}, S_{t+j}] = \sum_{i=0}^{\infty} \psi_i \gamma_{as}(i+j)$, so it follows that $g_{zs}(B) = \psi(B^{-1}) g_{as}(B)$. Therefore, the optimal linear filter in (9.4.9) is $h^{(\tau)}(B) = (1/\sigma_a^2)[B^{\tau-t} g_{zs}(B)/\psi(B^{-1})]_+$, and, hence, the optimal filter in terms of $\hat{S}_t = \upsilon^{(\tau)}(B) z_\tau$ is $\upsilon^{(\tau)}(B) = h^{(\tau)}(B)/\psi(B)$, which yields the result (9.4.8). The mean squared error of the optimal filter, since $\hat{S}_t = \sum_{j=0}^{\infty} h_j^{(\tau)} a_{\tau-j}$, is easily seen from the above derivation to be

$$E[(S_t - \hat{S}_t)^2] = E[S_t^2] - E[\hat{S}_t^2] = \text{var}[S_t] - \sigma_a^2 \sum_{j=0}^{\infty} \{h_j^{(\tau)}\}^2$$

In the smoothing case where $\tau = +\infty$, that is, we estimate S_t based on the infinite record of observations z_u, $-\infty < u < \infty$, by a linear filter $\hat{S}_t = \sum_{u=-\infty}^{\infty} \upsilon_u z_{t-u} = \upsilon(B) z_t$, the result (9.4.8) for the optimal filter reduces to

$$\upsilon(B) = \frac{g_{zs}(B)}{g_{zz}(B)} = \frac{g_{zs}(B)}{\sigma_a^2 \psi(B) \psi(B^{-1})} \tag{9.4.10}$$

For the signal extraction problem, we have $z_t = S_t + N_t$, where it is usually assumed that the signal $\{S_t\}$ and the noise process $\{N_t\}$ are independent. Thus, in this case we have $g_{zs}(B) = g_{ss}(B)$, and so in the smoothing case $\tau = +\infty$, we have $\upsilon(B) = g_{ss}(B)/g_{zz}(B)$ or $\upsilon(B) = g_{ss}(B)/[g_{ss}(B) + g_{nn}(B)]$.

Smoothing Relations for the Signal Plus Noise or Structural Components Model. The preceding results can be applied specifically to the model $z_t = S_t + N_t$, where we assume that the signal process $\{S_t\}$ and the noise process $\{N_t\}$ are independent and satisfy ARMA models, $\phi_s(B) S_t = \theta_s(B) b_t$ and $\phi_n(B) N_t = \theta_n(B) c_t$, where b_t and c_t are independent white noise processes with variances σ_b^2 and σ_c^2. It follows from Appendix A4.3 that the observed process z_t also satisfies an ARMA model $\phi(B) z_t = \theta(B) a_t$, where $\phi(B) = \phi_s(B) \phi_n(B)$, assuming no common factors in the AR operators. It then follows that the optimal linear ''smoother'' $\hat{S}_t = \sum_{u=-\infty}^{\infty} \upsilon_u z_{t-u} = \upsilon(B) z_t$ of S_t, based on the infinite set of values z_u, $-\infty < u < \infty$, has a filter given by

$$\upsilon(B) = \frac{g_{ss}(B)}{g_{zz}(B)} = \frac{\sigma_b^2 \phi(B) \phi(B^{-1}) \theta_s(B) \theta_s(B^{-1})}{\sigma_a^2 \theta(B) \theta(B^{-1}) \phi_s(B) \phi_s(B^{-1})} \tag{9.4.11}$$

In practice, since the series S_t and N_t are not observable, the models for S_t and N_t would usually not be known. Thus, the optimal filter would not be known in practice. However, by

developing a model for the observed series z_t and placing certain restrictions on the form of the models for S_t and N_t beyond those implied by the model for z_t, e.g., by assuming N_t is white noise with the largest possible variance, one may obtain reasonable approximations to the optimal filter $v(B)$. While optimal smoothing results, such as (9.4.10), have been derived for the case where S_t and N_t are stationary processes, Bell (1984) showed that the results extend to the nonstationary case under reasonable assumptions for the nonstationary signal S_t and noise N_t processes.

As noted earlier, an alternative to the classical filtering approach in the structural components models is to express the model in state-space form and use Kalman filtering and smoothing techniques to estimate the components, as illustrated, for example, by Kitagawa and Gersch (1984). For further discussion of this approach, see also Harvey (1989) and Durbin and Koopman (2012).

Seasonal Adjustments. The filtering and smoothing methods described above have applications to seasonal adjustments of economic and business time series (i.e., estimating and removing the seasonal component from the series). Approaches of the type discussed were used by Hillmer and Tiao (1982) to decompose a time series uniquely into mutually independent seasonal, trend, and irregular components. A model-based approach to seasonal adjustments was also considered by Cleveland and Tiao (1976). Seasonal adjustments are commonly performed by statistical agencies in the U.S. and elsewhere, and the methods used have received considerable attention in the literature. For an overview and further discussion, see, for example, Ghysels and Osborn (2001, Chapter 4), Bell and Sotiris (2010), Chu, Tiao, and Bell (2012), and Bell, Chu, and Tiao (2012).

9.5 REGRESSION MODELS WITH TIME SERIES ERROR TERMS

The previous discussion of deterministic components in Section 9.4.2 motivates consideration of time series models that include regression terms such as deterministic sine and cosine functions to represent seasonal behavior or stochastic predictor variables, in addition to a serially correlated ''noise'' or error term. We will assume that the noise series N_t follows a stationary ARMA process; otherwise, differencing may be need to be considered. Thus, letting w_t be a ''response'' series of interest, we wish to represent w_t in terms of its linear dependence on k explanatory or predictor time series variables $x_{t1}, \ldots, x_{tk}$ as follows:

$$w_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + N_t \qquad t = 1, \ldots, n \tag{9.5.1}$$

where the errors N_t follow a zero-mean ARMA(p, q) model, $\phi(B)N_t = \theta(B)a_t$. The traditional linear regression model was reviewed briefly in Appendix A7.2. Using similar notations with $\mathbf{w} = (w_1, \ldots, w_n)'$, $\boldsymbol{N} = (N_1, \ldots, N_n)'$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, the model (9.5.1) may be written in matrix form as $\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{N}$, and with covariance matrix $\mathbf{V} = \text{cov}[\boldsymbol{N}]$. In the standard regression model, the errors N_t are assumed to be uncorrelated with common variance σ_N^2, so that $\mathbf{V} = \sigma_N^2\mathbf{I}$, and the ordinary least squares (LS) estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}$ has well-known properties such as $\text{cov}[\hat{\boldsymbol{\beta}}] = \sigma_N^2(\mathbf{X}'\mathbf{X})^{-1}$. However, in the case of autocorrelated errors, this property no longer holds and the ordinary least-squares estimator has covariance matrix

$$\text{cov}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Moreover, standard inference procedures based on the t and F distributions are no longer valid due to the lack of independence.

When $\text{cov}[\boldsymbol{N}] = \mathbf{V} \neq \sigma_N^2 \mathbf{I}$, the best linear unbiased estimator of $\boldsymbol{\beta}$ is the *generalized least-squares* (GLS) estimator given by

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{w} \tag{9.5.2}$$

which has $\text{cov}[\hat{\boldsymbol{\beta}}_G] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. The estimator $\hat{\boldsymbol{\beta}}_G$ is the best linear unbiased estimator in the sense that $\text{var}[\boldsymbol{c}'\hat{\boldsymbol{\beta}}_G]$ is a minimum among all possible linear unbiased estimators of $\boldsymbol{\beta}$, for every arbitrary k-dimensional vector of constants $c' = (c_1, \ldots, c_k)$; in particular, $\text{var}[c'\hat{\boldsymbol{\beta}}_G] \leq \text{var}[c'\hat{\boldsymbol{\beta}}]$ holds relative to the ordinary LS estimator $\hat{\boldsymbol{\beta}}$. It follows that $\hat{\boldsymbol{\beta}}_G$ in (9.5.2) is the estimate of $\boldsymbol{\beta}$ obtained by minimizing the generalized sum of squares $S(\boldsymbol{\beta}; \mathbf{V}) = (\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})$ with $\mathbf{V}$ given. This estimator also corresponds to the maximum likelihood estimator under the assumption of normality of the errors when the covariance matrix $\mathbf{V}$ is known. Of course, a practical limitation to use of the GLS estimate $\hat{\boldsymbol{\beta}}_G$ is that the ARMA noise model and its parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ needed to determine $\mathbf{V}$ must be known, which is typically not true in practice. This motivates an iterative model building and estimation procedure discussed below.

9.5.1 Model Building, Estimation, and Forecasting Procedures for Regression Models

When a regression model is fitted to time series data, one should always consider the possibility that the errors are autocorrelated. Often, a reasonable approach to identify an appropriate model for the error N_t is first to obtain the least-squares estimate $\hat{\boldsymbol{\beta}}$, and then compute the corresponding regression model residuals

$$\hat{N}_t = w_t - \hat{\beta}_1 x_{t1} - \hat{\beta}_2 x_{t2} - \cdots - \hat{\beta}_k x_{tk} \tag{9.5.3}$$

This residual series can be examined by the usual time series methods, such as inspection of its sample ACF and PACF, to identify an appropriate ARMA model for N_t. This would typically be adequate to specify a tentative model for the error term N_t, especially when the explanatory variables x_{ti} are deterministic functions such as sine and cosine functions, or polynomial terms. In such cases, it is known (e.g., Anderson, 1971, Section 10.2) that the least-squares estimator for $\boldsymbol{\beta}$ is an asymptotically efficient estimator relative to the best linear estimator. In addition, it is known that the sample autocorrelations and partial autocorrelations calculated using the residuals from the preliminary least-squares fit are asymptotically equivalent to those obtained from the actual noise series N_t (e.g., Anderson, 1971, Section 10.3; Fuller, 1996, Section 9.3).

Hence, the complete model that we consider is

$$w_t = \boldsymbol{x}_t'\boldsymbol{\beta} + N_t \qquad \phi(B)(1-B)^d N_t = \theta(B)a_t \qquad t = 1, \ldots, n \tag{9.5.4}$$

where $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tk})'$. Estimates of all parameters can be obtained by maximum likelihood methods. The resulting estimate for $\boldsymbol{\beta}$ has the GLS form

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{w}$$

but where V is replaced by the estimate $\hat{\mathbf{V}}$ obtained from the MLEs $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$ of the ARMA parameters for N_t. Also, $\text{cov}[\hat{\boldsymbol{\beta}}_G] \simeq (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$. The estimation can be

performed iteratively, alternating between calculation of $\hat{\boldsymbol{\beta}}_G$ for given estimates $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$, and reestimation of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, given $\hat{\boldsymbol{\beta}}_G$ and the estimated noise series $\hat{N}_t = w_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_G$.

Transformed Model. With the ARMA model specified for N_t, the computation of the generalized least-squares estimator of $\boldsymbol{\beta}$ can be carried out in a computationally convenient manner as follows. Let $\mathbf{P}'$ be a lower triangular matrix, such that $\mathbf{P}'\mathbf{V}\mathbf{P} = \sigma_a^2\mathbf{I}$, that is, $\mathbf{V}^{-1} = \mathbf{P}\mathbf{P}'/\sigma_a^2$. Then, as in Appendix A7.2.5, the GLS estimator can be obtained from the transformed regression model

$$\mathbf{P}'\mathbf{w} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\boldsymbol{N} \tag{9.5.5}$$

or $\mathbf{w}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{a}$, where the transformed variables are $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$, $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$, and $\boldsymbol{a} = \mathbf{P}'\boldsymbol{N}$. Since the covariance matrix of the error vector $\boldsymbol{a} = \mathbf{P}'\boldsymbol{N}$ in the transformed model is

$$\text{cov}[\boldsymbol{a}] = \mathbf{P}'\text{cov}[\boldsymbol{N}]\mathbf{P} = \mathbf{P}'\mathbf{V}\mathbf{P} = \sigma_a^2\mathbf{I}$$

we can now use ordinary least-squares to estimate $\boldsymbol{\beta}$ in the transformed model. That is, the GLS estimator of $\boldsymbol{\beta}$ is obtained as the LS estimator in terms of the transformed variables $\mathbf{w}^*$ and $\mathbf{X}^*$ as

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{w}^* \qquad \text{with} \qquad \text{cov}[\hat{\boldsymbol{\beta}}_G] = \sigma_a^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} \tag{9.5.6}$$

However, since the ARMA parameters for N_t are not known in practice, one must still iterate between the computation of $\hat{\boldsymbol{\beta}}_G$ using the current estimates of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ to form the transformation matrix $\hat{\mathbf{P}}'$, and estimation of the ARMA parameters based on $\hat{N}_t = w_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_G$ constructed from the current estimate of $\boldsymbol{\beta}$. The computational procedure used to determine the exact sum-of-squares function for the specified ARMA model will also essentially determine the nature of the transformation matrix $\mathbf{P}'$. For instance, the innovations algorithm described in Section 7.4 gives the sum of squares for an ARMA model as $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sigma_a^2\mathbf{w}'\mathbf{V}^{-1}\mathbf{w} = \mathbf{e}'\mathbf{D}^{-1}\mathbf{e}$, where $\mathbf{e} = \mathbf{G}^{-1}\mathbf{L}_\phi\mathbf{w}$ and $\mathbf{D} = \text{diag}(v_1, \ldots, v_n)$, and $\mathbf{G}$ and $\mathbf{L}_\phi$ are specific lower triangular matrices. Hence, the innovations algorithm can be viewed as providing the transformation matrix $\mathbf{P}' = \mathbf{D}^{-1/2}\mathbf{G}^{-1}\mathbf{L}_\phi$ such that $\mathbf{w}^* = \mathbf{D}^{-1/2}\mathbf{G}^{-1}\mathbf{L}_\phi\mathbf{w} \equiv \mathbf{P}'\mathbf{w}$ has covariance matrix of the ''standard'' form

$$\text{cov}[\mathbf{w}^*] = \mathbf{P}'\text{cov}[\mathbf{w}]\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{G}^{-1}\mathbf{L}_\phi\,\text{cov}[\mathbf{w}]\mathbf{L}_\phi'\mathbf{G}'^{-1}\mathbf{D}^{-1/2} = \sigma_a^2\mathbf{I}$$

Therefore, the required transformed variables $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$ and $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$ in (9.5.6) can be obtained by applying the innovations algorithm recursive calculations (e.g.,(7.4.9)) to the series $\mathbf{w} = (w_1, \ldots, w_n)'$ and to each column, $\mathbf{x}_i' = (x_{1i}, \ldots, x_{ni})'$, $i = 1, \ldots, k$, of the matrix $\mathbf{X}$.

Example. We take the simple example of an AR(1) model, $(1 - \phi B)N_t = a_t$, for the noise N_t, for illustration. Then the covariance matrix $\mathbf{V}$ of $\boldsymbol{N}$ has (i, j)th element given by $\gamma_{i-j} = \sigma_a^2\phi^{|i-j|}/(1 - \phi^2)$. The $n \times n$ matrix $\mathbf{P}'$ such that $\mathbf{P}'\mathbf{V}\mathbf{P} = \sigma_a^2\mathbf{I}$ has its (1, 1) element equal to $(1 - \phi^2)^{1/2}$, its remaining diagonal elements equal to 1, its first subdiagonal elements equal to $-\phi$, and all remaining elements equal to zero. Hence, the transformed variables are $w_1^* = (1 - \phi^2)^{1/2}w_1$ and $w_t^* = w_t - \phi w_{t-1}, t = 2, 3, \ldots, n$, and similarly for the transformed explanatory variables x_{ti}^*. In effect, with AR(1) errors, the original model

(9.5.1) has been transformed by applying the AR(1) operator $(1 - \phi B)$ throughout the equation to obtain

$$w_t - \phi w_{t-1} = \beta_1(x_{t1} - \phi x_{t-1,1}) + \beta_2(x_{t2} - \phi x_{t-1,2}) + \cdots + \beta_k(x_{tk} - \phi x_{t-1,k}) + a_t \tag{9.5.7}$$

or, equivalently, $w_t^* = \beta_1 x_{t1}^* + \beta_2 x_{t2}^* + \cdots + \beta_k x_{tk}^* + a_t$, where the errors a_t now are uncorrelated. Thus, ordinary least-squares applies to the transformed regression model, and the resulting estimator is the same as the GLS estimator in the original regression model.

Generalization of the transformation procedure to higher order AR models is straightforward. Apart from special treatment for the initial p observations, the transformed variables are $w_t^* = \phi(B)w_t = w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p}$ and $x_{ti}^* = \phi(B)x_{ti} = x_{ti} - \phi_1 x_{t-1,i} - \cdots - \phi_p x_{t-p,i}$, $i = 1, \ldots, k$. The exact form of the transformation in the case of mixed ARMA models will be more complicated [an approximate form is $w_t^* \simeq \theta^{-1}(B)\phi(B)w_t$, and so on] but can be determined through the same procedure as is used to construct the exact sum-of-squares function for the ARMA model.

Forecasting. Forecasting for regression models with time series errors is straightforward when future values $x_{t+l,i}$ of the explanatory variables are known, as would be the case for deterministic functions such as sine and cosine functions, for example. Then, based on forecast origin t, the lead l forecast of

$$w_{t+l} = \beta_1 x_{t+l,1} + \cdots + \beta_k x_{t+l,k} + N_{t+l}$$

based on past values through time t, is

$$\hat{w}_t(l) = \beta_1 x_{t+l,1} + \beta_2 x_{t+l,2} + \cdots + \beta_k x_{t+l,k} + \hat{N}_t(l) \tag{9.5.8}$$

where $\hat{N}_t(l)$ is the usual l-step-ahead forecast of N_{t+l} from the ARMA(p, q) model, $\phi(B)N_t = \theta(B)a_t$, based on the past values of the noise series N_t. The forecast error is

$$e_t(l) = w_{t+l} - \hat{w}_t(l) = N_{t+l} - \hat{N}_t(l) = \sum_{i=0}^{l-1} \psi_i a_{t+l-i} \tag{9.5.9}$$

with $V(l) = \text{var}[e_t(l)] = \sigma_a^2 \sum_{i=0}^{l-1} \psi_i^2$, just the forecast error and its variance from the ARMA model for the noise series N_t, where the ψ_i are the coefficients in $\psi(B) = \phi^{-1}(B)\theta(B)$ for the noise model.

Example. For the model

$$w_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi t}{12}\right) + \beta_2 \sin\left(\frac{2\pi t}{12}\right) + N_t$$

where $(1 - \phi B)N_t = a_t$, the forecasts are

$$\hat{w}_t(l) = \beta_0 + \beta_1 \cos\left[\frac{2\pi(t+l)}{12}\right] + \beta_2 \sin\left[\frac{2\pi(t+l)}{12}\right] + \hat{N}_t(l)$$

with $\hat{N}_t(l) = \phi^l N_t$. Note that these forecasts are similar in functional form to those that would be obtained in an ARMA(1, 3) model (with zero constant term) for the series

$(1 - B)(1 - \sqrt{3}B + B^2)w_t$, except that the β coefficients in the forecast function for the regression model case are deterministic, not adaptive, as was noted at the end of Section 9.4.2.

In practice, estimates of $\boldsymbol{\beta}$ and the time series model parameters would be used to obtain the estimated noise series $\hat{N}_t$ from which forecasts of future values would be made. The effect of parameter estimation errors on the variance of the corresponding forecast error was investigated by Baillie (1979) for regression models with autoregressive errors, generalizing a similar study by Yamamoto (1976) conducted for pure autoregressive models.

More detailed discussions of regression analysis with time series errors are given by Harvey and Phillips (1979) and by Wincek and Reinsel (1986), who also consider the possibility of missing data. A state-space approach with associated Kalman filtering calculations, as discussed in Section 7.4, can be employed for the regression model with time series errors, and this corresponds to one particular choice for the transformation matrix $\mathbf{P}'$ in the above discussion. A specific application of the use of regression models with time series errors to model calendar effects in seasonal time series was given by Bell and Hillmer (1983), while Reinsel and Tiao (1987) used regression models with time series errors to model atmospheric ozone data for estimation of trends.

One common application of regression models for seasonal time series is where seasonality can be modeled as a *deterministic seasonal mean* model. Then, for monthly seasonal data, for example, we might consider a model of the form

$$z_t = \beta_0 + \sum_{j=1}^{6} \left[\beta_{1j} \cos\left(\frac{2\pi jt}{12}\right) + \beta_{2j} \sin\left(\frac{2\pi jt}{12}\right) \right] + N_t \tag{9.5.10}$$

where N_t is modeled as an ARIMA process. As an example, Reinsel and Tiao (1987) consider the time series z_t of monthly averages of atmospheric total column ozone measured at the station Aspendale, Australia, for the period from 1958 to 1984. This series is highly seasonal, and so in terms of ARIMA modeling, the seasonal differences $w_t = (1 - B^{12})z_t$, were considered. Based on the sample ACF and PACF of w_t, the following model was specified and estimated,

$$(1 - 0.48B - 0.22B^2)(1 - B^{12})z_t = (1 - 0.99\ B^{12})a_t$$

and the model was found to be adequate. We see that this model contains a near-common seasonal difference factor $(1 - B^{12})$, and consequently, it is equivalent to the model that contains a deterministic seasonal component, $z_t = S_t + N_t$, of exactly the form given in (9.5.10), and where N_t follows the AR(2) model, $(1 - 0.48B - 0.22B^2)N_t = a_t$. This model was estimated using regression methods similar to those discussed above.

Sometimes, the effects of a predictor variable $\{x_t\}$ on z_t are not confined to a single time period t, but the effects are more dynamic over time and are ''distributed'' over several time periods. With a single predictor variable, this would lead to models of the form

$$z_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \cdots + N_t$$

where N_t might be an ARIMA process. For parsimonious modeling, the regression coefficients β_i can be formulated as specific functions of a small number or unknown parameters. Such models are referred to as *transfer function* models or *dynamic regression* models, and will be considered in detail in Chapters 11 and 12.

Remark. Note that regression models with autocorrelated errors can be fitted to data using the arima() function in R with an argument xreg added to account for regression terms; type help(arima) for details. For further discussion, see also Venables and Ripley (2002). An alternative available in the MTS package of R is the function tfm1() that can be used to fit a regression model with a single input variable X_t. We demonstrate the use of this function to develop a dynamic regression model in Chapter 12. A similar function which allows for two input series is also available in the MTS package of R.

9.5.2 Restricted Maximum Likelihood Estimation for Regression Models

A detracting feature of the maximum likelihood estimator (MLE) of the ARMA parameters in the linear regression model (9.5.1) is that the MLE can produce a nonnegligible bias for small to moderate sample sizes. This bias could have significant impact on inferences of the regression parameters $\boldsymbol{\beta}$ based on the GLS estimation, through the approximation $\text{cov}[\hat{\boldsymbol{\beta}}_G] \simeq (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$, where $\hat{\mathbf{V}}$ involves the ML estimates of the ARMA parameters. One ''preventive'' approach for reducing the bias is to use the *restricted maximum likelihood* (REML) estimation procedure, also known as the *residual maximum likelihood* estimation procedure, for the ARMA model parameters.

The REML method has been popular and commonly used in the estimation of variance components in mixed-effects linear models. For ARMA models, this procedure has been used by Cooper and Thompson (1977) and Tunnicliffe Wilson (1989), among others. Cheang and Reinsel (2000, 2003) compared the ML and REML estimation methods, and bias characteristics in particular, for time series regression models with AR and ARMA noise (as well as fractional ARIMA noise, see Section 10.4). They established approximate bias characteristics for these estimators, and confirmed empirically that REML typically reduces the bias substantially over ML estimation. Consequently, the REML approach leads to more accurate inferences about the regression parameters.

The REML estimation of the parameters in the ARMA noise models differs from the ML estimation in that it explicitly takes into account the fact that the regression parameters $\boldsymbol{\beta}$ are unknown and must be estimated (i.e., estimation of ARMA parameters relies on the residuals $\hat{N}_t = w_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_G$ rather than the ''true'' noise $N_t = w_t - \boldsymbol{x}_t'\boldsymbol{\beta}$). In the REML estimation method, the estimates of $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 are determined so as to maximize the restricted likelihood function. This is the likelihood function based on observation of the ''residual vector'' of error contrasts $\mathbf{u} = \mathbf{H}'\mathbf{w}$ only, whose distribution is free of the regression parameters $\boldsymbol{\beta}$, rather than the likelihood based on the 'full' vector of observations $\mathbf{w}$. Here, $\mathbf{H}'$ is any $(n-k) \times n$ full rank matrix such that $\mathbf{H}'\mathbf{X} = \mathbf{0}$, so the regression effects are eliminated in $\mathbf{u} = \mathbf{H}'\mathbf{w}$ and its distribution is free of the parameters $\boldsymbol{\beta}$.

Assuming normality, the distribution of $\mathbf{w}$ is normal with mean vector $E(\mathbf{w}) = \mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\text{cov}[\mathbf{w}] = \mathbf{V}$, which we write as $\mathbf{V} = \sigma_a^2\mathbf{V}_*$ for convenience of notation. Then, $\mathbf{u} = \mathbf{H}'\mathbf{w}$ has normal distribution with zero mean vector and covariance matrix $\text{cov}[\mathbf{u}] = \sigma_a^2\mathbf{H}'\mathbf{V}_*\mathbf{H}$. Thus, the likelihood of $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 based on $\mathbf{u}$, that is, the density of $\mathbf{u}$, is

$$p(\mathbf{u}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n-k)/2}|\mathbf{H}'\mathbf{V}_*\mathbf{H}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2}\mathbf{u}'(\mathbf{H}'\mathbf{V}_*\mathbf{H})^{-1}\mathbf{u}\right]$$

It has been established (e.g., Harville, 1974, 1977), however, that this likelihood (i.e., density) can be expressed in an equivalent form that does not involve the particular choice of error contrast matrix $\mathbf{H}'$ as

$$\begin{aligned} L_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) &\equiv p(\mathbf{u}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) \\ &= (2\pi\sigma_a^2)^{-(n-k)/2} |\mathbf{X}'\mathbf{X}|^{1/2} |\mathbf{V}_*|^{-1/2} \\ &\quad \times |\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\hat{\boldsymbol{\beta}}_G, \boldsymbol{\phi}, \boldsymbol{\theta})\right] \end{aligned} \tag{9.5.11}$$

where

$$\begin{aligned} S(\hat{\boldsymbol{\beta}}_G, \boldsymbol{\phi}, \boldsymbol{\theta}) &= (\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)'\mathbf{V}_*^{-1}(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}_G) \\ &\equiv \mathbf{w}'(\mathbf{V}_*^{-1} - \mathbf{V}_*^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_\star^{-1})\mathbf{w} \end{aligned}$$

and $\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{w}$. Evaluation of the restricted likelihood (9.5.11) requires little additional computational effort beyond that of the ''full'' likelihood, only the additional factor $|\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|$. Therefore, numerical determination of the REML estimates of $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 is very similar to methods for ML estimation of the ARMA model parameters. However, one difference is that the REML estimate of σ_a^2 takes into account the loss in degrees of freedom that results from estimating the regression parameters and is given by $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\beta}}_G, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/(n-k)$ as opposed to $S(\hat{\boldsymbol{\beta}}_G, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/n$ for the ML estimate, although arguments can be put forth for use of the divisor $n-k-p-q$ rather than $n-k$ in the REML estimate $\hat{\sigma}_a^2$. For further discussion and details related to REML estimation, see Tunnicliffe Wilson (1989) and Cheang and Reinsel (2000, 2003).

APPENDIX A9.1 AUTOCOVARIANCES FOR SOME SEASONAL MODELS

See the following Table A9.1:

TABLE A9.1 Autocovariances for Some Seasonal Models

Model	(Autocovariances of $w_t)/\sigma_a^2$	Special Characteristics
(1) $w_t = (1 - \theta B)(1 - \Theta B^s)a_t$ $w_t = a_t - \theta a_{t-1} - \Theta a_{t-s} + \theta\Theta a_{t-s-1}$ $s \geq 3$	$\gamma_0 = (1 + \theta^2)(1 + \Theta^2)$ $\gamma_1 = -(1 + \Theta^2)$ $\gamma_{s-1} = \theta\Theta$ $\gamma_s = -\Theta(1 + \theta^2)$ $\gamma_{s+1} = \gamma_{s-1}$ All other autocovariances are zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\rho_{s-1} = \rho_{s+1} = \rho_1\rho_s$
(2) $(1 - \Phi B^s)w_t = (l - \theta B)(1 - \Theta B^s)a_t$ $w_t - \Phi w_{t-s} = a_t - \theta a_{t-l} - \Theta a_{t-s} + \theta\Theta a_{t-s-1}$ $s \geq 3$	$\gamma_0 = (1 + \theta^2)\left[1 + (\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_1 = -\theta\left[1 + (\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_{s-1} = \theta\left[\Theta - \Phi - \Phi(\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_s = -(1 + \theta^2)\left[\Theta - \Phi - \Phi(\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 2$ For $s \geq 4, \gamma_2, \gamma_3, \ldots, \gamma_{s-2}$ are all zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\gamma_j = \Phi\gamma_{j-s}\ \ j \geq s + 2$
(3) $w_t = (1 - \theta_1 B - \theta_2 B^2)$ $\times (1 - \Theta_1 B^s - \Theta_2 B^{2s})a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \Theta_1 a_{t-s}$ $+\theta_1\Theta_1 a_{t-s-1} + \theta_2\Theta_1 a_{t-s-2}$ $-\Theta_2 a_{t-2s} + \theta_1\Theta_2 a_{t-2s-1}$ $+\theta_2\Theta_2 a_{t-2s-2}$ $s \geq 5$	$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_1 = -\theta_1(1 - \theta_2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_2 = -\theta_2(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_{s-2} = \theta_2\Theta_1(1 - \Theta_2)$ $\gamma_{s-1} = \theta_1\Theta_1(1 - \theta_2)(1 - \Theta_2)$ $\gamma_s = -\Theta_1(1 - \Theta_2)(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{s+2} = \gamma_{s-2}$ $\gamma_{2s-2} = \theta_2\Theta_2$ $\gamma_{2s-1} = \theta_1(1 - \theta_2)\Theta_2$ $\gamma_{2s} = -\Theta_2(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{2s+1} = \gamma_{2s-1}$	(*a*) $\gamma_{s-2} = \gamma_{s+2}$ (*b*) $\gamma_{s-1} = \gamma_{s+1}$ (*c*) $\gamma_{s-2} = \gamma_{2s+2}$ (*d*) $\gamma_{2s-1} = \gamma_{2s+1}$

TABLE A9.1 (*continued*)

Model	(Autocovariances of $w_t)/\sigma_a^2$	Special Characteristics
	$\gamma_{2s+2} = \gamma_{2s-2}$ All other autocovariances are zero.	
(3a) *Special case of model 3* $w_t = (1 - \theta_1 B - \theta_2 B^2)(1 - \Theta B^s)a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \Theta a_{t-s}$ $+\theta_1 \Theta a_{t-s-1} + \theta_2 \Theta a_{t-s-2}$ $s \geq 5$	$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)(1 + \Theta^2)$ $\gamma_1 = -\theta_1(1 - \theta_2)(1 + \Theta^2)$ $\gamma_2 = -\theta_2(1 + \Theta^2)$ $\gamma_{s-2} = \theta_2 \Theta$ $\gamma_{s-1} = \theta_1(1 - \theta_2)\Theta$ $\gamma_s = -\Theta(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{s+2} = \gamma_{s-2}$ All other autocovariances are zero.	(a) $\gamma_{s-2} = \gamma_{s+2}$ (b) $\gamma_{s-1} = \gamma_{s+1}$
(3b) *Special case of model 3* $w_t = (1 - \theta B)(1 - \Theta_1 B^s - \Theta_2 B^{2s})a_t$ $w_t = a_t - \theta a_{t-1} - \Theta_1 a_{t-s} + \theta \Theta_1 a_{t-s-1}$ $-\Theta_2 a_{t-2s} + \theta \Theta_2 a_{t-2s-1}$ $s \geq 3$	$\gamma_0 = (1 + \theta^2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_1 = -\theta(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_{s-1} = \theta \Theta_1(1 - \Theta_2)$ $\gamma_s = -\Theta_1(1 - \Theta_2)(1 + \theta^2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{2s-1} = \theta \Theta_2$ $\gamma_{2s} = -\Theta_2(1 + \theta^2)$ $\gamma_{2s+1} = \gamma_{2s-1}$ All other autocovariances are zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\gamma_{2s-1} = \gamma_{2s+1}$
(4) $w_t = (1 - \theta_1 B - \theta_s B^s - \theta_{s+1} B^{s+1})a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s} - \theta_{s+1} a_{t-s-1}$ $s \geq 3$	$\gamma_0 = 1 + \theta_1^2 + \theta_s^2 + \theta_{s+1}^2$ $\gamma_1 = -\theta_1 + \theta_s \theta_{s+1}$ $\gamma_{s-1} = \theta_1 \theta_s$ $\gamma_s = -\theta_s + \theta_1 \theta_{s+1}$ $\gamma_{s+1} = -\theta_{s+1}$ All other autocovariances are zero	(a) In general, $\gamma_{s-1} \neq \gamma_{s+1}$ $\gamma_1 \gamma_s \neq \gamma_{s+1}$

(*continued*)

TABLE A9.1 (*continued*)

Model	(Autocovariances of $w_t)/\sigma_a^2$	Special Characteristics
(4a) *Special case of model 4* $w_t = (1 - \theta_1 B - \theta_s B^s)a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $s \geq 3$	$\gamma_0 = 1 + \theta_1^2 + \theta_s^2$ $\gamma_1 = -\theta_1$ $\gamma_{s-1} = \theta_1\theta_s$ $\gamma_s = -\theta_s$ All other autocovariances are zero.	(a) Unlike model 4, $\gamma_{s+1} = 0$
(5) $(1 - \Phi B^s)w_t = (1 - \theta_1 B - \theta_s B^s$ $-\theta_{s+1}B^{s+1})a_t$ $w_t - \Phi w_{t-s} = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $-\theta_{s+1}a_{t-s-1}$ $s \geq 3$	$\gamma_0 = 1 + \theta_1^2 + \dfrac{(\theta_s - \Phi)^2}{1 - \Phi^2} + \dfrac{(\theta_{s+1} + \theta_1\Phi)^2}{1 - \Phi^2}$ $\gamma_1 = -\theta_1 + \dfrac{(\theta_s - \Phi)(\theta_{s+1} + \theta_1\Phi)}{1 - \Phi^2}$ $\gamma_{s-1} = (\theta_s - \Phi)\left[\theta_1 + \Phi\dfrac{\theta_{s+1} + \theta_1\Phi}{1 - \Phi^2}\right]$ $\gamma_s = -(\theta_s - \Phi)\left[1 - \Phi\dfrac{\theta_s - \Phi}{1 - \Phi^2}\right]$ $+(\theta_{s+1} + \theta_1\Phi)\left[\theta_1 + \Phi\dfrac{\theta_{s+1} - \theta_1\Phi}{1 - \Phi^2}\right]$ $\gamma_{s+1} = -(\theta_{s+1} + \theta_1\Phi)\left[1 - \Phi\dfrac{\theta_s - \Phi}{1 - \Phi^2}\right]$ $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 2$ For $s \geq 4, \gamma_2, \ldots, \gamma_{s-2}$ are all zero.	(a) $\gamma_{s-1} \neq \gamma_{s+1}$ (b) $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 2$
(5a) *Special case of model 5* $(1 - \Phi B^s)w_t = (1 - \theta_1 B - \theta_s B^s)a_t$ $w_t - \Phi w_{t-s} = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $s \geq 3$	$\gamma_0 = 1 + \dfrac{\theta_1^2 + (\theta_s - \Phi)^2}{1 - \Phi^2}$ $\gamma_1 = -\theta_1\left[1 - \Phi\dfrac{\theta_s - \Phi}{1 - \Phi^2}\right]$ $\gamma_{s-1} = \dfrac{\theta_1(\theta_s - \Phi)}{1 - \Phi^2}$ $\gamma_s = \dfrac{\Phi\theta_1^2 - (\theta_s - \Phi)(1 - \Phi\theta_s)}{1 - \Phi^2}$ $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 1$ For $s \geq 4, \gamma_2, \ldots, \gamma_{s-2}$ are all zero.	(a) Unlike model 5, $\gamma_{s+1} = \Phi\gamma_1$

EXERCISES

9.1. Show that the seasonal difference operator $1 - B^{12}$, often useful in the analysis of monthly data, may be factorized as follows:

$$(1 - B^{12}) = (1 + B)(1 - \sqrt{3}B + B^2)(1 - B + B^2)(1 + B^2)(1 + B + B^2) \times (1 + \sqrt{3}B + B^2)(1 - B)$$

Plot the zeros of this expression in the unit circle and show by actual numerical calculation and plotting of the results that the factors in the order given above correspond to sinusoids with frequencies (in cycles per year) of 6, 5, 4, 3, 2, 1, together with a constant term. [For example, the difference equation $(1 - B + B^2)x_t = 0$ with arbitrary starting values $x_1 = 0, x_2 = 1$ yields $x_3 = 1, x_4 = 0, x_5 = -1$, and so on, generating a sine wave of frequency 2 cycles per year.]

9.2. A method that has sometimes been used for ''deseasonalizing'' monthly time series employs an equally weighted 12-month moving average:

$$\overline{z}_t = \frac{1}{12}(z_t + z_{t-1} + \cdots + z_{t-11})$$

(a) Using the decomposition $(1 - B^{12}) = (1 - B)(1 + B + B^2 + \cdots + B^{11})$, show that $12(\overline{z}_t - \overline{z}_{t-1}) = (1 - B^{12})z_t$.

(b) The exceedance for a given month over the previous moving average may be computed as $z_t - \overline{z}_{t-1}$. A quantity u_t may then be calculated that compares the current exceedance with the average of similar monthly exceedances experienced over the last k years. Show that u_t may be written as

$$u_t = \left(1 - \frac{B}{12}\frac{1 - B^{12}}{1 - B}\right)\left(1 - \frac{B^{12}}{k}\frac{1 - B^{12k}}{1 - B^{12}}\right) z_t$$

9.3. It has been shown (Tiao et al., 1975) that monthly averages for the (smog-producing) oxidant level in Azusa, California, may be represented by the model

$$(1 - B^{12})z_t = (1 + 0.2B)(1 - 0.9B^{12})a_t \qquad \sigma_a^2 = 1.0$$

(a) Compute and plot the ψ_j weights of this model.

(b) Compute and plot the π_j weights of this model.

(c) Calculate the standard deviations of the forecast errors 3 months and 12 months ahead.

(d) Obtain the eventual forecast function.

9.4. The monthly oxidant averages in parts per hundred million in Azusa from January 1969 to December 1972 were as follows:

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1969	2.1	2.6	4.1	3.9	6.7	5.1	7.8	9.3	7.5	4.1	2.9	2.6
1970	2.0	3.2	3.7	4.5	6.1	6.5	8.7	9.1	8.1	4.9	3.6	2.0
1971	2.4	3.3	3.3	4.0	3.6	6.2	7.7	6.8	5.8	4.1	3.0	1.6
1972	1.9	3.0	4.5	4.2	4.8	5.7	7.1	4.8	4.2	2.3	2.1	1.6

Using the model of Exercise 9.3, compute the forecasts for the next 24 months. (Approximate unknown a's by zeros.)

9.5. Thompson and Tiao (1971) have shown that the outward station movements of telephones (logged data) in Wisconsin are well represented by the model

$$(1 - 0.5B^3)(1 - B^{12})z_t = (1 - 0.2B^9 - 0.3B^{12} - 0.2B^{13})a_t$$

Obtain and plot the autocorrelation function of $w_t = (1 - B^{12})z_t$ for lags $1, 2, \ldots, 24$.

9.6. Consider the airline series analyzed earlier in this chapter. We have seen that the logarithm of the series is well represented by the multiplicative model $w_t = (1 - \theta B)(1 - \Theta_{12}B^{12})a_t$

(a) Compute and plot the 36-step-ahead forecasts and associated ± 2 forecast error limits for the logged series.

(b) Use the results in part (a) to obtain 12-step-ahead forecasts and associated forecast error limits for the original series. Plot the results.

9.7. Quarterly earnings per share of the U.S. company Johnson & Johnson are available for the period 1960–1980 as series 'JohnsonJohnson' in the R datasets package.

(a) Plot the time series using the graphics capabilities in R.

(b) Determine a variance stabilizing transformation for the series.

(c) Plot the autocorrelation functions and identify a suitable model (or models) for the series.

(d) Estimate the parameters of the model (or models) identified in part (**c**) and assess the statistical significance of the estimated parameters.

(e) Perform diagnostic checks to determine the adequacy of the fitted model.

(f) Compute and plot the l-step-ahead forecasts and associated two-standard-error prediction limits, $l = 1, \ldots, 4$, for this series.

9.8. Monthly Mauna Loa atmospheric CO_2 concentration readings for the period 1959–1997 are available as series 'co_2' in the R datasets package.

(a) Plot the time series and comment on the pattern in the data.

(b) Examine the autocorrelation structure and develop a suitable time series model for this series.

(c) Compute and plot the 12-step-ahead forecasts and associated two-standard-error prediction limits.

9.9. A time series representing the total monthly electricity generated in the United States (in millions of kilowatt-hours) for the period January 1970 to December 2005 is available as series 'electricity' in the R TSA package.

(a) Plot the series and comment. Is a variance stabilizing transformation needed for this case?

(b) Determine a suitable model for the series following the iterative three-stage procedure of model identification, parameter estimation, and diagnostics checking.

(c) Is there evidence of a deterministic seasonal pattern in this series? If so, how would this impact your choice of model for this series?

9.10. Consider the time series model $w_t = \beta_0 + N_t$ where N_t follows the AR(1) model $N_t = \phi N_{t-1} + a_t$. Assume that a series of length n is available for analysis.

(a) Assuming that the parameter ϕ is known, derive the generalized least-squares estimator of the constant β_0 in this model.

(b) Repeat the derivation in part (**a**) assuming that N_t follows the seasonal AR model $N_t = \phi_4 N_{t-4} + a_t$.

9.11. Suppose the quarterly seasonal process $\{z_t\}$ is represented as $z_t = S_t + a_{2t}$, where S_t follows a ''seasonal random walk'' model $(1 - B^4)S_t = \theta_0 + a_{1t}$, and a_{1t} and a_{2t} are independent white noise processes with variances $\sigma^2_{a_1}$ and $\sigma^2_{a_2}$, respectively. Show that z_t follows the seasonal ARIMA model $(1 - B^4)z_t = \theta_0 + (1 - \Theta B^4)a_t$, and determine expressions for Θ and σ^2_a in terms of the variance parameters of the other two processes. Discuss the implication if the resulting value of Θ is equal (or very close) to one, with regard to deterministic seasonal components.

9.12. Monthly averages of hourly ozone readings in downtown Los Angeles for the period from January 1955 to December 1972 are included as Series R in Part 5 of this book; see also http://pages.stat.wisc.edu/reinsel/bjr-data/.

(a) Plot the time series and comment.

(b) Develop a suitable time model for this time series. Discuss the adequacy of the selected model.

10

ADDITIONAL TOPICS AND EXTENSIONS

In previous chapters, the properties of *linear* autoregressive–moving average models have been examined extensively and it has been shown how these models can be used to represent stationary and nonstationary time series that arise in practice. This chapter will discuss additional topics that either supplement or extend the material presented in earlier chapters. We begin by discussing unit root tests that can be used as a supplementary tool to determine whether a time series is unit root nonstationary and can be transformed to a stationary series through differencing. This topic is discussed in Section 10.1. Unit root testing has received considerable attention in the econometrics literature, in particular, since it appears to be a common starting point for applied research in macroeconomics. For example, unit root tests are an integral part of the methodology used to detect long-term equilibrium relationships among nonstationary economic time series, commonly referred to as cointegration. In Section 10.2, we consider models for conditional heteroscedastic time series, which exhibit periods of differing degrees of volatility or variability depending on the past history of the series. Such behavior is common in many economic and financial time series, in particular. In Section 10.3, we introduce several classes of nonlinear time series models, which are capable of capturing some distinctive features in the behavior of processes that deviate from linear Gaussian time series. Finally, Section 10.4 looks at models for long memory processes, which are characterized by the much slower convergence to zero of their autocorrelation function ρ_k as $k \to \infty$ compared with the dependence structure of ARMA processes.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

10.1 TESTS FOR UNIT ROOTS IN ARIMA MODELS

As discussed in earlier chapters, the initial decision concerning the need for differencing is based, informally, on characteristics of the time series plot of z_t and of its sample autocorrelation function. In particular, a failure of the autocorrelations r_k to dampen out sufficiently quickly would indicate that the time series is nonstationary and needs to be differenced. This can be evaluated further using formal tests for unit roots in the autoregressive operator of the model. Testing for unit roots has received considerable attention in the time series literature motivated by econometric applications, in particular. Early contributions to this area include work by Dickey and Fuller (1979, 1981). These authors proposed tests based on the conditional least-squares estimator for an autoregressive process and the corresponding ''t-statistic.'' While the underlying concepts are fairly straightforward, a number of challenges arise in practice. In particular, the distribution theory for parameter estimates and associated test statistics developed for stationary time series do not apply when a unit root is present in the model. The asymptotic distributions are functions of standard Brownian motions and do not have convenient closed-form expressions. As a result, the percentiles of the distributions needed to perform the tests have to be evaluated using numerical approximations or by simulation. Moreover, the form of the test statistics and their asymptotic distributions are impacted by the presence of deterministic terms such as constants or time trends in the model. The size and power characteristics of unit root tests can also be a concern for shorter time series. This section provides a brief description of the tests proposed by Dickey and Fuller and summarizes some of the subsequent developments. For a more detailed discussion of unit root testing, see, for example, Hamilton (1994) and Fuller (1996). Reviews of unit root tests and their applications are provided by Dickey et al. (1986), Pantula et al. (1994), Phillips and Xiao (1998), and Haldrup et al. (2013), among others.

10.1.1 Tests for Unit Roots in AR Models

Simple AR(1) Model. To introduce unit root testing, we first examine the simple AR(1) model $z_t = \phi z_{t-1} + a_t, t = 1, 2, \ldots, n$, with $z_0 = 0$ and no constant term. We are interested in testing the hypothesis that $\phi = 1$ so that the series follows a random walk. The conditional least-squares (CLS) estimator of ϕ is given by

$$\hat{\phi} = \frac{\sum_{t=2}^{n} z_{t-1} z_t}{\sum_{t=2}^{n} z_{t-1}^2} = \phi + \frac{\sum_{t=2}^{n} z_{t-1} a_t}{\sum_{t=2}^{n} z_{t-1}^2}$$

In the stationary case with $|\phi| < 1$, the statistic $n^{1/2}(\hat{\phi} - \phi)$ has an approximate normal distribution with zero mean and variance $(1 - \phi^2)$. However, when $\phi = 1$, so that $z_t = \sum_{j=0}^{t-1} a_{t-j} + z_0$ in the integrated form, it can be shown that

$$n(\hat{\phi} - 1) = \frac{n^{-1} \sum_{t=2}^{n} z_{t-1} a_t}{n^{-2} \sum_{t=2}^{n} z_{t-1}^2} = O_p(1)$$

bounded in probability as $n \to \infty$, with both the numerator and denominator possessing nondegenerate and nonnormal limiting distributions. Hence, in the nonstationary case the estimator $\hat{\phi}$ approaches its true value $\phi = 1$ with increasing sample size n at a faster rate than in the stationary case.

The limiting distribution of $n(\hat{\phi}-1)$ was studied by Dickey and Fuller (1979) who showed that under the null hypothesis $\phi = 1$

$$n(\hat{\phi}-1) \xrightarrow{\mathcal{D}} \frac{\frac{1}{2}(\Lambda^2-1)}{\Gamma} \tag{10.1.1}$$

where $(\Gamma, \Lambda) = (\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2, \sum_{i=1}^{\infty} 2^{1/2}\gamma_i Z_i)$, with $\gamma_i = 2(-1)^{i+1}/[(2i-1)\pi]$, and the Z_i are iid $N(0, 1)$ distributed random variables. An equivalent representation for the distribution is given by

$$\begin{aligned} n(\hat{\phi}-1) &\xrightarrow{\mathcal{D}} \frac{\int_0^1 B(u)\,dB(u)}{\int_0^1 B(u)^2\,du} \\ &= \frac{\frac{1}{2}(B(1)^2-1)}{\int_0^1 B(u)^2\,du} \end{aligned} \tag{10.1.2}$$

where $B(u)$ is a (continuous-parameter) standard Brownian motion process on [0, 1]; see Chan and Wei (1988). Such a process is characterized by the properties that $B(0) = 0$, increments over nonoverlapping intervals are independent, and $B(u+s) - B(s)$ is distributed as normal $N(0, u)$. Basically, $B(u)$ is the limit as $n \to \infty$ of the process

$$\frac{n^{-1/2}}{\sigma_a} z_{[nu]} = \frac{n^{-1/2}}{\sigma_a} \sum_{t=1}^{[nu]} a_t$$

where $[nu]$ denotes the largest integer part of $nu, 0 < u < 1$.

By the functional central limit theorem (Billingsley, 1999; Hall and Heyde, 1980, Section 4.2), $n^{-1/2} z_{[nu]}/\sigma_a$ converges in law as $n \to \infty$ to the standard Brownian motion process $\{B(u), 0 < u < 1\}$. The random walk model $z_t = z_{t-1} + a_t$ with $z_0 = 0$ implies that $z_{t-1}a_t = \frac{1}{2}(z_t^2 - z_{t-1}^2 - a_t^2)$, so that

$$n^{-1}\sum_{t=2}^{n} z_{t-1}a_t = \frac{1}{2}\left[n^{-1}z_n^2 - n^{-1}\sum_{t=1}^{n} a_t^2\right] \xrightarrow{\mathcal{D}} \frac{\sigma_a^2}{2}[B(1)^2 - 1] \tag{10.1.3}$$

since $n^{-1}z_n^2 = \sigma_a^2(n^{-1/2}z_n/\sigma_a)^2 \xrightarrow{\mathcal{D}} \sigma_a^2 B(1)^2$ while $n^{-1}\sum_{t=1}^{n} a_t^2 \xrightarrow{\mathcal{P}} \sigma_a^2$ by the law of large numbers. In addition,

$$n^{-2}\sum_{t=2}^{n} z_{t-1}^2 = \sigma_a^2 \int_0^1 \left(\frac{n^{-1/2}z_{[nu]}}{\sigma_a}\right)^2 du + o_p(1) \xrightarrow{\mathcal{D}} \sigma_a^2 \int_0^1 B(u)^2\,du \tag{10.1.4}$$

by the continuous mapping theorem (Billingsley, 1999; Hall and Heyde, 1980, p. 276). Hence, these last two results establish the representation (10.1.2).

The limiting distribution of $n(\hat{\phi}-1)$ described above does not have a closed-form representation but it can be evaluated numerically using simulation. Tables for the percentiles of the limiting distribution are given by Fuller (1996, Appendix 10.A). Fuller also provides

tables for the limiting distribution of the ''Studentized'' statistic

$$\hat{\tau} = \frac{\hat{\phi} - 1}{s_a(\sum_{t=2}^{n} z_{t-1}^2)^{-1/2}} \tag{10.1.5}$$

where $s_a^2 = (n-2)^{-1}(\sum_{t=2}^{n} z_t^2 - \hat{\phi} \sum_{t=2}^{n} z_{t-1} z_t)$ is the residual mean square. These results can be used to test the random walk hypothesis that $\phi = 1$. Since the alternative hypothesis of stationarity is one-sided, the test rejects $\phi = 1$ when $\hat{\tau}$ is sufficiently negative. The test based on $\hat{\tau}$ is commonly referred to as the Dickey–Fuller (DF) test in the literature.

Higher Order AR Models. To extend the results to higher order models, we consider a generalized AR($p+1$) process $z_t = \sum_{j=1}^{p+1} \varphi_j z_{t-j} + a_t$, or $\varphi(B) z_t = a_t$, where $\varphi(B)$ contains a single unit root so that $\varphi(B) = \phi(B)(1-B)$ and $\phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$ is a *stationary* AR operator of order p. Hence,

$$\varphi(B) z_t = \phi(B)(1-B) z_t = z_t - z_{t-1} - \sum_{j=1}^{p} \phi_j (z_{t-j} - z_{t-j-1}) + a_t$$

Testing for a unit root in $\varphi(B)$ is then equivalent to testing $\rho = 1$ in the model

$$z_t = \rho z_{t-1} + \sum_{j=1}^{p} \phi_j (z_{t-j} - z_{t-j-1}) + a_t$$

or equivalently testing $\rho - 1 = 0$ in the model

$$(z_t - z_{t-1}) = (\rho - 1) z_{t-1} + \sum_{j=1}^{p} \phi_j (z_{t-j} - z_{t-j-1}) + a_t$$

In fact, for any generalized AR($p+1$) model $z_t = \sum_{j=1}^{p+1} \varphi_j z_{t-j} + a_t$, it is seen that the model can be written in an equivalent form as

$$w_t = (\rho - 1) z_{t-1} + \sum_{j=1}^{p} \phi_j w_{t-j} + a_t \tag{10.1.6}$$

where $w_t = z_t - z_{t-1}$, $\rho - 1 = -\varphi(1) = \sum_{j=1}^{p+1} \varphi_j - 1$, and $\phi_j = \sum_{i=1}^{p} \varphi_i - 1$. Hence, the existence of a unit root in the AR operator $\varphi(B)$ is equivalent to $\rho = \sum_{j=1}^{p+1} \varphi_j = 1$.

Based on this last form of the model, let $(\hat{\rho} - 1, \hat{\phi}_1, \ldots, \hat{\phi}_p)$ denote the usual conditional least-squares estimates of the parameters in (10.1.6) obtained by regressing w_t on $z_{t-1}, w_{t-1}, \ldots, w_{t-p}$. Then, under the unit root model where $\rho = 1$ and $\phi(B)$ is stationary, it follows from Fuller (1996, Theorem 10.1.2 and Corollary 10.1.2.1) that

$$(\hat{\rho} - 1) \Big/ \left\{ s_a \left(\sum_{t=p+2}^{n} z_{t-1}^2 \right)^{-1/2} \right\}$$

has the same limiting distribution as the Studentized statistic $\hat{\tau}$ in (10.1.5) for the AR(1) model, while $(n-p-1)(\hat{\rho}-1)c$, where $c=\sum_{j=0}^{\infty}\psi_j$ with $\psi(B)=\phi^{-1}(B)$, has approximately the same distribution as the statistic $n(\hat{\phi}-1)$ for the AR(1) model. Also, it follows that the statistic, denoted as $\hat{\tau}$, formed by dividing $(\hat{\rho}-1)$ by its estimated standard error from the least-squares regression will be asymptotically equivalent to the statistic $(\hat{\rho}-1)/\{s_a(\sum_{t=p+2}^{n}z_{t-1}^2)^{-1/2}\}$, and hence will have the same limiting distribution as the statistic $\hat{\tau}$ for the AR(1) case; see Said and Dickey (1984).

The test statistic $\hat{\tau}$ formed from the regression of w_t on $z_{t-1}, w_{t-1}, \ldots, w_{t-p}$ as described above can thus be used to test for a unit root in the AR$(p+1)$ model $\varphi(B)z_t=a_t$. This is the well-known augmented Dickey–Fuller (ADF) test. Furthermore, as shown by Fuller (1996, Theorem 10.1.2), the limiting distribution of the least-squares estimates $(\hat{\phi}_1,\ldots,\hat{\phi}_p)$ for the parameters of the stationary operator $\phi(B)$ in the model is the same as the standard asymptotic distribution for least-squares estimates obtained by regressing the stationary differenced series w_t on $w_{t-1},\ldots,w_{t-p}$. The estimation results for the stationary AR model discussed earlier in Section 7.2.6 are therefore valid in this case.

Inclusion of a Constant Term. The results described above extend with suitable modifications to the more practical case where a constant term θ_0 is included in the least-squares regression. Under stationarity, the constant is related to the mean of the process and equals $\theta_0=(1-\varphi_1-\cdots-\varphi_{p+1})\mu=(1-\rho)\mu$. The least-squares regression yields a test statistic analogous to $\hat{\tau}$ above denoted by $\hat{\tau}_\mu$, although the limiting distribution of this test statistic is derived under the assumption that $\theta_0=0$ under the null hypothesis $\phi=1$. For example, for the AR(1) model $z_t=\phi z_{t-1}+\theta_0+a_t$ with $\theta_0=(1-\phi)\mu$, the least-squares estimator for ϕ is

$$\hat{\phi}_\mu=\frac{\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})(z_t-\bar{z}_{(0)})}{\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})^2} \tag{10.1.7}$$

where $\bar{z}_{(i)}=(n-1)^{-1}\sum_{t=2}^{n}z_{t-i}, i=0,1,$ so that $\hat{\phi}_\mu=\phi+\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})a_t/\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})^2$. When $\phi=1$, the representation for the limiting distribution of $n(\hat{\phi}_\mu-1)$ analogous to (10.1.2) is given by

$$n(\hat{\phi}_\mu-1)\xrightarrow{\mathcal{D}}\frac{\int_0^1 B(u)dB(u)-\xi B(1)}{\int_0^1 B(u)^2du-\xi^2} \tag{10.1.8}$$

where $\xi=\int_0^1 B(u)du$, and it is assumed that $\theta_0=(1-\phi)\mu=0$ when $\phi=1$. The corresponding Studentized test statistic for $\phi=1$ in the AR(1) case is

$$\hat{\tau}_\mu=\frac{\hat{\phi}_\mu-1}{s_a[\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})^2]^{-1/2}} \tag{10.1.9}$$

The limiting distribution of $\hat{\tau}_\mu$ readily follows from the result in (10.1.8). Tables of percentiles of the distribution of $\hat{\tau}_\mu$ when $\phi=1$ are provided by Fuller (1996, p. 642). Note that under $\phi=1$, since $z_t=\sum_{j=0}^{t-1}a_{t-j}+z_0$ in the truncated random shock or integrated form, the terms $z_t-\bar{z}_{(0)}$ and $z_{t-1}-\bar{z}_{(1)}$ do not involve the initial value z_0. Therefore, the distribution theory for the least-squares estimator $\hat{\phi}_\mu$ does not depend on any assumption

concerning z_0. Also, the results for the first-order AR(1) model with a constant term extend to higher order autoregressive models in much the same way as it does when the constant term θ_0 is absent from the model. The tables developed for the percentiles of the limiting distribution of statistic $\hat{\tau}_\mu$ can thus be used for higher order AR models as well.

The procedures described above are based on conditional LS estimation or equivalently on the conditional likelihood assuming that the noise term a_t follows a normal distribution. Pantula et al. (1994) studied *unconditional* likelihood estimation for the AR model with a unit root. They showed that the limiting distributions of estimators and test statistics for unit root based on the unconditional likelihood are different from those based on the conditional approach. For example, in the simple AR(l) model $z_t = \phi z_{t-1} + a_t$ with no constant term included in the estimation, the unconditional log-likelihood is

$$l(\phi, \sigma_a^2) = -\frac{n}{2}\ln(\sigma_a^2) + \frac{1}{2}\ln(1-\phi^2) - \frac{1}{2\sigma_a^2}\left[\sum_{t=2}^{n}(z_t - \phi z_{t-1})^2 + (1-\phi^2)z_1^2\right]$$

as shown in Appendix A7.4. The unconditional ML estimator $\hat{\phi}$, which maximizes $l(\phi, \sigma_a^2)$, is a root of the cubic equation in $\hat{\phi}$ given by (A7.4.20). Pantula et al. (1994) derived the asymptotic distribution of $n(\hat{\phi}_1 - 1)$ and concluded, using Monte Carlo studies, that tests for unit root in AR models based on the unconditional maximum likelihood estimator are more powerful than those based on the conditional maximum likelihood estimator for moderate values of n.

Processes with Deterministic Linear Trend. The asymptotic distribution theory related to the least-squares estimator $\hat{\phi}_\mu$ in (10.1.7) depends heavily on the condition that the constant term θ_0 is zero under the null hypothesis $\phi = 1$, since the behavior of the process $z_t = z_{t-1} + \theta_0 + a_t$ differs fundamentally between the cases $\theta_0 = 0$ and $\theta_0 \neq 0$. When $\theta_0 = 0$, the process is a random walk with zero drift. When $\theta_0 \neq 0$, the model can be written as $z_t = \theta_0 t + z_0 + u_t$, where $u_t = u_{t-1} + a_t$. The process $\{z_t\}$ is now a random walk with drift and its long-term behavior in many respects is dominated by the deterministic linear trend term $\theta_0 t$ contained in z_t. If θ_0 has a nonzero value under the hypothesis $\phi = 1$, then $n^{3/2}(\hat{\phi}_\mu - 1)$ converges in distribution to $N(0, 12\sigma_a^2/\theta_0^2)$ as $n \to \infty$. Thus, when $\theta_0 \neq 0$ the asymptotic normal distribution theory applies to the least-squares estimator $\hat{\phi}_\mu$ and to the corresponding test statistic $\hat{\tau}_\mu$. For details, see Fuller (1996, Section 10.1.2) and Hamilton (1994, Section 17.4).

For a time series that exhibits a persistent trend, it is often of interest to determine whether the trend arises from the drift term of a random walk or it is due to a deterministic trend added to a stationary AR(1) model, for example. The previous formulation of the AR(1) model with nonzero constant $z_t = \phi z_{t-1} + \theta_0 + a_t$ does not allow this, since when $|\phi| < 1$ this model implies a process with constant mean $\mu = E[z_t] = \theta_0/(1-\phi)$, independent of time. An alternate formulation of the AR(1) model that allows for a deterministic linear time trend that is not linked to ϕ is

$$z_t = \alpha + \theta_0 t + u_t \qquad \text{where} \qquad u_t = \phi u_{t-1} + a_t \qquad t = 1, \ldots, n \tag{10.1.10}$$

This model has a linear trend with slope $\theta_0 \neq 0$ regardless of whether $\phi = 1$ or $\phi \neq 1$. It is of interest to note the relation between parameters in this form relative to the previous

form. Applying the operator $(1 - \phi B)$ to (10.1.10), the model can be expressed as

$$z_t = \phi z_{t-1} + \alpha_0 + \delta_0 t + a_t \tag{10.1.11}$$

where $\alpha_0 = \alpha(1 - \phi) + \phi\theta_0$ and $\delta_0 = \theta_0(1 - \phi)$. Hence, in this form $\alpha_0 = \theta_0$ and $\delta_0 = 0$ are obtained under $\phi = 1$, so that $z_t = z_{t-1} + \theta_0 + a_t$. The presence of the linear time trend in (10.1.10) thus leads to a model with a nonzero constant but a zero coefficient for the time trend under the null hypothesis $\phi = 1$. The constant θ_0 is referred to as a drift term and measures the expected change in the series when the time increases by one unit.

A common procedure to test for a unit root in this model is to perform least-squares estimation with the linear trend term t in addition to the constant included in the regression. The resulting estimator of ϕ, denoted as $\hat{\phi}_\tau$, is such that the limiting distribution of $n(\hat{\phi}_\tau - 1)$, under $\phi = 1$, does not depend on the value of the constant $\alpha_0 = \theta_0$ but still requires the coefficient δ_0 of the time variable t to be zero under the null hypothesis. Hence, this estimator $\hat{\phi}_\tau$ can be used as the basis of a valid test of $\phi = 1$ regardless of the value of the constant θ_0. Tables of percentiles of the null distribution of $n(\hat{\phi}_\tau - 1)$ and of the corresponding Studentized statistic $\hat{\tau}_\tau$ are available in Fuller (1996, p. 642).

Alternative procedures to test $\phi = 1$ in the presence of a possible deterministic linear trend, which are valid regardless of the value of the constant term, have been proposed by several authors. Bhargava (1986) developed a locally most powerful invariant test for unit roots. Schmidt and Phillips (1992) used a score (or Lagrange multiplier (LM)) test for the model (10.1.10), and Ahn (1993) extended this approach to allow for a more general ARMA model for the noise process u_t. Elliott et al. (1996) used a point optimal testing approach with maximum power against a local alternative for the same model. The power gains were obtained by a preliminary generalized least-squares (GLS) detrending procedure using a local alternative to $\phi = 1$, followed by use of the least-squares estimate $\hat{\phi}$ and corresponding test statistic $\hat{\tau}$ obtained from the detrended series. Subsequent contributions to this area include work by Ng and Perron (2001), Perron and Qu (2007), and Harvey et al. (2009), among others.

10.1.2 Extensions of Unit Root Testing to Mixed ARIMA Models

The test procedures described above and other similar ones have been extended to testing for unit roots in mixed ARIMA$(p, 1, q)$ models (e.g., see Said and Dickey (1984, 1985) and Solo (1984b)), as well as models with higher order differencing (e.g., see Dickey and Pantula (1987)). Said and Dickey (1984) showed that the Dickey–Fuller procedure, which was originally developed for autoregressive models of known order p, remains valid asymptotically for an ARIMA$(p, 1, q)$ model where p and q are unknown. The authors approximated the mixed model by an autoregressive model of sufficiently high order and applied the ADF test to the resulting AR model. The approximation assumes that the lag length of the autoregression increases with the length of the series, n, at a controlled rate less than $n^{1/3}$. Phillips (1987) and Phillips and Perron (1988) proposed a number of unit root tests that have become popular in the econometrics literature. These tests differ from the ADF tests in how they deal with serial correlation and heteroscedasticity in the error process. Thus, while the ADF tests approximate the ARMA structure by a high-order autoregression, the Phillips and Perron tests deal with serial correlation by directly modifying the test statistics to account for serial correlation. Likelihood ratio type of unit root tests have also been considered for the mixed ARIMA model based on both conditional and unconditional normal distribution likelihoods by Yap and Reinsel (1995) and Shin and

Fuller (1998), among others. Simulation studies suggest that these tests often perform better than $\hat{\tau}$-type test statistics for mixed ARIMA models.

Motivated by problems in macroeconomics and related fields, the literature has continued to grow and many other extensions have been developed. These include the use of bootstrap methods for statistical inference as discussed, for example, by Palm et al. (2008). The use of Bayesian methods for unit root models has also been considered. The problem of distinguishing unit root nonstationary series from series with structural breaks such as level shifts or trend changes has been considered by many researchers. The methodology has also been extended and modified to deal with more complex series involving nonlinearities, time-varying volatility, and fractionally integrated processes with long-range dependence. Tests with a null hypothesis of stationarity, rather than unit root nonstationarity, have also been proposed in the literature. For further discussion and references, see, for example, Phillips and Xiao (1998) and Haldrup et al. (2013).

Example: Series C. To illustrate unit root testing, consider the series of temperature readings referred to as Series C. Two potential models identified for this series in Chapter 6 were the ARIMA(1, 1, 0) and the ARIMA(0, 2, 0). Since there is some doubt about the need for the second differencing in the ARIMA(0, 2, 0) model, with the alternative model being a stationary AR(1) for the first differences, we investigate this more formally. The AR(1) model $\nabla z_t = \phi \nabla z_{t-1} + a_t$ for the first differences can be written as $\nabla^2 z_t = (\phi - 1)\nabla z_{t-1} + a_t$, and in this form the conditional least-squares regression estimate $\hat{\phi} - 1 = -0.187$ is obtained, with an estimated standard error of 0.038, and $\hat{\sigma}_a^2 = 0.018$. Note that this implies $\hat{\phi} = 0.813$ similar to results in Tables 6.5 and 7.6. The Studentized statistic to test $\phi = 1$ is $\hat{\tau} = -4.87$, which is far more negative than the lower one percentage point of -2.58 for the distribution of $\hat{\tau}$ in the tables of Fuller (1996). Also, $\hat{\tau}_\mu = -4.96$ was obtained when a constant term is included in the AR(1) model for ∇z_t. Hence, these estimation results do not support the need for second differencing and point to a preference for the ARIMA(1, 1, 0) model.

Implementation in R. Tests for unit roots can be performed using the package fUnitRoots available in the FinTS package in R. If z represents the time series of interest, the command used to perform the augmented Dickey–Fuller test is

```
> adfTest(z,lags,type=c("nc","c","ct")
```

where lags denotes the number of lags in the autoregressive model and type indicates whether or not a constant or trend should be included in the fitted model. The argument ''nc'' specifies that no constant should be included in the model, ''c'' is used for constant only, and ''ct'' specifies a trend plus a constant. For lags equal to 0, the test is the original Dickey–Fuller test. Otherwise, lags represents the order of the stationary autoregressive polynomial in (10.1.6). For a mixed ARMA model, it represents the order of the autoregressive approximation to this model.

The calculations for Series C described above can be performed in R as follows:

```
> library(fUnitRoots)
> adfTest(diff(ts(seriesC)),0,type=c("nc"))

Title: Augmented Dickey-Fuller Test
Test Results:
```

```
  PARAMETER:
  Lag Order: 0
  STATISTIC:  Dickey-Fuller: -4.8655
  P VALUE:  0.01

> adfTest(diff(ts(seriesC)),0,type=c("c"))

  Title: Augmented Dickey-Fuller Test
  Test Results:
  PARAMETER:
  Lag Order: 0
  STATISTIC:  Dickey-Fuller: -4.962
  P VALUE:  0.01
```

The values of the test statistics agree in both cases with those quoted in the example. Note that the output shows the p value but does not give the critical value for the test. If the critical values are needed, they can be obtained in R using the command

```
>adfTable(trend=c("nc","c","ct"), statistic=c("nc","c","ct"))
```

Example: Series A. For further illustration, consider Series A that represents concentration readings of a chemical process at 2-hour intervals and has $n = 197$ observations. In Chapters 6 and 7, two possible ARMA/ARIMA models were proposed for this series. One is the nearly nonstationary ARMA(1, 1) model, $(1 - \phi B)z_t = \theta_0 + (1 - \theta B)a_t$, with estimates $\hat{\phi} = 0.92$, $\hat{\theta} = 0.58$, $\hat{\theta}_0 = 1.45$, and $\hat{\sigma}_a^2 = 0.0974$. The second is the nonstationary ARIMA(0, 1, 1) model, $(1 - B)z_t = (1 - \theta B)a_t$, with estimates $\hat{\theta} = 0.71$ and $\hat{\sigma}_a^2 = 0.1004$. Below we use the ADF test to test the hypothesis that differencing is needed so that the series follows the ARIMA(0, 1, 1) model. To determine the order k of the autoregressive approximation to this model, we first use the R command ar(z) to select a suitable value for k based on the AIC criterion. The output suggests an AR(6) model, which is then used for the test. A slightly different choice of k does not alter the conclusion.

```
  > library(fUnitRoots)
  > ar(diff(ts(seriesA)),aic=TRUE)

  Call: ar(x = diff(ts(seriesA)), aic = TRUE)
  Coefficients:
  1 2 3 4 5 6
  -0.6098 -0.3984 -0.3585 -0.3175 -0.3142 -0.2139
  Order selected 6 sigma^2 estimated as 0.09941

  > adfTest(ts(seriesA),6,type=c("nc"))

  Title: Augmented Dickey-Fuller Test
  Test Results:
  PARAMETER:
  Lag Order: 6
  STATISTIC:  Dickey-Fuller: 0.6271
  P VALUE:  0.8151
```

The p values are large and the test does not reject the null hypothesis that the series needs to be differenced, suggesting that ARIMA(0, 1, 1) is the preferred model. A similar conclusion was reached by Solo (1984b) who used a Lagrange multiplier test to determine the need for differencing.

10.2 CONDITIONAL HETEROSCEDASTIC MODELS

This section presents an overview of some models that have been developed to describe time-varying variability or volatility in a time series. To first introduce some notation, we note that the ARMA(p, q) process $\phi(B)z_t = \theta_0 + \theta(B)a_t$ can be written as the sum of a predictable part and a prediction error as

$$z_t = E[z_t|F_{t-1}] + a_t$$

where F_{t-1} represents the past information available at time $t-1$ and a_t represents the prediction error. For the ARMA model, F_{t-1} is a function of past observations and past error terms, but could more generally include external regression variables X_t. The assumption made thus far is that the prediction errors a_t are *independent* random variables with a constant variance $\text{Var}[a_t] = \sigma_a^2$ that is independent of the past. However, this assumption appears inconsistent with the heteroscedasticity often seen for time series in business and economics, in particular. For example, financial time series such as stock returns often exhibit periods when the volatility is high and periods when it is lower. This characteristic feature, or *stylized fact*, is commonly referred to as volatility clustering. For illustration, Figure 10.1(a) shows the weekly S&P 500 Index over the period January 3, 2000 to May 27, 2014 for a total of 751 observations. The log returns calculated as $\ln(p_t/p_{t-1}) = \ln(p_t) - \ln(p_{t-1})$, where p_t represents the original time series, are shown in Figure 10.1(b). We note that while the original time series is nonstationary, the returns fluctuate around a stable mean level. However, the variability around the mean changes and volatility clusters

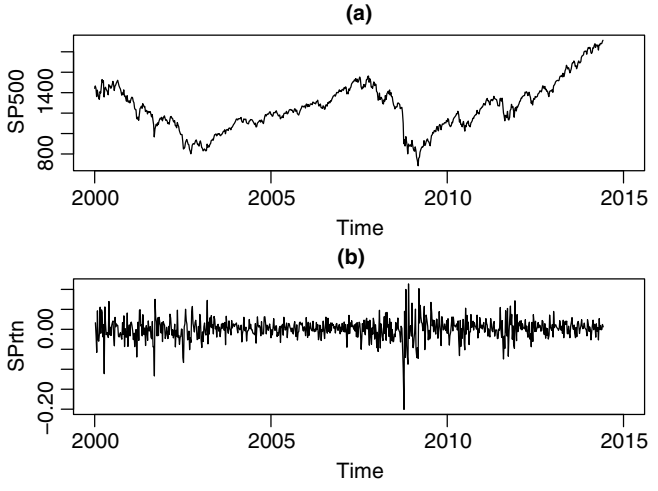

FIGURE 10.1 (a) Time plot of the weekly S&P 500 Index from January 3, 2000 to May 27, 2014, and (b) the weekly log returns on the S&P 500 Index.

are clearly visible. Note the high volatility during and following the 2008 financial crisis, in particular. Another common feature of financial time series is that the marginal distributions are leptokurtic and tend to have heavier tails than those of a normal distribution. A number of other stylized facts have been documented and investigated for financial data (for discussion and references, see, for example, Teräsvirta et al., 2010, Chapter 8).

The autoregressive conditional heteroscedastic (ARCH) model was introduced by Engle (1982) to describe time-varying variability in a series of inflation rates. An extension of this model called the generalized conditional heteroscedastic (GARCH) model was proposed by Bollerslev (1986). These models are capable of describing not only volatility clustering but also features such as heavy-tailed behavior that is common in many economic and financial time series. Still, there are other features related to volatility that are not captured by the basic ARCH and GARCH models. This has led to a number of extensions and alternative formulations aimed at addressing these issues. This section presents a brief description of the ARCH and GARCH models along with some extensions proposed in the literature. The literature in this area is extensive and only a select number of developments will be discussed. A more complete coverage can be found in survey papers by Bollerslev et al. (1992, 1994), Bera and Higgins (1993), Li et al. (2003), and Teräsvirta (2009), among others. Volatility modeling is also discussed in several time series texts, including Franses and van Dijk (2000), Mills and Markellos (2008), Teräsvirta et al. (2010), and Tsay (2010). Textbooks devoted to volatility modeling include Francq and Zakoïan (2010) and Xekalaki and Degiannakis (2010).

10.2.1 The ARCH Model

For a stationary ARMA process, the unconditional mean of the series is constant over time while the conditional mean $E[z_t|F_{t-1}]$ varies as a function of past observations. Parallel to this, the ARCH model assumes that the unconditional variance of the error process is constant over time but allows the conditional variance of a_t to vary as a function of past squared errors. Letting $\sigma_t^2 = \text{var}[a_t|F_{t-1}]$ denote the conditional variance of a_t, given the past F_{t-1}, the basic ARCH(s) model can be formulated as

$$a_t = \sigma_t e_t \tag{10.2.1}$$

where $\{e_t\}$ is a sequence of iid random variables with mean zero and variance 1, and

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_s a_{t-s}^2 \tag{10.2.2}$$

with $\alpha_0 > 0$, $\alpha_i \geq 0$, for $i = 1, \ldots, s-1$, and $\alpha_s > 0$. The parameter constraints are imposed to ensure that the conditional variance σ_t^2 is positive. The additional constraint $\sum_{i=1}^{s} \alpha_i < 1$ ensures that the a_t are covariance stationary with finite unconditional variance σ_a^2. For some time series, such as stock returns, the original observations are typically serially uncorrelated and the a_t are observed directly. Alternatively, the a_t can be the noise sequence associated with an ARMA or regression-type model. For modeling purposes, the e_t in (10.2.1) are usually assumed to follow a standard normal or a Student t-distribution.

The ARCH model was used by Engle (1982) to study the variance of UK inflation rates and by Engle (1983) to describe the variance of U.S. inflation rates. The ARCH model and its later extensions by Bollerslev (1986) and others quickly found other applications. For example, Diebold and Nerlove (1989) showed that the ARCH model may be used to generate statistically and economically meaningful measures of exchange rate volatility.

Bollerslev (1987) used the GARCH extension of the ARCH model to analyze the conditional volatility of financial returns observed at a monthly or higher frequency. In Weiss (1984), ARMA models with ARCH errors were used to model the time series behavior of 13 different U.S. macroeconomic time series. Bollerslev et al. (1992) describe a large number of other applications in their review of volatility models. While a majority of applications have been in finance and economics, the models have also been used in other fields. For example, Campbell and Diebold (2005) used volatility models in their analysis of the daily average temperatures for four U.S. cities. The models have also been used for variables such as wind speeds, air quality measurements, earthquake series, and in the analysis of speech signals. For selected references, see Francq and Zekoïan (2010, p. 12).

Some Properties of the ARCH Model. To establish some properties of the ARCH model, we first examine the ARCH(1) model where

$$\sigma_t^2 = \text{var}[a_t|F_{t-1}] = E[a_t^2 \mid F_{t-1}] = \alpha_0 + \alpha_1 a_{t-1}^2 \tag{10.2.3}$$

with $\alpha_0 > 0$ and $\alpha_1 > 0$. The form of the model shows that the conditional variance σ_t^2 will be large if a_{t-1} was large in absolute value and vice versa. A large (small) value of σ_t^2 will in turn tend to generate a large (small) value of a_t, thus giving rise to volatility clustering.

It follows from (10.2.1) that $E\,[a_t \mid F_{t-1}] = 0$. The *unconditional* mean of a_t is also zero since

$$E[a_t] = E\left[E[a_t \mid F_{t-1}]\right] = 0$$

Furthermore, the a_t are serially uncorrelated since for $j > 0$,

$$E[a_t a_{t-j}] = E\left[E[a_t a_{t-j} \mid F_{t-1}]\right] = E\left[a_{t-j} E[a_t \mid F_{t-1}]\right] = 0$$

But the a_t are not mutually independent since they are interrelated through their conditional variances. The lack of serial correlation is an important property that makes the ARCH model suitable for modeling asset returns that are expected to be uncorrelated by the efficient market hypothesis.

We also assume that the a_t have equal *unconditional* variances, $\text{var}[a_t] = E[a_t^2] = \sigma_a^2$, for all t, so that the process is weakly stationary. If $\alpha_1 < 1$, the unconditional variance exists and equals

$$\sigma_a^2 = \text{var}[a_t] = \frac{\alpha_0}{1 - \alpha_1} \tag{10.2.4}$$

This follows since

$$\sigma_a^2 = E[a_t^2] = E\left[E[a_t^2 \mid F_{t-1}]\right] = E[\alpha_0 + \alpha_1 a_{t-1}^2] = \alpha_0 + \alpha_1 \sigma_a^2$$

Further substituting $\alpha_0 = \sigma_a^2(1 - \alpha_1)$ from (10.2.4) into (10.2.3), we see that

$$\sigma_t^2 = \sigma_a^2 + \alpha_1(a_{t-1}^2 - \sigma_a^2) \tag{10.2.5}$$

or, equivalently, $\sigma_t^2 - \sigma_a^2 = \alpha_1(a_{t-1}^2 - \sigma_a^2)$. Hence, the conditional variance of a_t will be above the unconditional variance whenever a_{t-1}^2 is larger than the unconditional variance σ_a^2.

To study the tail behavior of a_t, we examine the fourth moment $\mu_4 = E[a_t^4]$. If a_t is normally distributed, conditional on the past, then

$$E[a_t^4 \mid F_{t-1}] = 3\sigma_t^4 = 3(\alpha_0 + \alpha_1 a_{t-1}^2)^2$$

Therefore, the fourth unconditional moment of a_t satisfies

$$E[a_t^4] = E\left[E[a_t^4 \mid F_{t-1}]\right] = 3\left[\alpha_0^2 + 2\alpha_0\alpha_1 E[a_{t-1}^2] + \alpha_1^2 E[a_{t-1}^4]\right]$$

Thus, if $\{a_t\}$ is fourth-order stationary so that $\mu_4 = E[a_t^4] = E[a_{t-1}^4]$, then

$$\mu_4 = \frac{3(\alpha_0^2 + 2\alpha_0\alpha_1\sigma_a^2)}{1 - 3\alpha_1^2} \equiv \frac{3\alpha_0^2(1-\alpha_1^2)}{(1-\alpha_1)^2(1-3\alpha_1^2)} \tag{10.2.6}$$

Since $\mu_4 = E[a_t^4] > 0$, this expression shows that α_1 must satisfy $0 < \alpha_1 < 1/\sqrt{3}$ in order for a_t to have finite fourth moment. Further, if κ denotes the unconditional kurtosis of a_t, then

$$\kappa = \frac{E[a_t^4]}{\left(E[a_t^2]\right)^2} = \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2}$$

This value exceeds 3, the kurtosis of the normal distribution. Hence, the marginal distribution of a_t has heavier tails than those of the normal distribution. This is an additional feature of the ARCH model that makes it useful for modeling financial asset returns where heavy-tailed behavior is the norm.

To derive an alternative form of the ARCH process, we let $v_t = a_t^2 - \sigma_t^2$, so that $a_t^2 = \sigma_t^2 + v_t$. The random variables v_t then have zero mean and they are serially uncorrelated since

$$\begin{aligned} E[(a_t^2 - \sigma_t^2)(a_{t-j}^2 - \sigma_{t-j}^2)] &= E[E\{(a_t^2 - \sigma_t^2)(a_{t-j}^2 - \sigma_{t-j}^2) \mid F_{t-1}\}] \\ &= E[(a_{t-j}^2 - \sigma_{t-j}^2)E\{(a_t^2 - \sigma_t^2) \mid F_{t-1}\}] = 0 \end{aligned}$$

Further, since $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2$, we find that the ARCH(1) model can be written as

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + v_t \tag{10.2.7}$$

This form reveals that the process of squared errors a_t^2 can be viewed as an AR(1) model with uncorrelated innovations v_t. The innovations are heteroscedastic and also non-Gaussian in this case, however.

For the ARCH(s) model in (10.2.2), we similarly have

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_s a_{t-s}^2 + v_t$$

so that the a_t^2 has the form of an AR(s) process. Other results related to the moments and the kurtosis of the ARCH(1) model also extend to higher order ARCH models. In particular, if $\sum_{i=1}^{s} \alpha_i < 1$, then the unconditional variance is

$$\sigma_a^2 = \frac{\alpha_0}{1 - \sum_{i=1}^{s} \alpha_i}$$

as shown by Engle (1982). Necessary and sufficient conditions for the existence of higher order even moments of the ARCH(s) process were given by Milhøj (1985).

Forecast Errors for the ARCH Model. Forecasts of a future value z_{t+l} generated from ARMA models with iid errors a_t have forecast errors that depend on the lead time l but are independent of the time origin t from which the forecasts are made. Baillie and Bollerslev (1992) showed that the minimum mean square error forecasts of z_{t+l} are the same irrespective of whether the shocks a_t are heteroscedastic or not. For an ARMA process with ARCH errors, this implies, in particular, that the one-step-ahead forecast error equals a_{t+1} while the l-step-ahead forecast error can be written as $e_t(l) = \sum_{j=0}^{l-1} \psi_j a_{t+l-j}$ with $\psi_0 = 1$. The presence of conditional heteroscedasticity will, however, impact the variance of the forecast errors.

For an ARCH(1) process, the conditional variance of the one-step-ahead forecast error a_{t+1} is given by (10.2.5) as

$$E[e_t^2(1)|F_t] = \sigma_{t+1}^2 = \sigma_a^2 + \alpha_1(a_t^2 - \sigma_a^2) \tag{10.2.8}$$

The conditional variance of the one-step-ahead forecast error can thus be smaller or larger than the unconditional variance depending on the difference between the last squared error a_t^2 and σ_a^2.

Conditional variances of multistep-ahead forecast errors $e_t(l)$ can also be shown to depend on the past squared errors based on

$$E[e_t^2(l) \mid F_t] = \sum_{j=0}^{l-1} \psi_j^2 E[a_{t+l-j}^2 \mid F_t]$$

where for the ARCH(1) model

$$\begin{aligned} E[a_{t+h}^2|F_t] &= E[E(a_{t+h}^2 \mid F_t)] \\ &= \alpha_0 + \alpha_1 E[a_{t+h-1}^2 \mid F_t] \\ &= \alpha_0(1 + \alpha_1 + \cdots + \alpha_1^{h-1}) + \alpha_1^h a_t^2 \qquad \text{for} \quad h > 0 \end{aligned}$$

From this and using (10.2.4) it can be verified that

$$E[e_t^2(l) \mid F_t] = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2 + \sum_{j=0}^{l-1} \psi_j^2 \alpha_1^{l-j}(a_t^2 - \sigma_a^2) \tag{10.2.9}$$

which simplifies to (10.2.8), for $l = 1$. The first term on the right-hand side of this expression is the conventional prediction error variance assuming that the errors a_t are homoscedastic while the second term reflects the impact of the ARCH effects. This term varies over time and can again be positive or negative depending on the difference $a_t^2 - \sigma_a^2$. The variance of the predicted values thus varies over time and can be larger or smaller than that under homoscedasticity. For the general ARCH(s) model, the second term on the right-hand side will be a function of s past values $a_t^2, \ldots, a_{t-s+1}^2$.

If the time series z_t follows an AR(1) model, the ψ weights are given by $\psi_j = \phi^{j-1}$. If ϕ equals zero, so that the mean of the series is a constant independent of the past, expression (10.2.9) simplifies to $\sigma_a^2 + \alpha_1^l(a_t^2 - \sigma_a^2)$. We note that this is the conditional l-step-ahead forecast of the conditional variance σ_{t+l}^2 for the ARCH(1) model. This forecast

could be calculated more directly as $E[\sigma^2_{t+l}|F_t] = \alpha_0 + \alpha_1 E[a^2_{t+l-1}|F_t]$, where $E[a^2_{t+l-1}|F_t]$ can be generated recursively from the AR model for a^2_t. The result follows by setting $\alpha_0 = \sigma^2_a(1-\alpha_1)$.

10.2.2 The GARCH Model

The ARCH model has a disadvantage in that it often requires a high lag order s to adequately describe the evolution of volatility over time. An extension of the ARCH model called the *generalized* ARCH, or GARCH, model was introduced by Bollerslev (1986) to overcome this issue. The GARCH(s, r) model assumes that $a_t = \sigma_t e_t$, where the $\{e_t\}$ again are iid random variables with mean zero and variance 1, and where σ_t is given by

$$\sigma^2_t = \alpha_0 + \sum_{i=1}^{s} \alpha_i a^2_{t-i} + \sum_{j=1}^{r} \beta_j \sigma^2_{t-j} \tag{10.2.10}$$

with $\alpha_0 > 0$, $\alpha_i \geq 0$, $i = 1, \ldots, s-1$, $\alpha_s > 0$, $\beta_j \geq 0$, $j = 1, \ldots, r-1$, and $\beta_r > 0$. These parameter constraints are sufficient for the conditional variance σ^2_t to be positive. Nelson and Cao (1992) showed that these constraints can be relaxed slightly to allow some of the parameters to be negative while the conditional variance still remains positive. The additional constraint $\sum_{i=1}^{m}(\alpha_i + \beta_i) < 1$, where $m = \max(s, r)$ with $\alpha_i = 0$, for $i > s$, and $\beta_j = 0$, for $j > r$, ensures that the unconditional variance σ^2_a is finite.

The simplest and most widely used model in this class is the GARCH(1, 1) model where

$$\sigma^2_t = E[a^2_t \mid F_{t-1}] = \alpha_0 + \alpha_1 a^2_{t-1} + \beta_1 \sigma^2_{t-1}$$

Since the constants α_1 and β_1 are positive, we see that a large value of a^2_{t-1} or σ^2_{t-1} results in a large value of σ^2_t. As for the ARCH process, this model therefore accounts for volatility clustering.

Assuming that $\alpha_1 + \beta_1 < 1$, the unconditional variance of a_t is

$$\sigma^2_a = \text{var}[a_t] = \alpha_0/[1 - (\alpha_1 + \beta_1)]$$

Also, assuming that the conditional distributions are normal, the fourth unconditional moment of a_t is finite provided that $(\alpha_1 + \beta_1)^2 + 2\alpha^2_1 < 1$ (Bollerslev, 1986). In addition, the kurtosis of the marginal distribution of a_t equals

$$\kappa = \frac{E(a^4_t)}{[E(a^2_t)]^2} = \frac{3[1 - (\alpha_1 + \beta_1)^2]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha^2_1} > 3$$

As in the ARCH case, the unconditional distribution of a_t thus has heavier tails than the normal distribution and is expected to give rise to a higher frequency of extreme observations or ''outliers'' than would be the case under normality.

Now let $v_t = a^2_t - \sigma^2_t$ so that $\sigma^2_t = a^2_t - v_t$, where the v_t have zero mean and are serially uncorrelated. We then see that the GARCH(1, 1) model can be rearranged as $a^2_t - v_t = \alpha_0 + \alpha_1 a^2_{t-1} + \beta_1(a^2_{t-1} - v_{t-1})$, or

$$a^2_t = \alpha_0 + (\alpha_1 + \beta_1)a^2_{t-1} + v_t - \beta_1 v_{t-1} \tag{10.2.11}$$

The process of *squared errors* thus has the form of an ARMA(1, 1) model with uncorrelated innovations v_t. The v_t are in general heteroscedastic, however. In the special case of $\beta_1 = 0$, the model reduces to $a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + v_t$, which is the AR(1) form of the ARCH(1) model. For the general GARCH(s, r) process, expression (10.2.11) generalizes to

$$a_t^2 = \alpha_0 + \sum_{i=1}^{m}(\alpha_i + \beta_i)a_{t-i}^2 + v_t - \sum_{i=1}^{s}\beta_i v_{t-i}$$

which has the form of an ARMA process for a_t^2 with AR order equal to $m = \max(r, s)$. The autocorrelation structure of a_t^2 also mimics that of the ARMA process provided that fourth unconditional moment of a_t is finite (Bollerslev, 1988).

The necessary and sufficient condition for second-order stationarity of the GARCH(s, r) process is

$$\sum_{i=1}^{s}\alpha_i + \sum_{i=1}^{r}\beta_i = \sum_{i=1}^{m}(\alpha_i + \beta_i) < 1$$

When this condition is met, the unconditional variance is

$$\sigma_a^2 = \text{var}[a_t] = \alpha_0 \Big/ \left[1 - \sum_{i=1}^{m}(\alpha_i + \beta_i)\right]$$

This was shown by Bollerslev (1986) who also gave necessary and sufficient conditions for the existence of all higher order moments for the GARCH(1, 1) model and the fourth-order moments for GARCH(1, 2) and GARCH(2, 1) models. Extensions of these results have been given by He and Teräsvirta (1999) and Ling and McAleer (2002), among others. The expressions for the higher order moments and the constraints on the parameters needed to ensure their existence become more complex for the higher order models. The model specification also becomes more difficult. On the other hand, numerous studies have shown that low-order models such as the GARCH(1, 1), GARCH(2, 1), and GARCH(1, 2) models are often adequate in practice, with the GARCH(1, 1) model being the most popular.

10.2.3 Model Building and Parameter Estimation

Testing for ARCH/GARCH Effects. The preceding results motivate the use of the ACF and PACF of the squares a_t^2 for model specification and for basic preliminary checking for the presence of ARCH/GARCH effects in the errors a_t. For an ARMA model with heteroscedastic errors, a starting point for the analysis is an examination of the sample ACF and PACF of the squared residuals $\hat{a}_t^2$ obtained from fitting an ARMA model to the observed series. In particular, let $r_k(\hat{a}^2)$ denote the sample autocorrelations of the squared residuals $\hat{a}_t^2$ so that

$$r_k(\hat{a}^2) = \sum_{t=1}^{n-k}(\hat{a}_t^2 - \hat{\sigma}_a^2)(\hat{a}_{t+k}^2 - \hat{\sigma}_a^2) \Big/ \sum_{t=1}^{n}(\hat{a}_t^2 - \hat{\sigma}_a^2)^2$$

where $\hat{\sigma}_a^2 = n^{-1}\sum_{t=1}^{n}\hat{a}_t^2$ is the residual variance estimate. Analogous to the modified portmanteau statistic described in Section 8.2.2, McLeod and Li (1983) proposed the

portmanteau statistic

$$\tilde{Q}(\hat{a}^2) = n(n+2)\sum_{k=1}^{K} r_k^2(\hat{a}^2)/(n-k) \tag{10.2.12}$$

to detect departures from the ARMA assumptions. As a portmanteau test, this test does not assume a specific alternative, but the type of departures for which $\tilde{Q}(\hat{a}^2)$ can be useful includes conditional heteroscedasticity in the form of ARCH/GARCH effects, and bilinear type of nonlinearity in the conditional mean of the process (see Section 10.3 for discussion of bilinear models). McLeod and Li (1983) showed that the statistic $\tilde{Q}(\hat{a}^2)$ has approximately the χ^2 distribution with K degrees of freedom under the assumption that the ARMA model alone is adequate. The distribution is similar to that of the usual portmanteau statistic $\tilde{Q}$ based on the residuals $\hat{a}_t$, with the exception that the degrees of freedom in the case of (10.2.12) are *not affected* by the fact that $p+q$ ARMA parameters have been estimated. The potentially more powerful portmanteau statistics by Peña and Rodríguez (2002, 2006) discussed in Section 8.2 could also be applied to the squared residuals $\hat{a}_t^2$.

An alternative test for ARCH effects is the score or Lagrange multiplier test proposed by Engle (1982). The score statistic Λ for testing the null hypothesis H_0: $\alpha_i = 0$, $i = 1, \ldots, s$, has a convenient form and can be expressed as n times the coefficient of determination in the least-squares fitting of the auxiliary regression equation

$$\hat{a}_t^2 = \alpha_0 + \alpha_1\hat{a}_{t-1}^2 + \alpha_2\hat{a}_{t-2}^2 + \cdots + \alpha_s\hat{a}_{t-s}^2 + \varepsilon_t$$

Assuming normality of the a_t's, the score statistic Λ has an asymptotic χ^2 distribution with s degrees of freedom under the null model of no ARCH effects. The test procedure is thus to fit a time series model to the observed series, save the residuals $\hat{a}_t$, and regress the squared residuals on a constant and s lagged values of the $\hat{a}_t^2$. The resulting value of nR^2 is then referred to a χ^2 distribution with s degrees of freedom. Even though this test was derived for the ARCH(s) model, it has been shown to be useful for detecting other forms of conditional heteroscedasticity as well. Also, the test is asymptotically equivalent to the McLeod–Li portmanteau test based on the autocorrelations of the squared residuals (see Luukkonen et al., 1988b). Thus, although the latter was derived as a pure significance test, it is also a LM test against ARCH effects.

Parameter Estimation. The parameter estimation for models with ARCH or GARCH errors is typically performed using the conditional maximum likelihood method. For estimation of an ARMA model $\phi(B)z_t = \theta_0 + \theta(B)a_t$ with ARCH or GARCH errors a_t, we assume that a_t is conditionally normally distributed as $N(0, \sigma_t^2)$. The z_t are then conditionally normal, given $z_{t-1}, z_{t-2}, \ldots$, and from the joint density function $p(\boldsymbol{z}) = \prod_{t=1}^{n} p(z_t \mid z_{t-1}, \ldots, z_1)$ we obtain the log-likelihood function

$$l = \log(L) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log(\sigma_t^2) - \frac{1}{2}\sum_{t=1}^{n} a_t^2/\sigma_t^2 \tag{10.2.13}$$

where $a_t = z_t - \sum_{i=1}^{p}\phi_i z_{t-i} - \theta_0 + \sum_{i=1}^{q}\theta_i a_{t-i}$ and σ_t^2 is given by (10.2.2) or (10.2.10). A discussion of the iterative maximization of the likelihood function along with other results related to the parameter estimation can be found, for example, in Engle (1982), Weiss (1984, 1986), and Bollerslev (1986). When an ARMA model with ARCH or GARCH errors

is fitted to the series, the information matrix of the log-likelihood is block diagonal with respect to the conditional mean and variance parameters, so that iterations can be carried out separately with respect to the two sets of parameters. The so-called BHHH algorithm by Berndt, Hall, Hall, and Hausman (1974) provides a convenient method to perform the calculations. This algorithm has the advantage that only first-order derivatives are needed for the optimization. These derivatives can be evaluated numerically or analytically. Use of analytical first derivatives is often recommended as they improve the precision of the parameter estimates. Provided that the fourth-order moment of the process is finite, the resulting estimates of the ARMA–ARCH parameters are consistent and asymptotically normal as shown by Weiss (1986).

The normal distribution was originally proposed by Engle (1982) to model the conditional distribution of the disturbances a_t. As discussed earlier, the conditional normal distribution results in a leptokurtic unconditional distribution. Nevertheless, in financial applications the normal distribution sometimes fails to capture the excess kurtosis that is present in stock returns and other variables. To overcome this drawback, Bollerslev (1987) suggested using a standardized Student t-distribution with $\nu > 2$ degrees of freedom for the estimation. The density function of the t-distribution is

$$f(x|\nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\left(1+\frac{x^2}{(\nu-2)}\right)^{-(\nu+1)/2}$$

where $\Gamma(\nu) = \int_0^\infty e^{-x}x^{\nu-1}dx$ is the Gamma function and ν measures the tail thickness. As is well known, the distribution is symmetric around zero and approaches a normal distribution as $\nu \to \infty$. For $\nu > 4$, the fourth moment exists and the conditional kurtosis equals $3(\nu-2)/(\nu-4)$. Since this value exceeds 3, the tails are heavier than those of the normal distribution. The log-likelihood function based on the t-distribution is given by

$$\begin{aligned} l = \log(L) = n\left[\log\Gamma\left(\frac{\nu+1)}{2}\right) - \log\left(\frac{\nu}{2}\right) - \frac{1}{2}\log(\pi(\nu-2))\right] \\ -\frac{1}{2}\sum_{t=1}^{n}\left[\log(\sigma_t^2) + (1+\nu)\log\left(1+\frac{a_t^2}{(\nu-2)\sigma_t^2}\right)\right] \end{aligned}$$

Here, ν is either prespecified or estimated jointly with other parameters. If ν is specified in advance, values between 5 and 8 are often used; see Tsay (2010). With ν prespecified, the conditional likelihood function is maximized by minimizing the second term of the likelihood function given above.

Nelson (1991) suggested using the generalized error distribution (GED) for the estimation. The density function of a GED random variable normalized to have mean zero and variance one is given by

$$f(x|\eta) = \frac{\eta\exp(-0.5|x/\lambda|^\eta)}{\lambda 2^{(1+1/\eta)}\Gamma(1/\eta)}$$

where $\lambda = [2^{(-2/\eta)}\Gamma(1/\eta)/\Gamma(3/\eta)]^{1/2}$. For the tail thickness parameter $\eta = 2$, the distribution equals the normal distribution used in (10.2.13). For $\eta < 2$, the distribution has thicker tails than the normal distribution. The reverse is true for $\eta > 2$. Box and Tiao (1973) call the GED distribution an exponential power distribution.

In addition to having excess kurtosis, the distribution of a_t may also be skewed. A discussion of potential sources for skewness can be found in He et al. (2008). To allow for skewness as well as heavy tails, the likelihood calculations can be based on skewed versions for the Student t-distribution and the GED distributions available in software packages such as R. Other forms of skewed distributions have also been considered.

In practice, it is often difficult to know whether the specified probability distribution is the correct one. An alternative approach is to continue to base the parameter estimation on the normal likelihood function in (10.2.13). This method is commonly referred to as the quasi-maximum likelihood (QML) estimation. The asymptotic properties of the resulting QML estimator for the ARCH, GARCH, and ARMA–GARCH models have been studied by many authors with early contributions provided by Weiss (1986) and Bollerslev and Wooldridge (1992). For further discussion and references, see, for example, Francq and Zakoïan (2009, 2010).

Diagnostic Checking. Methods for model checking include informal graphical checks using time series plots and Q–Q plots of the residuals along with a study of their dependence structure. The assumption underlying the ARCH and GARCH models is that the standardized innovations a_t/σ_t are independent and identically distributed. Having estimated the parameters of model, the adequacy of the mean value function can be checked by examining the autocorrelation and partial autocorrelation functions of the standardized residuals $\hat{a}_t/\hat{\sigma}_t$. Similar checks on the autocorrelation and partial autocorrelations of the squared standardized residuals are useful for examining the adequacy of the volatility model. These checks are often supplemented by the portmanteau test proposed by McLeod and Li (1983) or the score test proposed by Engle (1982). However, while these statistics can provide useful indications of lack of fit, their asymptotic distributions are impacted by the estimation of the ARCH or GARCH parameters. Li and Mak (1994) derived an alternative portmanteau statistic that asymptotically follows the correct χ^2_K distribution. This statistic is a quadratic form in the first m autocorrelations of the squared standardized residuals but has a more complex form than the $\tilde{Q}$ statistic in (10.2.12). Analogous modifications of Engle's score test based on ARCH residuals were discussed by Lundbergh and Teräsvirta (2002). More recent contributions to model checking include work by Wong and Ling (2005), Ling and Tong (2011), Fisher and Gallagher (2012), and many others.

10.2.4 An Illustrative Example: Weekly S&P 500 Log Returns

To demonstrate the model building process, we consider the weekly log returns on the S&P 500 Index displayed in Figure 10.1(b) for the period January 3, 2000 to May 27, 2014. Figure 10.2 shows the ACF of the returns along with the ACF of the squared returns. We note that there is little, if any, serial correlation in the returns themselves. The mean value function μ_t will thus be taken as a constant. However, the squared returns are clearly correlated and show a pattern consistent with that of an ARCH or a GARCH model. The PACF of the squared returns (not shown) has a pattern that persists over several lags suggesting that a GARCH may be appropriate for the volatility.

The parameters can be estimated in R using the function garchFit() in the fGarch package. The normal distribution is the default error distribution for the ARCH or GARCH models. Other options include the Student t-distribution and the GED distributions along with skewed versions of these distributions. For demonstration, we will fit a GARCH(1, 1)

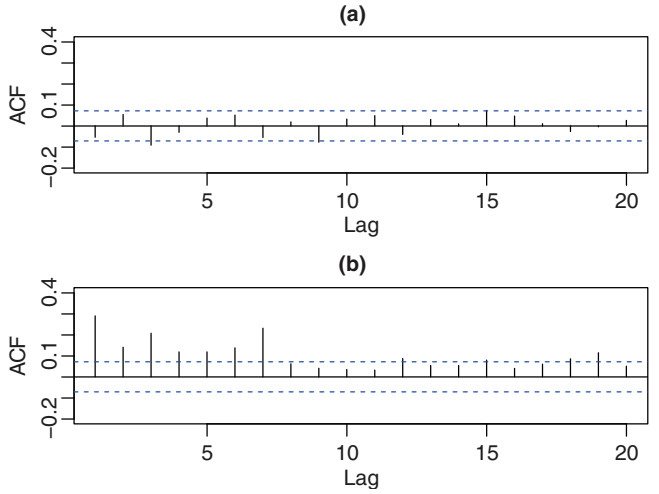

FIGURE 10.2 Autocorrelation functions for (a) the S&P 500 weekly log returns and (b) the squared weekly log returns.

model with normal errors to the returns. The R commands and a partial model output are provided below, where the log returns are denoted by SPrtn:

```
  >library(fGarch)
  >m1=garchFit(~garch(1,1),data=SPrtn,trace=F)
  >summary(m1) % Retrieve model output

Title: GARCH Modelling
Call: garchFit(formula=~garch(1,1),data=SPrtn, trace=F)

Mean and Variance Equation:  data ~ garch(1,1)
Conditional Distribution:  norm

Coefficient(s):
     mu       omega      alpha1       beta1
 2.1875e-03  3.5266e-05  2.1680e-01  7.3889e-01

 Error Analysis:
        Estimate   Std. Error   t value   Pr(>|t|)
 mu     2.187e-03   6.875e-04    3.182    0.00146 **
 omega  3.527e-05   1.153e-05    3.058    0.00223 **
 alpha1 2.168e-01   4.189e-02    5.176   2.27e-07 ***
 beta1  7.389e-01   4.553e-02   16.230    < 2e-16 ***

 Standardised Residuals Tests:
                                Statistic    p-Value
 Jarque-Bera Test      Chi^2    77.92548       0
 Shapiro-Wilk Test  R    W       0.9815283 3.990011e-08
 Ljung-Box Test     R    Q(10)   6.910052  0.7339084
 Ljung-Box Test     R    Q(20)   16.43491  0.689303
```

```
Ljung-Box Test      R^2  Q(10)    12.64346  0.244295
Ljung-Box Test      R^2  Q(20)    18.15442  0.5772367
LM Arch Test        R    TR^2     14.05565  0.297169

Information Criterion Statistics:
  AIC         BIC         SIC       HQIC
-4.751772 -4.727132 -4.751829 -4.742278
```

Letting w_t denote the log returns, the fitted model is

$$w_t = 0.002187 + a_t, \quad \sigma_t^2 = 0.000035 + 0.2168a_{t-1}^2 + 0.7389\sigma_{t-1}^2$$

where all the parameter estimates are statistically significant. The portmanteau tests for serial correlation in the standardized residuals and in their squared values indicate no lack of fit. However, the Jarque–Bera and Shapiro–Wilk tests for normality suggest that the model is not fully adequate. To examine this issue, the Student t-distribution and its skewed version were tested by adding the argument cond.dist="std" and cond.dist ="sstd", respectively, to the garchFit command. The GED distribution and its skewed version were also tested. Although these modifications improved the fit, the results are for simplicity not shown here.

The standardized residuals from the fitted model and the ACF of the squared standardized residuals are shown in Figure 10.3. A normal Q–Q plot is also included in this graph. Visual inspection of the standardized residuals and the Q–Q plot confirms the results of the normality tests discussed above. The ACF of the squared residuals indicates no lack of fit although a marginally significant correlation is present at lag 1. This value would be reduced by fitting a GARCH(1, 2) model to the data. But this potential refinement is not pursued here. Finally, estimates of the conditional standard deviation σ_t are displayed in Figure 10.4(a). Figure 10.4(b) displays the volatility shown earlier in Figure 10.1(b) with two standard deviation limits now superimposed around the series. A variety of other graphs can be generated using the R command plot(m1), where m1 refers to the fitted model. In addition, l-step-ahead forecasts of future volatility based on the conditional standard deviations shown in Figure 10.4 can be generated using the R command predict(m1,l).

10.2.5 Extensions of the ARCH and GARCH Models

While the ARCH and GARCH models allow for volatility clustering and capture thick-tailed behavior of the underlying unconditional distributions, they do not account for certain other features that are commonly observed in financial data. For example, so-called leverage effects are often observed in stock returns, where a negative innovation tends to increase the volatility more than a positive innovation of the same magnitude. In symmetric ARCH and GARCH models, on the other hand, the variance depends on the magnitude of the innovations but not their signs. Another limitation of the basic ARCH and GARCH models is the assumption that the conditional mean of the process is unaffected by the volatility. This assumption ignores the so-called risk premium that relates to the fact that investors expect to receive higher returns as compensation for taking on riskier assets. The presence of this feature would generate a positive relationship between expected return and volatility. Below we describe some extensions and modifications of the ARCH and GARCH models that have been proposed to address such issues.