

6

Nonstationary and Seasonal Time Series Models

- 6.1 ARIMA Models for Nonstationary Time Series
- 6.2 Identification Techniques
- 6.3 Unit Roots in Time Series Models
- 6.4 Forecasting ARIMA Models
- 6.5 Seasonal ARIMA Models
- 6.6 Regression with ARMA Errors

In this chapter we shall examine the problem of finding an appropriate model for a given set of observations $\{x_1, \dots, x_n\}$ that are not necessarily generated by a stationary time series. If the data (a) exhibit no apparent deviations from stationarity and (b) have a rapidly decreasing autocovariance function, we attempt to fit an ARMA model to the mean-corrected data using the techniques developed in Chapter 5. Otherwise, we look first for a transformation of the data that generates a new series with the properties (a) and (b). This can frequently be achieved by differencing, leading us to consider the class of ARIMA (autoregressive integrated moving-average) models, defined in Section 6.1. We have in fact already encountered ARIMA processes. The model fitted in Example 5.1.1 to the Dow Jones Utilities Index was obtained by fitting an AR model to the differenced data, thereby effectively fitting an ARIMA model to the original series. In Section 6.1 we shall give a more systematic account of such models.

In Section 6.2 we discuss the problem of finding an appropriate transformation for the data and identifying a satisfactory ARMA(p, q) model for the transformed data. The latter can be handled using the techniques developed in Chapter 5. The sample ACF and PACF and the preliminary estimators $\hat{\phi}_m$ and $\hat{\theta}_m$ of Section 5.1 can provide useful guidance in this choice. However, our prime criterion for model selection will be the AICC statistic discussed in Section 5.5.2. To apply this criterion we compute maximum likelihood estimators of ϕ , θ , and σ^2 for a variety of competing p and q values and choose the fitted model with smallest AICC value. Other techniques, in particular those that use the R and S arrays of Gray et al. (1978), are discussed in the survey of model identification by de Gooijer et al. (1985). If the fitted model is

satisfactory, the residuals (see Section 5.3) should resemble white noise. Tests for this were described in Section 5.3 and should be applied to the minimum AICC model to make sure that the residuals are consistent with their expected behavior under the model. If they are not, then competing models (models with AICC value close to the minimum) should be checked until we find one that passes the goodness of fit tests. In some cases a small difference in AICC value (say less than 2) between two satisfactory models may be ignored in the interest of model simplicity. In Section 6.3 we consider the problem of testing for a unit root of either the autoregressive or moving-average polynomial. An autoregressive unit root suggests that the data require differencing, and a moving-average unit root suggests that they have been overdifferenced. Section 6.4 considers the prediction of ARIMA processes, which can be carried out using an extension of the techniques developed for ARMA processes in Sections 3.3 and 5.4. In Section 6.5 we examine the fitting and prediction of seasonal ARIMA (SARIMA) models, whose analysis, except for certain aspects of model identification, is quite analogous to that of ARIMA processes. Finally, we consider the problem of regression, allowing for dependence between successive residuals from the regression. Such models are known as regression models with time series residuals and often occur in practice as natural representations for data containing both trend and serially dependent errors.

6.1 ARIMA Models for Nonstationary Time Series

We have already discussed the importance of the class of ARMA models for representing stationary series. A generalization of this class, which incorporates a wide range of nonstationary series, is provided by the ARIMA processes, i.e., processes that reduce to ARMA processes when differenced finitely many times.

Definition 6.1.1

If d is a nonnegative integer, then $\{X_t\}$ is an **ARIMA(p, d, q) process** if $Y_t := (1 - B)^d X_t$ is a causal ARMA(p, q) process.

This definition means that $\{X_t\}$ satisfies a difference equation of the form

$$\phi^*(B)X_t \equiv \phi(B)(1 - B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad (6.1.1)$$

where $\phi(z)$ and $\theta(z)$ are polynomials of degrees p and q , respectively, and $\phi(z) \neq 0$ for $|z| \leq 1$. The polynomial $\phi^*(z)$ has a zero of order d at $z = 1$. The process $\{X_t\}$ is stationary if and only if $d = 0$, in which case it reduces to an ARMA(p, q) process.

Notice that if $d \geq 1$, we can add an arbitrary polynomial trend of degree $(d - 1)$ to $\{X_t\}$ without violating the difference equation (6.1.1). ARIMA models are therefore useful for representing data with trend (see Sections 1.5 and 6.2). It should be noted, however, that ARIMA processes can also be appropriate for modeling series with no trend. Except when $d = 0$, the mean of $\{X_t\}$ is not determined by equation (6.1.1), and it can in particular be zero (as in Example 1.3.3). Since for $d \geq 1$, equation (6.1.1) determines the second-order properties of $\{(1 - B)^d X_t\}$ but not those of $\{X_t\}$ (Problem 6.1), estimation of ϕ , θ , and σ^2 will be based on the observed differences $(1 - B)^d X_t$. Additional assumptions are needed for prediction (see Section 6.4).

Example 6.1.1 $\{X_t\}$ is an ARIMA(1,1,0) process if for some $\phi \in (-1, 1)$,

$$(1 - \phi B)(1 - B)X_t = Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

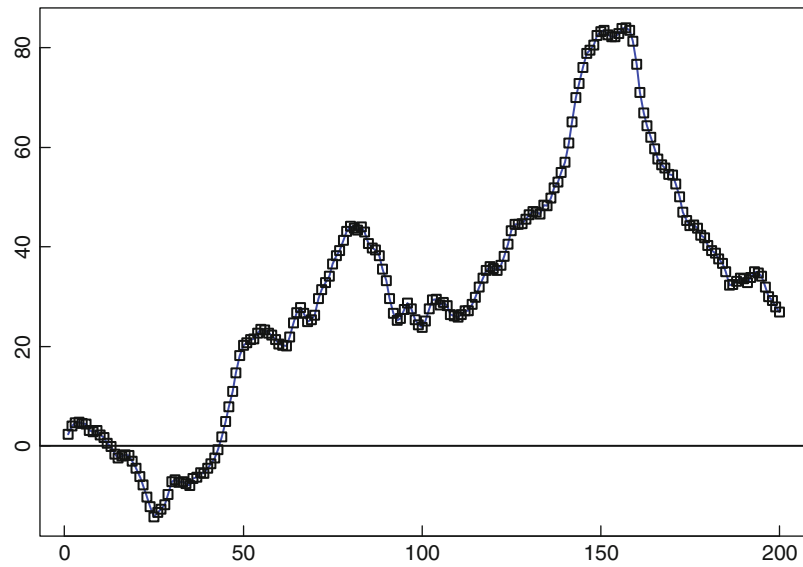


Figure 6-1
200 observations of the
ARIMA(1,1,0) series
 X_t of Example 6.1.1

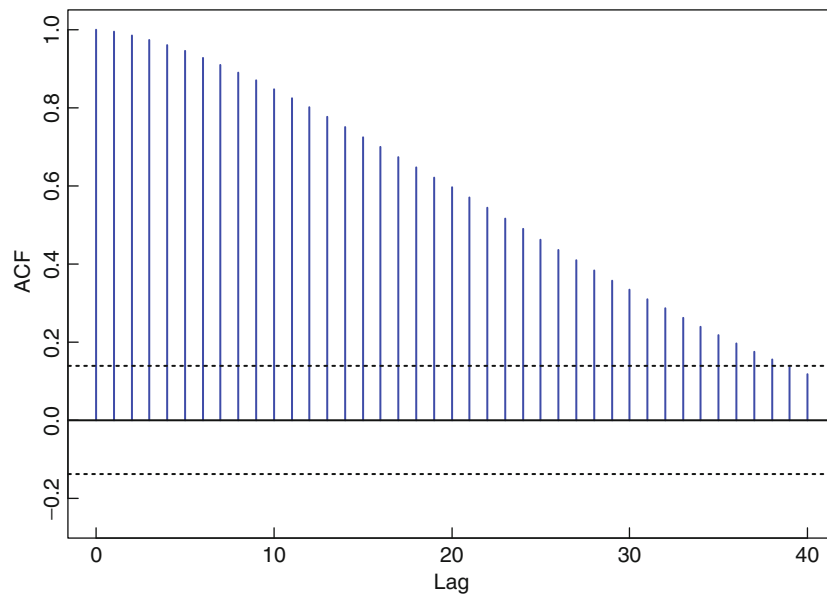


Figure 6-2
The sample ACF of the
data in Figure 6-1

We can then write

$$X_t = X_0 + \sum_{j=1}^t Y_j, \quad t \geq 1,$$

where

$$Y_t = (1 - B)X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

A realization of $\{X_1, \dots, X_{200}\}$ with $X_0 = 0$, $\phi = 0.8$, and $\sigma^2 = 1$ is shown in Figure 6-1, with the corresponding sample autocorrelation and partial autocorrelation functions in Figures 6-2 and 6-3, respectively. \square

A distinctive feature of the data that suggests the appropriateness of an ARIMA model is the slowly decaying positive sample autocorrelation function in Figure 6-2.

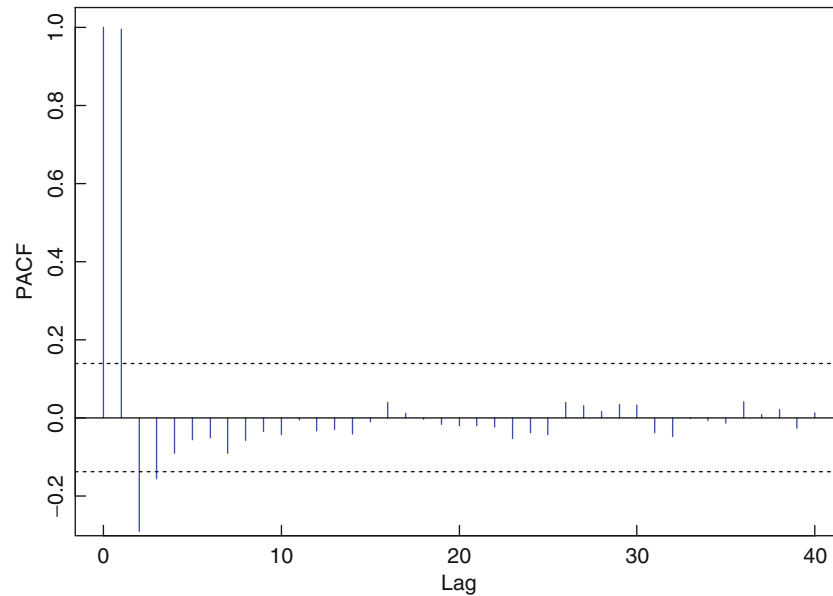


Figure 6-3
The sample PACF of
the data in Figure 6-1

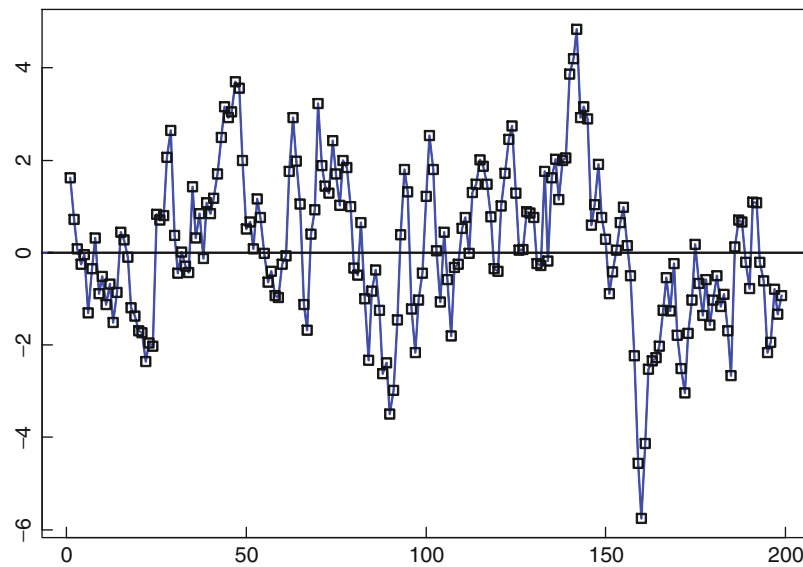


Figure 6-4
199 observations of the
series $Y_t = \nabla X_t$ with
 $\{X_t\}$ as in Figure 6-1

If, therefore, we were given only the data and wished to find an appropriate model, it would be natural to apply the operator $\nabla = 1 - B$ repeatedly in the hope that for some j , $\{\nabla^j X_t\}$ will have a rapidly decaying sample autocorrelation function compatible with that of an ARMA process with no zeros of the autoregressive polynomial near the unit circle. For this particular time series, one application of the operator ∇ produces the realization shown in Figure 6-4, whose sample ACF and PACF (Figures 6-5 and 6-6) suggest an AR(1) [or possibly AR(2)] model for $\{\nabla X_t\}$. The maximum likelihood estimates of ϕ and σ^2 obtained from ITSM under the assumption that $E(\nabla X_t) = 0$ (found by *not* subtracting the mean after differencing the data) are 0.808 and 0.978, respectively, giving the model

$$(1 - 0.808B)(1 - B)X_t = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.978), \quad (6.1.2)$$

which bears a close resemblance to the true underlying process,

$$(1 - 0.8B)(1 - B)X_t = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 1). \quad (6.1.3)$$

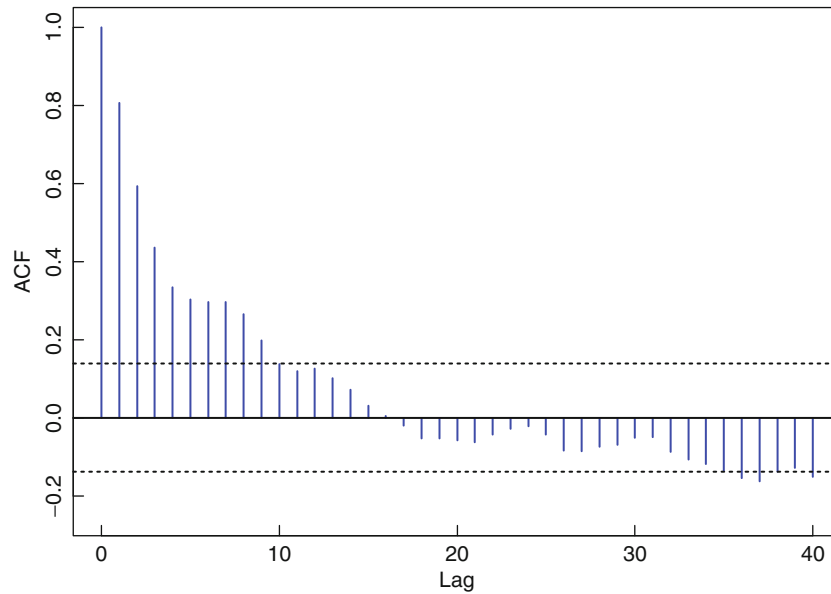


Figure 6-5
The sample ACF of the series $\{Y_t\}$ in Figure 6-4

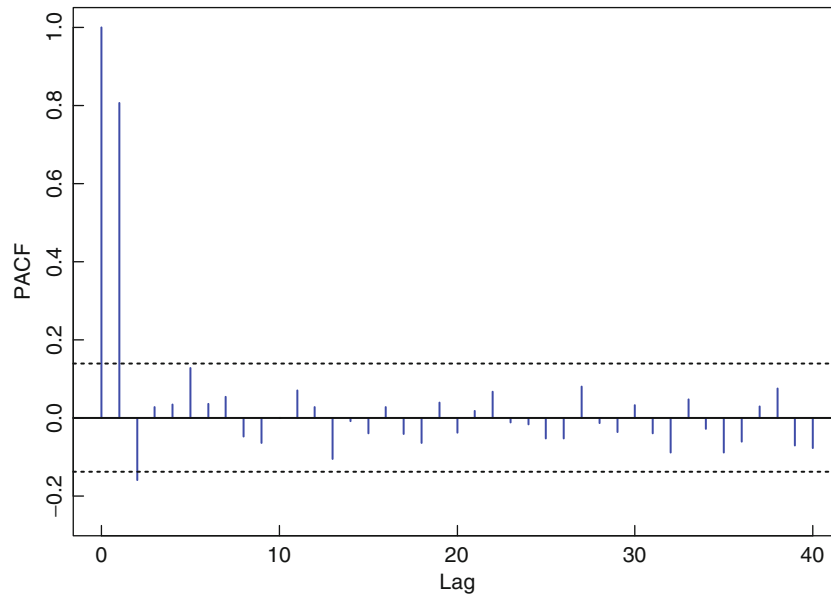


Figure 6-6
The sample PACF of the series $\{Y_t\}$ in Figure 6-4

Instead of differencing the series in Figure 6-1 we could proceed more directly by attempting to fit an AR(2) process as suggested by the sample PACF of the original series in Figure 6-3. Maximum likelihood estimation, carried out using ITSM after fitting a preliminary model with Burg's algorithm and assuming that $EX_t = 0$, gives the model

$$(1 - 1.808B + 0.811B^2)X_t = (1 - 0.825B)(1 - 0.983B)X_t = Z_t, \\ \{Z_t\} \sim \text{WN}(0, 0.970), \quad (6.1.4)$$

which, although stationary, has coefficients closely resembling those of the true nonstationary process (6.1.3). (To obtain the model (6.1.4), two optimizations were carried out using the `Model>Estimation>Max likelihood` option of ITSM, the first with the default settings and the second after setting the accuracy parameter to 0.00001.) From a sample of finite length it will be extremely difficult to distinguish between a nonstationary process such as (6.1.3), for which $\phi^*(1) = 0$, and a process such as (6.1.4), which has very similar coefficients but for which ϕ^* has all of its

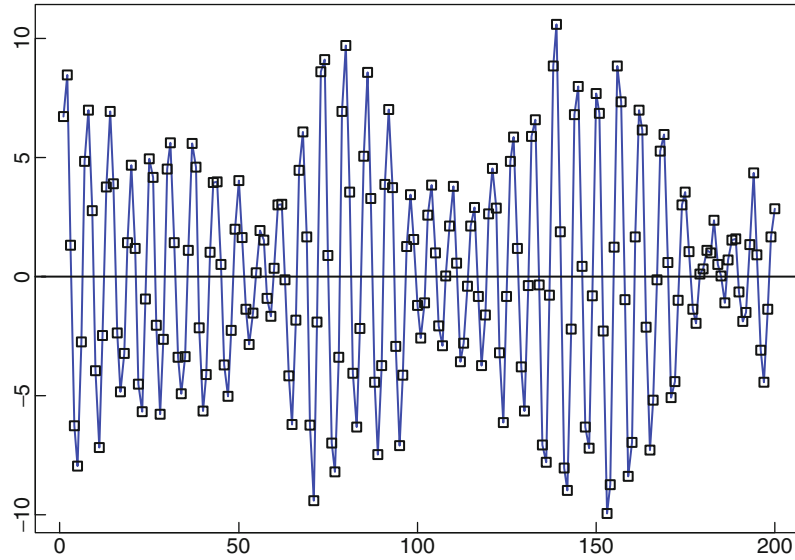


Figure 6-7
200 observations of
the AR(2) process
defined by (6.1.6) with
 $r = 1.005$ and $\omega = \pi/3$

zeros outside the unit circle. In either case, however, if it is possible by differencing to generate a series with rapidly decaying sample ACF, then the differenced data set can be fitted by a low-order ARMA process whose autoregressive polynomial ϕ^* has zeros that are comfortably outside the unit circle. This means that the fitted parameters will be well away from the boundary of the allowable parameter set. This is desirable for numerical computation of parameter estimates and can be quite critical for some methods of estimation. For example, if we apply the Yule–Walker equations to fit an AR(2) model to the data in Figure 6-1, we obtain the model

$$(1 - 1.282B + 0.290B^2)X_t = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 6.435), \quad (6.1.5)$$

which bears little resemblance to either the maximum likelihood model (6.1.4) or the true model (6.1.3). In this case the matrix \hat{R}_2 appearing in (5.1.7) is nearly singular.

An obvious limitation in fitting an ARIMA(p, d, q) process $\{X_t\}$ to data is that $\{X_t\}$ is permitted to be nonstationary only in a very special way, i.e., by allowing the polynomial $\phi^*(B)$ in the representation $\phi^*(B)X_t = Z_t$ to have a zero of multiplicity d at the point 1 on the unit circle. Such models are appropriate when the sample ACF is a slowly decaying positive function as in Figure 6-2, since sample autocorrelation functions of this form are associated with models $\phi^*(B)X_t = \theta(B)Z_t$ in which ϕ^* has a zero either at or close to 1.

Sample autocorrelations with slowly decaying oscillatory behavior as in Figure 6-8 are associated with models $\phi^*(B)X_t = \theta(B)Z_t$ in which ϕ^* has a zero close to $e^{i\omega}$ for some $\omega \in (-\pi, \pi]$ other than 0. Figure 6-8 is the sample ACF of the series of 200 observations in Figure 6-7, obtained from ITSM by simulating the AR(2) process

$$X_t - (2r^{-1} \cos \omega)X_{t-1} + r^{-2}X_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 1), \quad (6.1.6)$$

with $r = 1.005$ and $\omega = \pi/3$, i.e.,

$$X_t - 0.9950X_{t-1} + 0.9901X_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 1).$$

The autocorrelation function of the model (6.1.6) can be derived by noting that

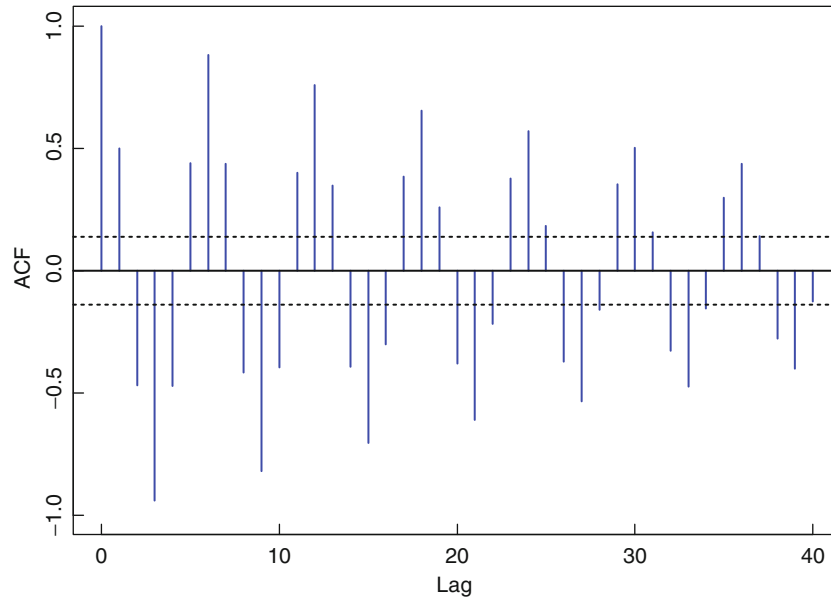


Figure 6-8
The sample ACF of the
data in Figure 6-7

$$1 - (2r^{-1} \cos \omega) B + r^{-2} B^2 = (1 - r^{-1} e^{i\omega} B) (1 - r^{-1} e^{-i\omega} B) \quad (6.1.7)$$

and using (3.2.12). This gives

$$\rho(h) = r^{-h} \frac{\sin(h\omega + \psi)}{\sin \psi}, \quad h \geq 0, \quad (6.1.8)$$

where

$$\tan \psi = \frac{r^2 + 1}{r^2 - 1} \tan \omega. \quad (6.1.9)$$

It is clear from these equations that

$$\rho(h) \rightarrow \cos(h\omega) \text{ as } r \downarrow 1. \quad (6.1.10)$$

With $r = 1.005$ and $\omega = \pi/3$ as in the model generating Figure 6-7, the model ACF (6.1.8) is a damped sine wave with damping ratio $1/1.005$ and period 6. These properties are reflected in the sample ACF shown in Figure 6-8. For values of r closer to 1, the damping will be even slower as the model ACF approaches its limiting form (6.1.10).

If we were simply given the data shown in Figure 6-7, with no indication of the model from which it was generated, the slowly damped sinusoidal sample ACF with period 6 would suggest trying to make the sample ACF decay more rapidly by applying the operator (6.1.7) with $r = 1$ and $\omega = \pi/3$, i.e., $(1 - B + B^2)$. If it happens, as in this case, that the period $2\pi/\omega$ is close to some integer s (in this case 6), then the operator $1 - B^s$ can also be applied to produce a series with more rapidly decaying autocorrelation function (see also Section 6.5). Figures 6-9 and 6-10 show the sample autocorrelation functions obtained after applying the operators $1 - B + B^2$ and $1 - B^6$, respectively, to the data shown in Figure 6-7. For either one of these two differenced series, it is then not difficult to fit an ARMA model $\phi(B)X_t = \theta(B)Z_t$ for which the zeros of ϕ are well outside the unit circle. Techniques for identifying and determining such ARMA models have already been introduced in Chapter 5. For convenience we shall collect these together in the following sections with a number of illustrative examples.

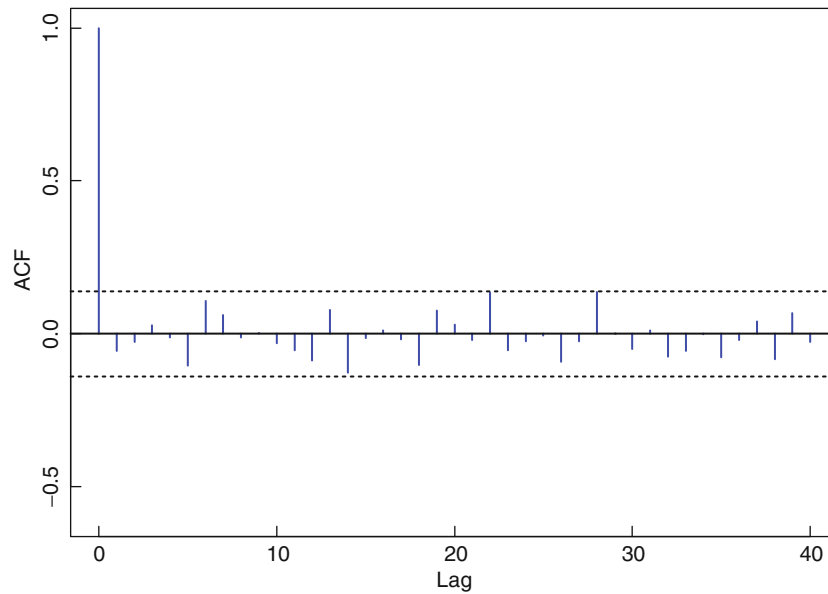


Figure 6-9
The sample ACF of $(1 - B + B^2)X_t$ with $\{X_t\}$ as in Figure 6-7

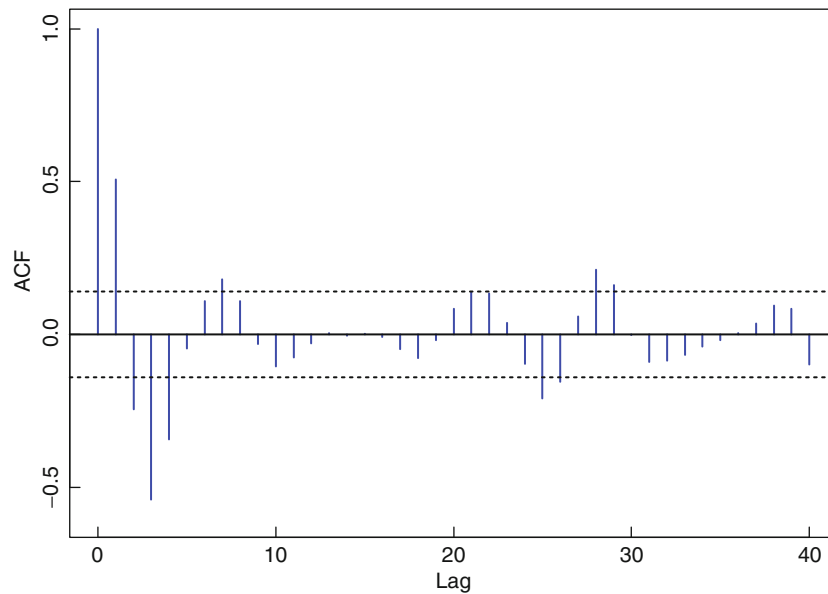


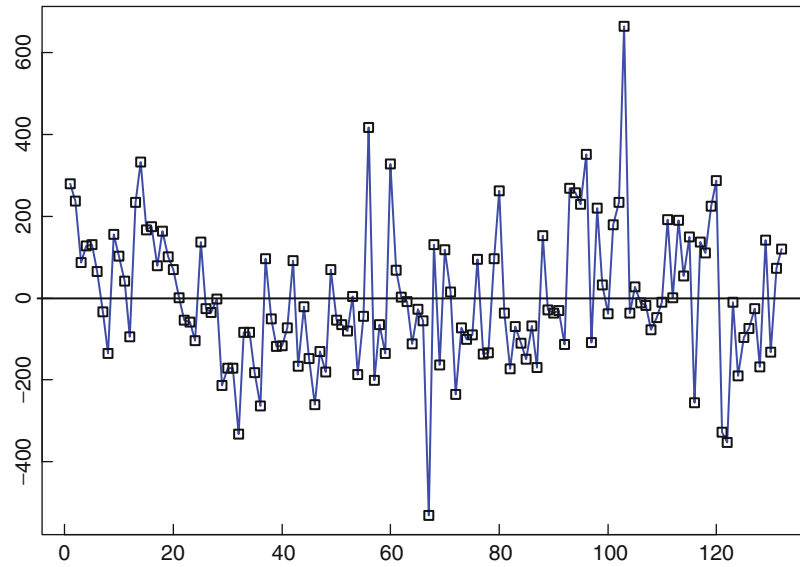
Figure 6-10
The sample ACF of $(1 - B^6)X_t$ with $\{X_t\}$ as in Figure 6-7

6.2 Identification Techniques

(a) *Preliminary Transformations.* The estimation methods of Chapter 5 enable us to find, for given values of p and q , an $ARMA(p, q)$ model to fit a given series of data. For this procedure to be meaningful it must be at least plausible that the data are in fact a realization of an ARMA process and in particular a realization of a stationary process. If the data display characteristics suggesting nonstationarity (e.g., trend and seasonality), then it may be necessary to make a transformation so as to produce a new series that is more compatible with the assumption of stationarity.

Deviations from stationarity may be suggested by the graph of the series itself or by the sample autocorrelation function or both.

Figure 6-11
The Australian red wine data after taking natural logarithms and removing a seasonal component of period 12 and a linear trend



Inspection of the graph of the series will occasionally reveal a strong dependence of variability on the level of the series, in which case the data should first be transformed to reduce or eliminate this dependence. For example, Figure 1-1 shows the Australian monthly red wine sales from January 1980 through October 1991, and Figure 1-17 shows how the increasing variability with sales level is reduced by taking natural logarithms of the original series. The logarithmic transformation $V_t = \ln U_t$ used here is in fact appropriate whenever $\{U_t\}$ is a series whose standard deviation increases linearly with the mean. For a systematic account of a general class of variance-stabilizing transformations, we refer the reader to Box and Cox (1964). The defining equation for the general Box–Cox transformation f_λ is

$$f_\lambda(U_t) = \begin{cases} \lambda^{-1}(U_t^\lambda - 1), & U_t \geq 0, \lambda > 0, \\ \ln U_t, & U_t > 0, \lambda = 0, \end{cases}$$

and the program ITSM provides the option (Transform>Box-Cox) of applying f_λ (with $0 \leq \lambda \leq 1.5$) prior to the elimination of trend and/or seasonality from the data. In practice, if a Box–Cox transformation is necessary, it is often the case that either f_0 or $f_{0.5}$ is adequate.

Trend and seasonality are usually detected by inspecting the graph of the (possibly transformed) series. However, they are also characterized by autocorrelation functions that are slowly decaying and nearly periodic, respectively. The elimination of trend and seasonality was discussed in Section 1.5, where we described two methods:

- (i) “classical decomposition” of the series into a trend component, a seasonal component, and a random residual component, and
- (ii) differencing.

The program ITSM (in the Transform option) offers a choice between these techniques. The results of applying methods (i) and (ii) to the transformed red wine data $V_t = \ln U_t$ in Figure 1-17 are shown in Figures 6-11 and 6-12, respectively. Figure 6-11 was obtained from ITSM by estimating and removing from $\{V_t\}$ a linear trend component and a seasonal component with period 12. Figure 6-12 was obtained by applying the operator $(1 - B^{12})$ to $\{V_t\}$. Neither of the two resulting series displays

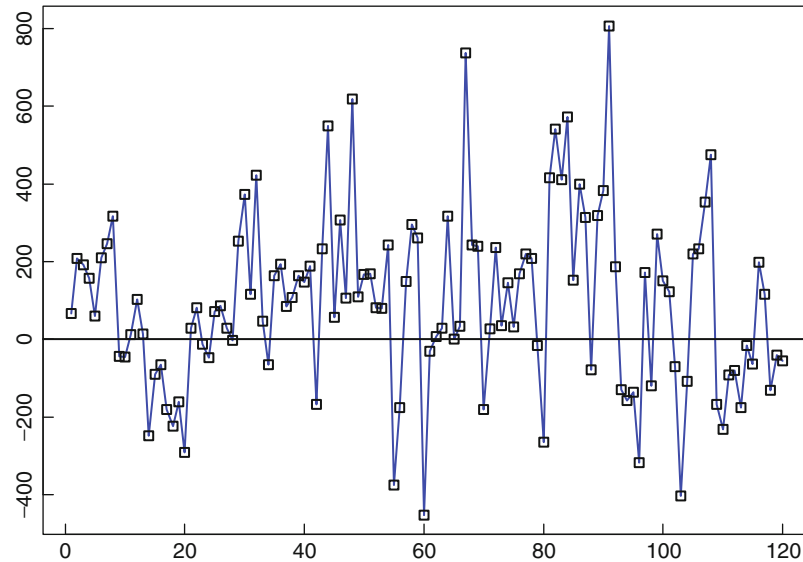


Figure 6-12
The Australian red wine data after taking natural logarithms and differencing at lag 12

any apparent deviations from stationarity, nor do their sample autocorrelation functions. The sample ACF and PACF of $\{(1 - B^{12})V_t\}$ are shown in Figures 6-13 and 6-14, respectively.

After the elimination of trend and seasonality, it is still possible that the sample autocorrelation function may appear to be that of a nonstationary (or nearly nonstationary) process, in which case further differencing may be carried out.

In Section 1.5 we also mentioned a third possible approach:

- (iii) fitting a sum of harmonics and a polynomial trend to generate a noise sequence that consists of the residuals from the regression.

In Section 6.6 we discuss the modifications to classical least squares regression analysis that allow for dependence among the residuals from the regression. These modifications are implemented in the ITSM option `Regression>Estimation>Generalized LS`.

(b) Identification and Estimation. Let $\{X_t\}$ be the mean-corrected transformed series found as described in (a). The problem now is to find the most satisfactory $\text{ARMA}(p, q)$ model to represent $\{X_t\}$. If p and q were known in advance, this would be a straightforward application of the estimation techniques described in Chapter 5. However, this is usually not the case, so it becomes necessary also to identify appropriate values for p and q .

It might appear at first sight that the higher the values chosen for p and q , the better the resulting fitted model will be. However, as pointed out in Section 5.5, estimation of too large a number of parameters introduces estimation errors that adversely affect the use of the fitted model for prediction as illustrated in Section 5.4. We therefore minimize one of the model selection criteria discussed in Section 5.5 in order to choose the values of p and q . Each of these criteria includes a penalty term to discourage the fitting of too many parameters. We shall base our choice of p and q primarily on the minimization of the AICC statistic, defined as

$$\text{AICC}(\phi, \theta) = -2 \ln L(\phi, \theta, S(\phi, \theta)/n) + 2(p + q + 1)n/(n - p - q - 2), \quad (6.2.1)$$

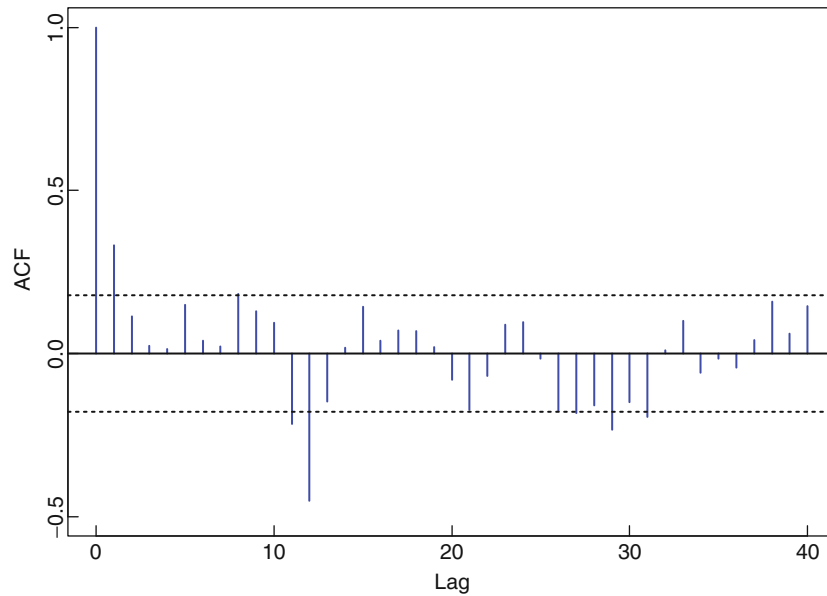


Figure 6-13
The sample ACF of the data in Figure 6-12

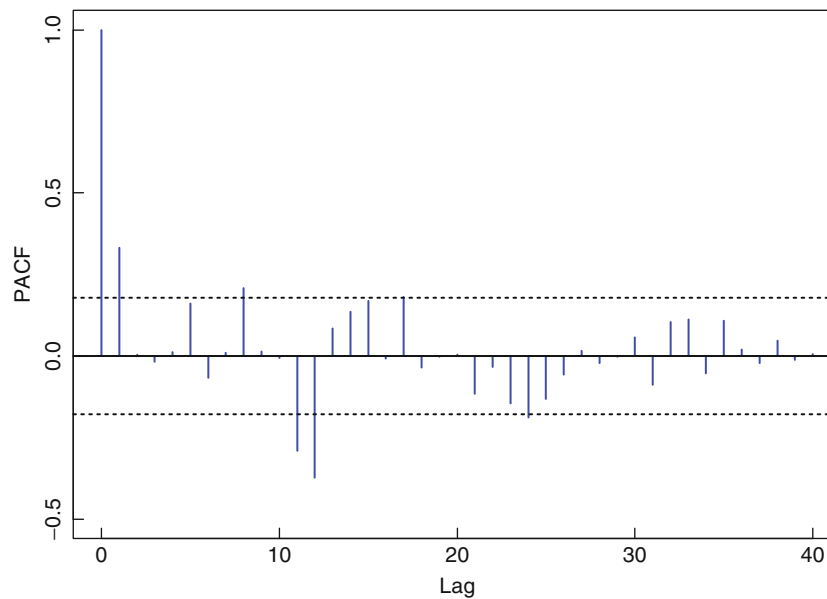


Figure 6-14
The sample PACF of the data in Figure 6-12

where $L(\phi, \theta, \sigma^2)$ is the likelihood of the data under the Gaussian ARMA model with parameters (ϕ, θ, σ^2) , and $S(\phi, \theta)$ is the residual sum of squares defined in (5.2.11). Once a model has been found that minimizes the AICC value, it is then necessary to check the model for goodness of fit (essentially by checking that the residuals are like white noise) as discussed in Section 5.3.

For any fixed values of p and q , the maximum likelihood estimates of ϕ and θ are the values that minimize the AICC. Hence, the minimum AICC model (over any given range of p and q values) can be found by computing the maximum likelihood estimators for each fixed p and q and choosing from these the maximum likelihood model with the smallest value of AICC. This can be done with the program ITSM by using the option `Model>Estimation>Autofit`. When this option is selected and upper and lower bounds for p and q are specified, the program fits maximum likelihood models for each pair (p, q) in the range specified and

selects the model with smallest AICC value. If some of the coefficient estimates are small compared with their estimated standard deviations, maximum likelihood subset models (with those coefficients set to zero) can also be explored.

The steps in model identification and estimation can be summarized as follows:

- After transforming the data (if necessary) to make the fitting of an ARMA(p, q) model reasonable, examine the sample ACF and PACF to get some idea of potential p and q values. Preliminary estimation using the ITSM option `Model>Estimation>Preliminary` is also useful in this respect. Burg's algorithm with AICC minimization rapidly fits autoregressions of all orders up to 27 and selects the one with minimum AICC value. For preliminary estimation of models with $q > 0$, each pair (p, q) must be considered separately.
- Select the option `Model>Estimation>Autofit` of ITSM. Specify the required limits for p and q , and the program will then use maximum likelihood estimation to find the minimum AICC model with p and q in the range specified.
- Examination of the fitted coefficients and their standard errors may suggest that some of them can be set to zero. If this is the case, then a **subset model** can be fitted by clicking on the button `Constrain optimization` in the `Maximum Likelihood Estimation` dialog box and setting the selected coefficients to zero. Optimization will then give the maximum likelihood model with the chosen coefficients constrained to be zero. The constrained model is assessed by comparing its AICC value with those of the other candidate models.
- Check the candidate model(s) for goodness of fit as described in Section 5.3. These tests can be performed by selecting the option `Statistics>Residual Analysis`.

Example 6.2.1 The Australian Red Wine Data

Let $\{X_1, \dots, X_{130}\}$ denote the series obtained from the red wine data of Example 1.1.1 after taking natural logarithms, differencing at lag 12, and subtracting the mean (0.0681) of the differences. The data prior to mean correction are shown in Figure 6-12. The sample PACF of $\{X_t\}$, shown in Figure 6-14, suggests that an AR(12) model might be appropriate for this series. To explore this possibility we use the ITSM option `Model>Estimation>Preliminary` with Burg's algorithm and AICC minimization. As anticipated, the fitted Burg models do indeed have minimum AICC when $p = 12$. The fitted model is

$$(1 - 0.245B - 0.069B^2 - 0.012B^3 - 0.021B^4 - 0.200B^5 + 0.025B^6 + 0.004B^7 - 0.133B^8 + 0.010B^9 - 0.095B^{10} + 0.118B^{11} + 0.384B^{12})X_t = Z_t,$$

with $\{Z_t\} \sim \text{WN}(0, 0.0135)$ and AICC value -158.77 . Selecting the option `Model>Estimation>Max likelihood` then gives the maximum likelihood AR(12) model, which is very similar to the Burg model and has AICC value -158.87 . Inspection of the standard errors of the coefficient estimators suggests the possibility of setting those at lags 2,3,4,6,7,9,10, and 11 equal to zero. If we do this by clicking on the `Constrain optimization` button in the `Maximum Likelihood Estimation` dialog box and then reoptimize, we obtain the model,

$$(1 - 0.270B - 0.224B^5 - 0.149B^8 + 0.099B^{11} + 0.353B^{12})X_t = Z_t,$$

with $\{Z_t\} \sim \text{WN}(0, 0.0138)$ and AICC value -172.49 .

In order to check more general ARMA(p, q) models, select the option `Model>Estimation>Autofit` and specify the minimum and maximum values of p and

q to be zero and 15, respectively. (The sample ACF and PACF suggest that these limits should be more than adequate to include the minimum AICC model.) In a few minutes (depending on the speed of your computer) the program selects an ARMA(1,12) model with AICC value -172.74 , which is slightly better than the subset AR(12) model just found. Inspection of the estimated standard deviations of the MA coefficients at lags 1, 3, 4, 6, 7, 9, and 11 suggests setting them equal to zero and reestimating the values of the remaining coefficients. If we do this by clicking on the **Constrain optimization** button in the **Maximum Likelihood Estimation** dialog box, setting the required coefficients to zero and then reoptimizing, we obtain the model,

$$(1 - 0.286B)X_t = \left(1 + 0.127B^2 + 0.183B^5 + 0.177B^8 + 0.181B^{10} - 0.554B^{12}\right)Z_t,$$

with $\{Z_t\} \sim \text{WN}(0, 0.0120)$ and AICC value -184.09 .

The subset ARMA(1,12) model easily passes all the goodness of fit tests in the `Statistics>Residual Analysis` option. In view of this and its small AICC value, we accept it as a plausible model for the transformed red wine series. \square

Example 6.2.2 The Lake Data

Let $\{Y_t, t = 1, \dots, 99\}$ denote the lake data of Example 1.3.5. We have seen already in Example 5.2.5 that the ITSM option `Model>Estimation>Autofit` gives the minimum-AICC model

$$X_t - 0.7446X_{t-1} = Z_t + 0.3213Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 0.4750),$$

for the mean-corrected series $X_t = Y_t - 9.0041$. The corresponding AICC value is 212.77. Since the model passes all the goodness of fit tests, we accept it as a reasonable model for the data. \square

6.3 Unit Roots in Time Series Models

The unit root problem in time series arises when either the autoregressive or moving-average polynomial of an ARMA model has a root on or near the unit circle. A unit root in either of these polynomials has important implications for modeling. For example, a root near 1 of the autoregressive polynomial suggests that the data should be differenced before fitting an ARMA model, whereas a root near 1 of the moving-average polynomial indicates that the data were overdifferenced. In this section, we consider inference procedures for detecting the presence of a unit root in the autoregressive and moving-average polynomials.

6.3.1 Unit Roots in Autoregressions

In Section 6.1 we discussed the use of differencing to transform a nonstationary time series with a slowly decaying sample ACF and values near 1 at small lags into one with a rapidly decreasing sample ACF. The degree of differencing of a time series $\{X_t\}$ was largely determined by applying the difference operator repeatedly until the sample ACF of $\{\nabla^d X_t\}$ decays quickly. The differenced time series could then be modeled by a low-order ARMA(p, q) process, and hence the resulting ARIMA(p, d, q) model

for the original data has an autoregressive polynomial $(1 - \phi_1 z - \dots - \phi_p z^p)(1 - z)^d$ [see (6.1.1)] with d roots on the unit circle. In this subsection we discuss a more systematic approach to testing for the presence of a unit root of the autoregressive polynomial in order to decide whether or not a time series should be differenced. This approach was pioneered by Dickey and Fuller (1979).

Let X_1, \dots, X_n be observations from the AR(1) model

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad (6.3.1)$$

where $|\phi_1| < 1$ and $\mu = EX_t$. For large n , the maximum likelihood estimator $\hat{\phi}_1$ of ϕ_1 is approximately $N(\phi_1, (1 - \phi_1^2)/n)$. For the unit root case, this normal approximation is no longer applicable, even asymptotically, which precludes its use for testing the unit root hypothesis $H_0 : \phi_1 = 1$ vs. $H_1 : \phi_1 < 1$. To construct a test of H_0 , write the model (6.3.1) as

$$\nabla X_t = X_t - X_{t-1} = \phi_0^* + \phi_1^* X_{t-1} + Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad (6.3.2)$$

where $\phi_0^* = \mu(1 - \phi_1)$ and $\phi_1^* = \phi_1 - 1$. Now let $\hat{\phi}_1^*$ be the ordinary least squares (OLS) estimator of ϕ_1^* found by regressing ∇X_t on 1 and X_{t-1} . The estimated standard error of $\hat{\phi}_1^*$ is

$$\widehat{\text{SE}}(\hat{\phi}_1^*) = S \left(\sum_{t=2}^n (X_{t-1} - \bar{X})^2 \right)^{-1/2},$$

where $S^2 = \sum_{t=2}^n (\nabla X_t - \hat{\phi}_0^* - \hat{\phi}_1^* X_{t-1})^2 / (n - 3)$ and \bar{X} is the sample mean of X_1, \dots, X_{n-1} . Dickey and Fuller derived the limit distribution as $n \rightarrow \infty$ of the t -ratio

$$\hat{\tau}_\mu := \hat{\phi}_1^* / \widehat{\text{SE}}(\hat{\phi}_1^*) \quad (6.3.3)$$

under the unit root assumption $\phi_1^* = 0$, from which a test of the null hypothesis $H_0 : \phi_1 = 1$ can be constructed. The 0.01, 0.05, and 0.10 quantiles of the limit distribution of $\hat{\tau}_\mu$ (see Table 8.5.2 of Fuller 1976) are -3.43 , -2.86 , and -2.57 , respectively. The augmented Dickey–Fuller test then rejects the null hypothesis of a unit root, at say, level 0.05 if $\hat{\tau}_\mu < -2.86$. Notice that the cutoff value for this test statistic is much smaller than the standard cutoff value of -1.645 obtained from the normal approximation to the t -distribution, so that the unit root hypothesis is less likely to be rejected using the correct limit distribution.

The above procedure can be extended to the case where $\{X_t\}$ follows the AR(p) model with mean μ given by

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \dots + \phi_p(X_{t-p} - \mu) + Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

This model can be rewritten as (see Problem 6.2)

$$\nabla X_t = \phi_0^* + \phi_1^* X_{t-1} + \phi_2^* \nabla X_{t-1} + \dots + \phi_p^* \nabla X_{t-p+1} + Z_t, \quad (6.3.4)$$

where $\phi_0 = \mu(1 - \phi_1 - \dots - \phi_p)$, $\phi_1^* = \sum_{i=1}^p \phi_i - 1$, and $\phi_j^* = -\sum_{i=j}^p \phi_i$, $j = 2, \dots, p$. If the autoregressive polynomial has a unit root at 1, then $0 = \phi(1) = -\phi_1^*$, and the differenced series $\{\nabla X_t\}$ is an AR($p - 1$) process. Consequently, testing the hypothesis of a unit root at 1 of the autoregressive polynomial is equivalent to testing $\phi_1^* = 0$. As in the AR(1) example, ϕ_1^* can be estimated as the coefficient of X_{t-1} in the OLS regression of ∇X_t onto 1, X_{t-1} , ∇X_{t-1} , \dots , ∇X_{t-p+1} . For large n the t -ratio

$$\hat{\tau}_\mu := \hat{\phi}_1^* / \widehat{\text{SE}}(\hat{\phi}_1^*), \quad (6.3.5)$$

where $\widehat{\text{SE}}(\hat{\phi}_1^*)$ is the estimated standard error of $\hat{\phi}_1^*$, has the same limit distribution as the test statistic in (6.3.3). The augmented Dickey–Fuller test in this case is applied in exactly the same manner as for the AR(1) case using the test statistic (6.3.5) and the cutoff values given above.

Example 6.3.1 Consider testing the time series of Example 6.1.1 (see Figure 6-1) for the presence of a unit root in the autoregressive operator. The sample PACF in Figure 6-3 suggests fitting an AR(2) or possibly an AR(3) model to the data. Regressing ∇X_t on $1, X_{t-1}, \nabla X_{t-1}, \nabla X_{t-2}$ for $t = 4, \dots, 200$ using OLS gives

$$\nabla X_t = 0.1503 - 0.0041X_{t-1} + 0.9335\nabla X_{t-1} - 0.1548\nabla X_{t-2} + Z_t,$$

$$(0.1135) \quad (0.0028) \quad (0.0707) \quad (0.0708)$$

where $\{Z_t\} \sim \text{WN}(0, 0.9639)$. The test statistic for testing the presence of a unit root is

$$\hat{\tau}_\mu = \frac{-0.0041}{0.0028} = -1.464.$$

Since $-1.464 > -2.57$, the unit root hypothesis is not rejected at level 0.10. In contrast, if we had mistakenly used the t -distribution with 193 degrees of freedom as an approximation to $\hat{\tau}_\mu$, then we would have rejected the unit root hypothesis at the 0.10 level (p -value is 0.074). The t -ratios for the other coefficients, ϕ_0^* , ϕ_2^* , and ϕ_3^* , have an approximate t -distribution with 193 degrees of freedom. Based on these t -ratios, the intercept should be 0, while the coefficient of ∇X_{t-2} is barely significant. The evidence is much stronger in favor of a unit root if the analysis is repeated without a mean term. The fitted model without a mean term is

$$\nabla X_t = 0.0012X_{t-1} + 0.9395\nabla X_{t-1} - 0.1585\nabla X_{t-2} + Z_t,$$

$$(0.0018) \quad (0.0707) \quad (0.0709)$$

where $\{Z_t\} \sim \text{WN}(0, 0.9677)$. The 0.01, 0.05, and 0.10 cutoff values for the corresponding test statistic when a mean term is excluded from the model are -2.58 , -1.95 , and -1.62 (see Table 8.5.2 of Fuller 1976). In this example, the test statistic is

$$\hat{\tau} = \frac{-0.0012}{0.0018} = -0.667,$$

which is substantially larger than the 0.10 cutoff value of -1.62 . □

Further extensions of the above test to AR models with $p = O(n^{1/3})$ and to ARMA(p, q) models can be found in Said and Dickey (1984). However, as reported in Schwert (1987) and Pantula (1991), this test must be used with caution if the underlying model orders are not correctly specified.

6.3.2 Unit Roots in Moving Averages

A unit root in the moving-average polynomial can have a number of interpretations depending on the modeling application. For example, let $\{X_t\}$ be a causal and invertible ARMA(p, q) process satisfying the equations

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

Then the differenced series $Y_t := \nabla X_t$ is a noninvertible ARMA($p, q + 1$) process with moving-average polynomial $\theta(z)(1 - z)$. Consequently, testing for a unit root in the moving-average polynomial is equivalent to testing that the time series has been overdifferenced.

As a second application, it is possible to distinguish between the competing models

$$\nabla^k X_t = a + V_t$$

and

$$X_t = c_0 + c_1 t + \cdots + c_k t^k + W_t,$$

where $\{V_t\}$ and $\{W_t\}$ are invertible ARMA processes. For the former model the differenced series $\{\nabla^k X_t\}$ has no moving-average unit roots, while for the latter model $\{\nabla^k X_t\}$ has a multiple moving-average unit root of order k . We can therefore distinguish between the two models by using the observed values of $\{\nabla^k X_t\}$ to test for the presence of a moving-average unit root.

We confine our discussion of unit root tests to first-order moving-average models, the general case being considerably more complicated and not fully resolved. Let X_1, \dots, X_n be observations from the MA(1) model

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{IID}(0, \sigma^2).$$

Davis and Dunsmuir (1996) showed that under the assumption $\theta = -1$, $n(\hat{\theta} + 1)$ ($\hat{\theta}$ is the maximum likelihood estimator) converges in distribution. A test of $H_0 : \theta = -1$ vs. $H_1 : \theta > -1$ can be fashioned on this limiting result by rejecting H_0 when

$$\hat{\theta} > -1 + c_\alpha/n,$$

where c_α is the $(1 - \alpha)$ quantile of the limit distribution of $n(\hat{\theta} + 1)$. (From Table 3.2 of Davis et al. (1995), $c_{0.01} = 11.93$, $c_{0.05} = 6.80$, and $c_{0.10} = 4.90$.) In particular, if $n = 50$, then the null hypothesis is rejected at level 0.05 if $\hat{\theta} > -1 + 6.80/50 = -0.864$.

The likelihood ratio test can also be used for testing the unit root hypothesis. The likelihood ratio for this problem is $L(-1, S(-1)/n)/L(\hat{\theta}, \hat{\sigma}^2)$, where $L(\theta, \sigma^2)$ is the Gaussian likelihood of the data based on an MA(1) model, $S(-1)$ is the sum of squares given by (5.2.11) when $\theta = -1$, and $\hat{\theta}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators of θ and σ^2 . The null hypothesis is rejected at level α if

$$\lambda_n := -2 \ln \left(\frac{L(-1, S(-1)/n)}{L(\hat{\theta}, \hat{\sigma}^2)} \right) > c_{\text{LR}, \alpha}$$

where the cutoff value is chosen such that $P_{\theta=-1}[\lambda_n > c_{\text{LR}, \alpha}] = \alpha$. The limit distribution of λ_n was derived by Davis et al. (1995), who also gave selected quantiles of the limit. It was found that these quantiles provide a good approximation to their finite-sample counterparts for time series of length $n \geq 50$. The limiting quantiles for λ_n under H_0 are $c_{\text{LR}, 0.01} = 4.41$, $c_{\text{LR}, 0.05} = 1.94$, and $c_{\text{LR}, 0.10} = 1.00$.

Example 6.3.2 For the overshoot data $\{X_t\}$ of Example 3.2.8, the maximum likelihood MA(1) model for the mean corrected data $\{Y_t = X_t + 4.035\}$ was (see Example 5.4.1)

$$Y_t = Z_t - 0.818Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 2040.75).$$

In the structural formulation of this model given in Example 3.2.8, the moving-average parameter θ was related to the measurement error variances σ_U^2 and σ_V^2 through the equation

$$\frac{\theta}{1 + \theta^2} = \frac{-\sigma_U^2}{2\sigma_U^2 + \sigma_V^2}.$$

(These error variances correspond to the daily measured amounts of fuel in the tank and the daily measured adjustments due to sales and deliveries.) A value of $\theta = -1$ indicates that there is no appreciable measurement error due to sales and deliveries (i.e., $\sigma_V^2 = 0$), and hence testing for a unit root in this case is equivalent to testing that $\sigma_U^2 = 0$. Assuming that the mean is known, the unit root hypothesis is rejected at $\alpha = 0.05$, since $-0.818 > -1 + 6.80/57 = -0.881$. The evidence against H_0 is stronger using the likelihood ratio statistic. Using ITSM and entering the MA(1) model $\theta = -1$ and $\sigma^2 = 2203.12$, we find that $-2 \ln L(-1, 2203.12) = 604.584$, while $-2 \ln L(\hat{\theta}, \hat{\sigma}^2) = 597.267$. Comparing the likelihood ratio statistic $\lambda_n = 604.584 - 597.267 = 7.317$ with the cutoff value $c_{LR,0.01}$, we reject H_0 at level $\alpha = 0.01$ and conclude that the measurement error associated with sales and deliveries is nonzero.

In the above example it was assumed that the mean was known. In practice, these tests should be adjusted for the fact that the mean is also being estimated.

Tanaka (1990) proposed a locally best invariant unbiased (LBIU) test for the unit root hypothesis. It was found that the LBIU test has slightly greater power than the likelihood ratio test for alternatives close to $\theta = -1$ but has less power for alternatives further away from -1 (see Davis et al. 1995). The LBIU test has been extended to cover more general models by Tanaka (1990) and Tam and Reinsel (1995). Similar extensions to tests based on the maximum likelihood estimator and the likelihood ratio statistic have been explored in Davis et al. (1996). □

6.4 Forecasting ARIMA Models

In this section we demonstrate how the methods of Sections 3.3 and 5.4 can be adapted to forecast the future values of an ARIMA(p, d, q) process $\{X_t\}$. (The required numerical calculations can all be carried out using the program ITSM.)

If $d \geq 1$, the first and second moments EX_t and $E(X_{t+h}X_t)$ are not determined by the difference equations (6.1.1). We cannot expect, therefore, to determine best linear predictors for $\{X_t\}$ without further assumptions.

For example, suppose that $\{Y_t\}$ is a causal ARMA(p, q) process and that X_0 is any random variable. Define

$$X_t = X_0 + \sum_{j=1}^t Y_j, \quad t = 1, 2, \dots$$

Then $\{X_t, t \geq 0\}$ is an ARIMA($p, 1, q$) process with mean $EX_t = EX_0$ and autocovariances $E(X_{t+h}X_t) - (EX_0)^2$ that depend on $\text{Var}(X_0)$ and $\text{Cov}(X_0, Y_j), j = 1, 2, \dots$. The best linear predictor of X_{n+1} based on $\{1, X_0, X_1, \dots, X_n\}$ is the same as the best linear predictor in terms of the set $\{1, X_0, Y_1, \dots, Y_n\}$, since each linear combination of the latter is a linear combination of the former and vice versa. Hence, using P_n to denote best linear predictor in terms of either set and using the linearity of P_n , we can write

$$P_n X_{n+1} = P_n(X_0 + Y_1 + \dots + Y_{n+1}) = P_n(X_n + Y_{n+1}) = X_n + P_n Y_{n+1}.$$

To evaluate $P_n Y_{n+1}$ it is necessary (see Section 2.5) to know $E(X_0 Y_j), j = 1, \dots, n+1$, and EX_0^2 . However, if we assume that X_0 is uncorrelated with $\{Y_t, t \geq 1\}$, then $P_n Y_{n+1}$ is the same (Problem 6.5) as the best linear predictor \hat{Y}_{n+1} of Y_{n+1} in terms of $\{1, Y_1, \dots, Y_n\}$, which can be calculated as described in Section 3.3. The assumption

that X_0 is uncorrelated with Y_1, Y_2, \dots is therefore sufficient to determine the best linear predictor $P_n X_{n+1}$ in this case.

Turning now to the general case, we shall assume that our observed process $\{X_t\}$ satisfies the difference equations

$$(1 - B)^d X_t = Y_t, \quad t = 1, 2, \dots,$$

where $\{Y_t\}$ is a causal ARMA(p, q) process, and that the random vector (X_{1-d}, \dots, X_0) is uncorrelated with $Y_t, t > 0$. The difference equations can be rewritten in the form

$$X_t = Y_t - \sum_{j=1}^d \binom{d}{j} (-1)^j X_{t-j}, \quad t = 1, 2, \dots \quad (6.4.1)$$

It is convenient, by relabeling the time axis if necessary, to assume that we observe $X_{1-d}, X_{2-d}, \dots, X_n$. (The observed values of $\{Y_t\}$ are then Y_1, \dots, Y_n .) As usual, we shall use P_n to denote best linear prediction in terms of the observations up to time n (in this case $1, X_{1-d}, \dots, X_n$ or equivalently $1, X_{1-d}, \dots, X_0, Y_1, \dots, Y_n$).

Our goal is to compute the best linear predictors $P_n X_{n+h}$. This can be done by applying the operator P_n to each side of (6.4.1) (with $t = n+h$) and using the linearity of P_n to obtain

$$P_n X_{n+h} = P_n Y_{n+h} - \sum_{j=1}^d \binom{d}{j} (-1)^j P_n X_{n+h-j}. \quad (6.4.2)$$

Now the assumption that (X_{1-d}, \dots, X_0) is uncorrelated with $Y_t, t > 0$, enables us to identify $P_n Y_{n+h}$ with the best linear predictor of Y_{n+h} in terms of $\{1, Y_1, \dots, Y_n\}$, and this can be calculated as described in Section 3.3. The predictor $P_n X_{n+1}$ is obtained directly from (6.4.2) by noting that $P_n X_{n+1-j} = X_{n+1-j}$ for each $j \geq 1$. The predictor $P_n X_{n+2}$ can then be found from (6.4.2) using the previously calculated value of $P_n X_{n+1}$. The predictors $P_n X_{n+3}, P_n X_{n+4}, \dots$ can be computed recursively in the same way.

To find the mean squared error of prediction it is convenient to express $P_n Y_{n+h}$ in terms of $\{X_j\}$. For $n \geq 0$ we denote the one-step predictors by $\hat{Y}_{n+1} = P_n Y_{n+1}$ and $\hat{X}_{n+1} = P_n X_{n+1}$. Then from (6.4.1) and (6.4.2) we have

$$X_{n+1} - \hat{X}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}, \quad n \geq 1,$$

and hence from (3.3.12), if $n > m = \max(p, q)$ and $h \geq 1$, we can write

$$P_n Y_{n+h} = \sum_{i=1}^p \phi_i P_n Y_{n+h-i} + \sum_{j=h}^q \theta_{n+h-1,j} (X_{n+h-j} - \hat{X}_{n+h-j}). \quad (6.4.3)$$

Setting $\phi^*(z) = (1 - z)^d \phi(z) = 1 - \phi_1^* z - \dots - \phi_{p+d}^* z^{p+d}$, we find from (6.4.2) and (6.4.3) that

$$P_n X_{n+h} = \sum_{j=1}^{p+d} \phi_j^* P_n X_{n+h-j} + \sum_{j=h}^q \theta_{n+h-1,j} (X_{n+h-j} - \hat{X}_{n+h-j}), \quad (6.4.4)$$

which is analogous to the h -step prediction formula (3.3.12) for an ARMA process. As in (3.3.13), the mean squared error of the h -step predictor is

$$\sigma_n^2(h) = E(X_{n+h} - P_n X_{n+h})^2 = \sum_{j=0}^{h-1} \left(\sum_{r=0}^j \chi_r \theta_{n+h-r-1, j-r} \right)^2 v_{n+h-j-1}, \quad (6.4.5)$$

where $\theta_{n0} = 1$,

$$\chi(z) = \sum_{r=0}^{\infty} \chi_r z^r = (1 - \phi_1^* z - \dots - \phi_{p+d}^* z^{p+d})^{-1},$$

and

$$v_{n+h-j-1} = E(X_{n+h-j} - \hat{X}_{n+h-j})^2 = E(Y_{n+h-j} - \hat{Y}_{n+h-j})^2.$$

The coefficients χ_j can be found from the recursions (3.3.14) with ϕ_j^* replacing ϕ_j . For large n we can approximate (6.4.5), provided that $\theta(\cdot)$ is invertible, by

$$\sigma_n^2(h) = \sum_{j=0}^{h-1} \psi_j^2 \sigma^2, \quad (6.4.6)$$

where

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = (\phi^*(z))^{-1} \theta(z).$$

6.4.1 The Forecast Function

Inspection of equation (6.4.4) shows that for fixed $n > m = \max(p, q)$, the h -step predictors

$$g(h) := P_n X_{n+h},$$

satisfy the homogeneous linear difference equations

$$g(h) - \phi_1^* g(h-1) - \dots - \phi_{p+d}^* g(h-p-d) = 0, \quad h > q, \quad (6.4.7)$$

where $\phi_1^*, \dots, \phi_{p+d}^*$ are the coefficients of z, \dots, z^{p+d} in

$$\phi^*(z) = (1-z)^d \phi(z).$$

The solution of (6.4.7) is well known from the theory of linear difference equations (see Brockwell and Davis (1991), Section 3.6). If we assume that the zeros of $\phi(z)$ (denoted by ξ_1, \dots, ξ_p) are all distinct, then the solution is

$$g(h) = a_0 + a_1 h + \dots + a_{d-1} h^{d-1} + b_1 \xi_1^{-h} + \dots + b_p \xi_p^{-h}, \quad h > q - p - d, \quad (6.4.8)$$

where the coefficients a_0, \dots, a_{d-1} and b_1, \dots, b_p can be determined from the $p+d$ equations obtained by equating the right-hand side of (6.4.8) for $q-p-d < h \leq q$ with the corresponding value of $g(h)$ computed numerically (for $h \leq 0$, $P_n X_{n+h} = X_{n+h}$, and for $1 \leq h \leq q$, $P_n X_{n+h}$ can be computed from (6.4.4) as already described). Once the constants a_i and b_i have been evaluated, the algebraic expression (6.4.8) gives the predictors for all $h > q - p - d$. In the case $q = 0$, the values of $g(h)$ in the equations for $a_0, \dots, a_{d-1}, b_1, \dots, b_p$ are simply the *observed* values $g(h) = X_{n+h}$, $-p-d \leq h \leq 0$, and the expression (6.4.6) for the mean squared error is exact.

The calculation of the forecast function is easily generalized to deal with more complicated ARIMA processes. For example, if the observations $X_{-13}, X_{-12}, \dots, X_n$ are differenced at lags 12 and 1, and $(1-B)(1-B^{12})X_t$ is modeled as a causal invertible ARMA(p, q) process with mean μ and $\max(p, q) < n$, then $\{X_t\}$ satisfies an equation of the form

$$\phi(B)[(1-B)(1-B^{12})X_t - \mu] = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad (6.4.9)$$

and the forecast function $g(h) = P_n X_{n+h}$ satisfies the analogue of (6.4.7), namely,

$$\phi(B)(1-B)(1-B^{12})g(h) = \phi(1)\mu, \quad h > q. \quad (6.4.10)$$

To find the general solution of these inhomogeneous linear difference equations, it suffices (see Brockwell and Davis (1991), Section 3.6) to find one particular solution of (6.4.10) and then add to it the general solution of the same equations with the right-hand side set equal to zero. A particular solution is easily found (by trial and error) to be

$$g(h) = \frac{\mu h^2}{24},$$

and the general solution is therefore

$$g(h) = \frac{\mu h^2}{24} + a_0 + a_1 h + \sum_{j=1}^{11} c_j e^{ij\pi/6} + b_1 \xi_1^{-h} + \dots + b_p \xi_p^{-h}, \quad (6.4.11)$$

$$h > q - p - 13.$$

(The terms a_0 and $a_1 h$ correspond to the double root $z = 1$ of the equation $\phi(z)(1-z)(1-z^{12}) = 0$, and the subsequent terms to each of the other roots, which we assume to be distinct.) For $q - p - 13 < h \leq 0$, $g(h) = X_{n+h}$, and for $1 \leq h \leq q$, the values of $g(h) = P_n X_{n+h}$ can be determined recursively from the equations

$$P_n X_{n+h} = \mu + P_n X_{n-1} + P_n X_{n-12} - P_n X_{n-13} + P_n Y_{n+h},$$

where $\{Y_t\}$ is the ARMA process $Y_t = (1-B)(1-B^{12})X_t - \mu$. Substituting these values of $g(h)$ into (6.4.11), we obtain a set of $p + 13$ equations for the coefficients a_i, b_j , and c_k . Solving these equations then completes the determination of $g(h)$.

The large-sample approximation to the mean squared error is again given by (6.4.6), with ψ_j redefined as the coefficient of z^j in the power series expansion of $\theta(z)/[(1-z)(1-z^{12})\phi(z)]$.

Example 6.4.1 An ARIMA(1,1,0) Model

In Example 5.2.4 we found the maximum likelihood AR(1) model for the mean-corrected differences X_t of the Dow Jones Utilities Index (August 28–December 18, 1972). The model was

$$X_t - 0.4471X_{t-1} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.1455), \quad (6.4.12)$$

where $X_t = D_t - D_{t-1} - 0.1336$, $t = 1, \dots, 77$, and $\{D_t, t = 0, 1, 2, \dots, 77\}$ is the original series. The model for $\{D_t\}$ is thus

$$(1 - 0.4471B)[(1-B)D_t - 0.1336] = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.1455).$$

The recursions for $g(h)$ therefore take the form

$$(1 - 0.4471B)(1-B)g(h) = 0.5529 \times 0.1336 = 0.07387, \quad h > 0. \quad (6.4.13)$$

A particular solution of these equations is $g(h) = 0.1336h$, so the general solution is

$$g(h) = 0.1336h + a + b(0.4471)^h, \quad h > -2. \quad (6.4.14)$$

Substituting $g(-1) = D_{76} = 122$ and $g(0) = D_{77} = 121.23$ in the equations with $h = -1$ and $h = 0$, and solving for a and b gives

$$g(h) = 0.1366h + 120.50 + 0.7331(0.4471)^h.$$

Setting $h = 1$ and $h = 2$ gives

$$P_{77}D_{78} = 120.97 \quad \text{and} \quad P_{77}D_{79} = 120.94.$$

From (6.4.5) we find that the corresponding mean squared errors are

$$\sigma_{77}^2(1) = v_{77} = \sigma^2 = 0.1455$$

and

$$\sigma_{77}^2(2) = v_{78} + \phi_1^{*2}v_{77} = \sigma^2(1 + 1.4471^2) = 0.4502.$$

(Notice that the approximation (6.4.6) is exact in this case.) The predictors and their mean squared errors are easily obtained from the program ITSM by opening the file DOWJ.TSM, differencing at lag 1, fitting a preliminary AR(1) model to the mean-corrected data with Burg's algorithm, and selecting Model>Estimation>Max likelihood to find the maximum likelihood AR(1) model. Predicted values and their mean squared errors are then found using the option Forecasting>ARMA. \square

6.5 Seasonal ARIMA Models

We have already seen how differencing the series $\{X_t\}$ at lag s is a convenient way of eliminating a seasonal component of period s . If we fit an ARMA(p, q) model $\phi(B)Y_t = \theta(B)Z_t$ to the differenced series $Y_t = (1 - B^s)X_t$, then the model for the original series is $\phi(B)(1 - B^s)X_t = \theta(B)Z_t$. This is a special case of the general seasonal ARIMA (SARIMA) model defined as follows.

Definition 6.5.1

If d and D are nonnegative integers, then $\{X_t\}$ is a **seasonal ARIMA(p, d, q) \times (P, D, Q) $_s$ process with period s** if the differenced series $Y_t = (1 - B)^d(1 - B^s)^D X_t$ is a causal ARMA process defined by

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad (6.5.1)$$

where $\phi(z) = 1 - \phi_1z - \dots - \phi_pz^p$, $\Phi(z) = 1 - \Phi_1z - \dots - \Phi_Pz^P$, $\theta(z) = 1 + \theta_1z + \dots + \theta_qz^q$, and $\Theta(z) = 1 + \Theta_1z + \dots + \Theta_Qz^Q$.

Remark 1. Note that the process $\{Y_t\}$ is causal if and only if $\phi(z) \neq 0$ and $\Phi(z) \neq 0$ for $|z| \leq 1$. In applications D is rarely more than one, and P and Q are typically less than three. \square

Remark 2. Equation (6.5.1) satisfied by the differenced process $\{Y_t\}$ can be rewritten in the equivalent form

$$\phi^*(B)Y_t = \theta^*(B)Z_t, \quad (6.5.2)$$

where $\phi^*(\cdot)$, $\theta^*(\cdot)$ are polynomials of degree $p + sP$ and $q + sQ$, respectively, whose coefficients can all be expressed in terms of ϕ_1, \dots, ϕ_p , Φ_1, \dots, Φ_P , $\theta_1, \dots, \theta_q$, and

$\Theta_1, \dots, \Theta_Q$. Provided that $p < s$ and $q < s$, the constraints on the coefficients of $\phi^*(\cdot)$ and $\theta^*(\cdot)$ can all be expressed as multiplicative relations

$$\phi_{is+j}^* = \phi_{is}^* \phi_j^*, \quad i = 1, 2, \dots; \quad j = 1, \dots, s-1,$$

and

$$\theta_{is+j}^* = \theta_{is}^* \theta_j^*, \quad i = 1, 2, \dots; \quad j = 1, \dots, s-1.$$

In Section 1.5 we discussed the classical decomposition model incorporating trend, seasonality, and random noise, namely, $X_t = m_t + s_t + Y_t$. In modeling real data it might not be reasonable to assume, as in the classical decomposition model, that the seasonal component s_t repeats itself precisely in the same way cycle after cycle. Seasonal ARIMA models allow for randomness in the seasonal pattern from one cycle to the next. \square

Example 6.5.1 Suppose we have r years of monthly data, which we tabulate as follows:

Year/Month	1	2	...	12
1	Y_1	Y_2	...	Y_{12}
2	Y_{13}	Y_{14}	...	Y_{24}
3	Y_{25}	Y_{26}	...	Y_{36}
\vdots	\vdots	\vdots	\vdots	
r	$Y_{1+12(r-1)}$	$Y_{2+12(r-1)}$...	$Y_{12+12(r-1)}$

Each column in this table may itself be viewed as a realization of a time series. Suppose that each one of these twelve time series is generated by the same ARMA(P, Q) model, or more specifically, that the series corresponding to the j th month, Y_{j+12t} , $t = 0, \dots, r-1$, satisfies a difference equation of the form

$$Y_{j+12t} = \Phi_1 Y_{j+12(t-1)} + \dots + \Phi_P Y_{j+12(t-P)} + U_{j+12t} + \Theta_1 U_{j+12(t-1)} + \dots + \Theta_Q U_{j+12(t-Q)}, \quad (6.5.3)$$

where

$$\{U_{j+12t}, t = \dots, -1, 0, 1, \dots\} \sim \text{WN}(0, \sigma_U^2). \quad (6.5.4)$$

Then since the same ARMA(P, Q) model is assumed to apply to each month, (6.5.3) holds for each $j = 1, \dots, 12$. (Notice, however, that $E(U_t U_{t+h})$ is not necessarily zero except when h is an integer multiple of 12.) We can thus write (6.5.3) in the compact form

$$\Phi(B^{12}) Y_t = \Theta(B^{12}) U_t, \quad (6.5.5)$$

where $\Phi(z) = 1 - \Phi_1 z - \dots - \Phi_P z^P$, $\Theta(z) = 1 + \Theta_1 z + \dots + \Theta_Q z^Q$, and $\{U_{j+12t}, t = \dots, -1, 0, 1, \dots\} \sim \text{WN}(0, \sigma_U^2)$ for each j . We refer to the model (6.5.5) as the between-year model. \square

Example 6.5.2 Suppose $P = 0$, $Q = 1$, and $\Theta_1 = -0.4$ in (6.5.5). Then the series for any particular month is a moving-average of order 1. If $E(U_t U_{t+h}) = 0$ for all h , i.e., if the white noise sequences for different months are uncorrelated with each other, then the columns themselves are uncorrelated. The correlation function for such a process is shown in Figure 6-15. \square

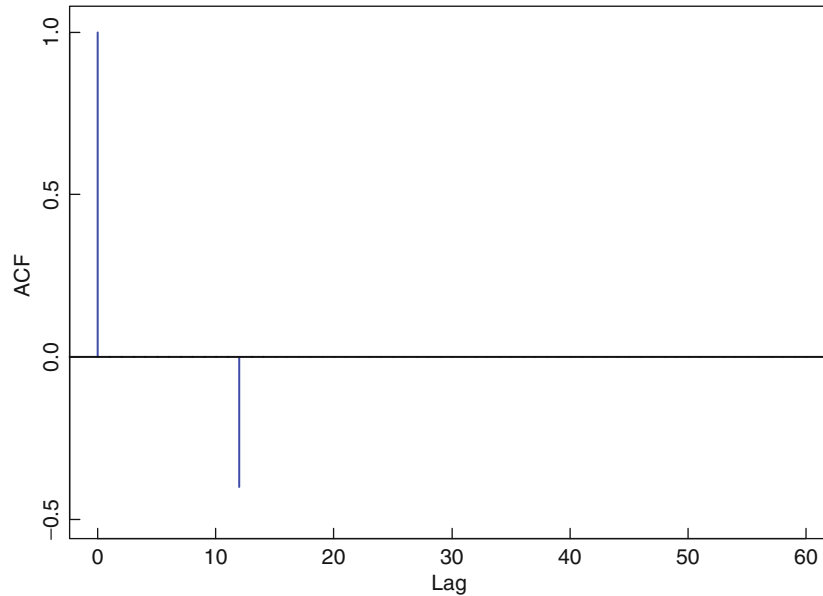


Figure 6-15
The ACF of the model
 $X_t = U_t - 0.4U_{t-12}$
of Example 6.5.2

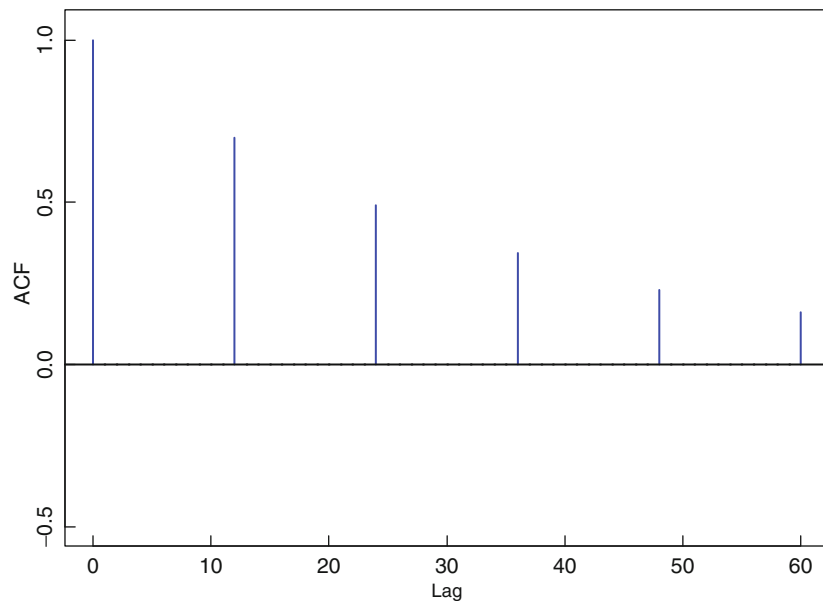


Figure 6-16
The ACF of the model
 $X_t - 0.7X_{t-12} = U_t$
of Example 6.5.3

Example 6.5.3 Suppose $P = 1$, $Q = 0$, and $\Phi_1 = 0.7$ in (6.5.5). In this case the 12 series (one for each month) are AR(1) processes that are uncorrelated if the white noise sequences for different months are uncorrelated. A graph of the autocorrelation function of this process is shown in Figure 6-16. □

In each of the Examples 6.5.1–6.5.3, the 12 series corresponding to the different months are uncorrelated. To incorporate dependence between these series we allow the process $\{U_t\}$ in (6.5.5) to follow an ARMA(p, q) model,

$$\phi(B)U_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2). \quad (6.5.6)$$

This assumption implies possible nonzero correlation not only between consecutive values of U_t , but also within the 12 sequences $\{U_{j+12t}, t = \dots, -1, 0, 1, \dots\}$, each of which was assumed to be uncorrelated in the preceding examples. In this case (6.5.4)

may no longer hold; however, the coefficients in (6.5.6) will frequently have values such that $E(U_t U_{t+12j})$ is small for $j = \pm 1, \pm 2, \dots$. Combining the two models (6.5.5) and (6.5.6) and allowing for possible differencing leads directly to Definition 6.5.1 of the general SARIMA model as given above.

The first steps in identifying SARIMA models for a (possibly transformed) data set are to find d and D so as to make the differenced observations

$$Y_t = (1 - B)^d (1 - B^s)^D X_t$$

stationary in appearance (see Sections 6.1–6.3). Next we examine the sample ACF and PACF of $\{Y_t\}$ at lags that are multiples of s for an indication of the orders P and Q in the model (6.5.5). If $\hat{\rho}(\cdot)$ is the sample ACF of $\{Y_t\}$, then P and Q should be chosen such that $\hat{\rho}(ks)$, $k = 1, 2, \dots$, is compatible with the ACF of an ARMA(P , Q) process. The orders p and q are then selected by trying to match $\hat{\rho}(1), \dots, \hat{\rho}(s-1)$ with the ACF of an ARMA(p , q) process. Ultimately, the AICC criterion (Section 5.5) and the goodness of fit tests (Section 5.3) are used to select the best SARIMA model from competing alternatives.

For given values of p , d , q , P , D , and Q , the parameters ϕ , θ , Φ , Θ , and σ^2 can be found using the maximum likelihood procedure of Section 5.2. The differences $Y_t = (1 - B)^d (1 - B^s)^D X_t$ constitute an ARMA($p + sP$, $q + sQ$) process in which some of the coefficients are zero and the rest are functions of the $(p + P + q + Q)$ -dimensional vector $\beta' = (\phi', \Phi', \theta', \Theta')$. For any fixed β the reduced likelihood $\ell(\beta)$ of the differences Y_{t+d+sD}, \dots, Y_n is easily computed as described in Section 5.2. The maximum likelihood estimator of β is the value that minimizes $\ell(\beta)$, and the maximum likelihood estimate of σ^2 is given by (5.2.10). The estimates can be found using the program ITSM by specifying the required multiplicative relationships among the coefficients as given in Remark 2 above.

A more direct approach to modeling the differenced series $\{Y_t\}$ is simply to fit a subset ARMA model of the form (6.5.2) without making use of the multiplicative form of $\phi^*(\cdot)$ and $\theta^*(\cdot)$ in (6.5.1).

Example 6.5.4 Monthly Accidental Deaths

In Figure 1-27 we showed the series $\{Y_t = (1 - B^{12})(1 - B)X_t\}$ obtained by differencing the accidental deaths series $\{X_t\}$ once at lag 12 and once at lag 1. The sample ACF of $\{Y_t\}$ is shown in Figure 6-17.

□

The values $\hat{\rho}(12) = -0.333$, $\hat{\rho}(24) = -0.099$, and $\hat{\rho}(36) = 0.013$ suggest a moving-average of order 1 for the between-year model (i.e., $P = 0$ and $Q = 1$). Moreover, inspection of $\hat{\rho}(1), \dots, \hat{\rho}(11)$ suggests that $\rho(1)$ is the only short-term correlation different from zero, so we also choose a moving-average of order 1 for the between-month model (i.e., $p = 0$ and $q = 1$). Taking into account the sample mean (28.831) of the differences $\{Y_t\}$, we therefore arrive at the model

$$Y_t = 28.831 + (1 + \theta_1 B)(1 + \Theta_1 B^{12})Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2), \quad (6.5.7)$$

for the series $\{Y_t\}$. The maximum likelihood estimates of the parameters are obtained from ITSM by opening the file DEATHS.TSM and proceeding as follows. After differencing (at lags 1 and 12) and then mean-correcting the data, choose the option `Model>Specify`. In the dialog box enter an MA(13) model with $\theta_1 = -0.3$, $\theta_{12} = -0.3$, $\theta_{13} = 0.09$, and all other coefficients zero. (This corresponds to the initial

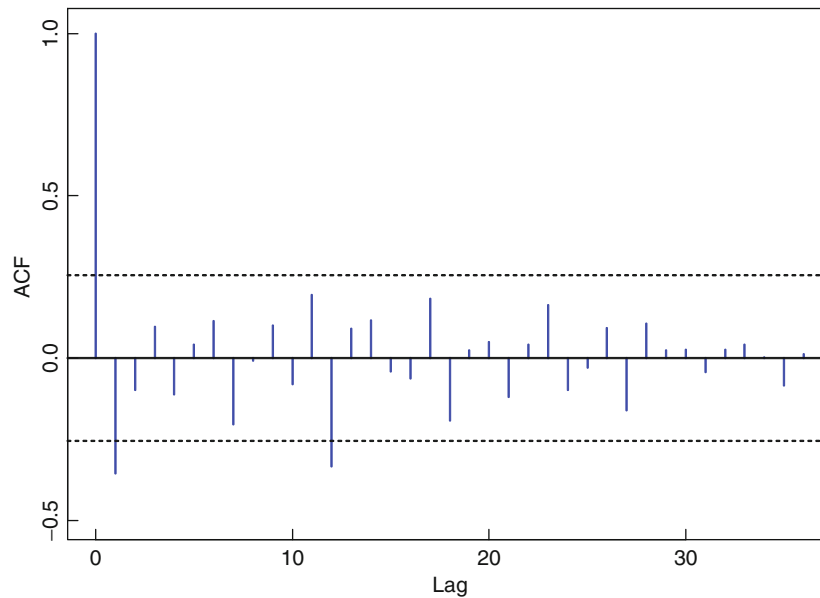


Figure 6-17
The sample ACF of the differenced accidental deaths $\{\nabla\nabla_{12}X_t\}$

guess $Y_t = (1 - 0.3B)(1 - 0.3B^{12})Z_t$.) Then choose Model>Estimation>Max likelihood and click on the button Constrain optimization. Specify the number of multiplicative relations (one in this case) in the box provided and define the relationship by entering 1, 12, 13 to indicate that $\theta_1 \times \theta_{12} = \theta_{13}$. Click OK to return to the Maximum Likelihood dialog box. Click OK again to obtain the parameter estimates

$$\hat{\theta}_1 = -0.478,$$

$$\Theta_1 = -0.591,$$

and

$$\hat{\sigma}^2 = 94,255,$$

with AICC value 855.53. The corresponding fitted model for $\{X_t\}$ is thus the SARIMA $(0, 1, 1) \times (0, 1, 1)_{12}$ process

$$\nabla\nabla_{12}X_t = 28.831 + (1 - 0.478B)(1 - 0.591B^{12})Z_t, \quad (6.5.8)$$

where $\{Z_t\} \sim \text{WN}(0, 94390)$.

If we adopt the alternative approach of fitting a subset ARMA model to $\{Y_t\}$ without seeking a multiplicative structure for the operators $\phi^*(B)$ and $\theta^*(B)$ in (6.5.2), we begin by fitting a preliminary MA(13) model (as suggested by Figure 6-17) to the series $\{Y_t\}$. We then fit a maximum likelihood MA(13) model and examine the standard errors of the coefficient estimators. This suggests setting the coefficients at lags 2, 3, 8, 10, and 11 equal to zero, since these are all less than one standard error from zero. To do this select Model>Estimation>Max likelihood and click on the button Constrain optimization. Then highlight the coefficients to be set to zero and click on the button Set to zero. Click OK to return to the Maximum Likelihood Estimation dialog box and again to carry out the constrained optimization. The coefficients that have been set to zero will be held at that value, and the optimization will be with respect to the remaining coefficients. This gives a model with substantially smaller AICC than the unconstrained MA(13) model. Examining the standard errors again we see that the coefficients at lags 4, 5, and 7 are promising

candidates to be set to zero, since each of them is less than one standard error from zero. Setting these coefficients to zero in the same way and reoptimizing gives a further reduction in AICC. Setting the coefficient at lag 9 to zero and reoptimizing again gives a further reduction in AICC (to 855.61) and the fitted model

$$\begin{aligned} \nabla \nabla_{12} X_t &= 28.831 + Z_t - 0.596Z_{t-1} - 0.407Z_{t-6} - 0.685Z_{t-12} + 0.460Z_{t-13}, \\ \{Z_t\} &\sim \text{WN}(0, 71240). \end{aligned} \quad (6.5.9)$$

The AICC value 855.61 is quite close to the value 855.53 for the model (6.5.8). The residuals from the two models are also very similar, the randomness tests (with the exception of the difference-sign test) yielding high p -values for both.

6.5.1 Forecasting SARIMA Processes

Forecasting SARIMA processes is completely analogous to the forecasting of ARIMA processes discussed in Section 6.4. Expanding out the operator $(1 - B)^d (1 - B^s)^D$ in powers of B , rearranging the equation

$$(1 - B)^d (1 - B^s)^D X_t = Y_t,$$

and setting $t = n + h$ gives the analogue

$$X_{n+h} = Y_{n+h} + \sum_{j=1}^{d+Ds} a_j X_{n+h-j} \quad (6.5.10)$$

of equation (6.4.2). Under the assumption that the first $d + Ds$ observations $X_{-d-Ds+1}, \dots, X_0$ are uncorrelated with $\{Y_t, t \geq 1\}$, we can determine the best linear predictors $P_n X_{n+h}$ of X_{n+h} based on $\{1, X_{-d-Ds+1}, \dots, X_n\}$ by applying P_n to each side of (6.5.10) to obtain

$$P_n X_{n+h} = P_n Y_{n+h} + \sum_{j=1}^{d+Ds} a_j P_n X_{n+h-j}. \quad (6.5.11)$$

The first term on the right is just the best linear predictor of the (possibly nonzero-mean) ARMA process $\{Y_t\}$ in terms of $\{1, Y_1, \dots, Y_n\}$, which can be calculated as described in Section 3.3. The predictors $P_n X_{n+h}$ can then be computed recursively for $h = 1, 2, \dots$ from (6.5.11), if we note that $P_n X_{n+1-j} = X_{n+1-j}$ for each $j \geq 1$.

An argument analogous to the one leading to (6.4.5) gives the prediction mean squared error as

$$\sigma_n^2(h) = E(X_{n+h} - P_n X_{n+h})^2 = \sum_{j=0}^{h-1} \left(\sum_{r=0}^j \chi_r \theta_{n+h-r-1, j-r} \right)^2 v_{n+h-j-1}, \quad (6.5.12)$$

where θ_{nj} and v_n are obtained by applying the innovations algorithm to the differenced series $\{Y_t\}$ and

$$\chi(z) = \sum_{r=0}^{\infty} \chi_r z^r = \left[\phi(z) \Phi(z^s) (1 - z)^d (1 - z^s)^D \right]^{-1}, \quad |z| < 1.$$

Table 6.1 Predicted values of the Accidental Deaths series for $t = 73, \dots, 78$, the standard deviations σ_t of the prediction errors, and the corresponding observed values of X_t for the same period

	t	73	74	75	76	77	78
Model (6.5.8)							
Predictors		8441	7704	8549	8885	9843	10279
σ_t		308	348	383	415	445	474
Model (6.5.9)							
Predictors		8345	7619	8356	8742	9795	10179
σ_t		292	329	366	403	442	486
Observed values							
X_t		7798	7406	8363	8460	9217	9316

For large n we can approximate (6.5.12), if $\theta(z)\Theta(z^s)$ is nonzero for all $|z| \leq 1$, by

$$\sigma_n^2(h) = \sum_{j=0}^{h-1} \psi_j^2 \sigma^2, \quad (6.5.13)$$

where

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)\Theta(z^s)}{\phi(z)\Phi(z^s)(1-z)^d(1-z^s)^D}, \quad |z| < 1.$$

The required calculations can all be carried out with the aid of the program ITSM. The mean squared errors are computed from the large-sample approximation (6.5.13) if the fitted model is invertible. If the fitted model is not invertible, ITSM computes the mean squared errors by converting the model to the equivalent (in terms of Gaussian likelihood) invertible model and then using (6.5.13).

Example 6.5.5 Monthly Accidental Deaths

Continuing with Example 6.5.4, we next use ITSM to predict six future values of the Accidental Deaths series using the fitted models (6.5.8) and (6.5.9). First fit the desired model as described in Example 6.5.4 or enter the data and model directly by opening the file DEATHS.TSM, differencing at lags 12 and 1, subtracting the mean, and then entering the MA(13) coefficients and white noise variance using the option `Model>Specify`. Select `Forecasting>ARMA`, and you will see the ARMA Forecast dialog box. Enter 6 for the number of predicted values required. You will notice that the default options in the dialog box are set to generate predictors of the *original* series by reversing the transformations applied to the data. If for some reason you wish to predict the transformed data, these check marks can be removed. If you wish to include prediction bounds in the graph of the predictors, check the appropriate box and specify the desired coefficient (e.g., 95%). Click OK, and you will see a graph of the data with the six predicted values appended. For numerical values of the predictors and prediction bounds, right-click on the graph and then on `Info`. The prediction bounds are computed under the assumption that the white noise sequence in the ARMA model for the transformed data is Gaussian. Table 6.1 shows the predictors and standard deviations of the prediction errors under both models (6.5.8) and (6.5.9) for the Accidental Deaths series.

□

6.6 Regression with ARMA Errors

6.6.1 OLS and GLS Estimation

In standard linear regression, the errors (or deviations of the observations from the regression function) are assumed to be independent and identically distributed. In many applications of regression analysis, however, this assumption is clearly violated, as can be seen by examination of the residuals from the fitted regression and their sample autocorrelations. It is often more appropriate to assume that the errors are observations of a zero-mean second-order stationary process. Since many autocorrelation functions can be well approximated by the autocorrelation function of a suitably chosen ARMA(p, q) process, it is of particular interest to consider the model

$$Y_t = \mathbf{x}'_t \boldsymbol{\beta} + W_t, \quad t = 1, \dots, n, \quad (6.6.1)$$

or in matrix notation,

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{W}, \quad (6.6.2)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)'$ is the vector of observations at times $t = 1, \dots, n$, X is the design matrix whose t th row, $\mathbf{x}'_t = (x_{t1}, \dots, x_{tk})$, consists of the values of the explanatory variables at time t , $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ is the vector of regression coefficients, and the components of $\mathbf{W} = (W_1, \dots, W_n)'$ are values of a causal zero-mean ARMA(p, q) process satisfying

$$\phi(B)W_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2). \quad (6.6.3)$$

The model (6.6.1) arises naturally in trend estimation for time series data. For example, the explanatory variables $x_{t1} = 1$, $x_{t2} = t$, and $x_{t3} = t^2$ can be used to estimate a quadratic trend, and the variables $x_{t1} = 1$, $x_{t2} = \cos(\omega t)$, and $x_{t3} = \sin(\omega t)$ can be used to estimate a sinusoidal trend with frequency ω . The columns of X are not necessarily simple functions of t as in these two examples. Any specified column of relevant variables, e.g., temperatures at times $t = 1, \dots, n$, can be included in the design matrix X , in which case the regression is conditional on the observed values of the variables included in the matrix.

The **ordinary least squares** (OLS) estimator of $\boldsymbol{\beta}$ is the value, $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, which minimizes the sum of squares

$$(\mathbf{Y} - X\boldsymbol{\beta})'(\mathbf{Y} - X\boldsymbol{\beta}) = \sum_{t=1}^n (Y_t - \mathbf{x}'_t \boldsymbol{\beta})^2.$$

Equating to zero the partial derivatives with respect to each component of $\boldsymbol{\beta}$ and assuming (as we shall) that $X'X$ is nonsingular, we find that

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (X'X)^{-1} X' \mathbf{Y}. \quad (6.6.4)$$

(If $X'X$ is singular, $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is not uniquely determined but still satisfies (6.6.4) with $(X'X)^{-1}$ any generalized inverse of $X'X$.) The OLS estimate also maximizes the likelihood of the observations when the errors W_1, \dots, W_n are iid and Gaussian. If the design matrix X is nonrandom, then even when the errors are non-Gaussian and dependent, the OLS estimator is unbiased (i.e., $E(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \boldsymbol{\beta}$) and its covariance matrix is

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = (X'X)^{-1} X' \Gamma_n X (X'X)^{-1}, \quad (6.6.5)$$

where $\Gamma_n = E(\mathbf{W}\mathbf{W}')$ is the covariance matrix of \mathbf{W} .

The **generalized least squares** (GLS) estimator of β is the value $\hat{\beta}_{GLS}$ that minimizes the *weighted* sum of squares

$$(\mathbf{Y} - X\beta)' \Gamma_n^{-1} (\mathbf{Y} - X\beta). \quad (6.6.6)$$

Differentiating partially with respect to each component of β and setting the derivatives equal to zero, we find that

$$\hat{\beta}_{GLS} = (X' \Gamma_n^{-1} X)^{-1} X' \Gamma_n^{-1} \mathbf{Y}. \quad (6.6.7)$$

If the design matrix X is nonrandom, the GLS estimator is unbiased and has covariance matrix

$$\text{Cov}(\hat{\beta}_{GLS}) = (X' \Gamma_n^{-1} X)^{-1}. \quad (6.6.8)$$

It can be shown that the GLS estimator is the best linear unbiased estimator of β , i.e., for any k -dimensional vector \mathbf{c} and for any unbiased estimator $\hat{\beta}$ of β that is a linear function of the observations Y_1, \dots, Y_n ,

$$\text{Var}(\mathbf{c}' \hat{\beta}_{GLS}) \leq \text{Var}(\mathbf{c}' \hat{\beta}).$$

In this sense the GLS estimator is therefore superior to the OLS estimator. However, it can be computed only if ϕ and θ are known.

Let $V(\phi, \theta)$ denote the matrix $\sigma^{-2} \Gamma_n$ and let $T(\phi, \theta)$ be any square root of V^{-1} (i.e., a matrix such that $T'T = V^{-1}$). Then we can multiply each side of (6.6.2) by T to obtain

$$T\mathbf{Y} = TX\beta + T\mathbf{W}, \quad (6.6.9)$$

a regression equation with coefficient vector β , data vector $T\mathbf{Y}$, design matrix TX , and error vector $T\mathbf{W}$. Since the latter has uncorrelated, zero-mean components, each with variance σ^2 , the best linear estimator of β in terms of $T\mathbf{Y}$ (which is clearly the same as the best linear estimator of β in terms of \mathbf{Y} , i.e., $\hat{\beta}_{GLS}$) can be obtained by applying OLS estimation to the transformed regression equation (6.6.9). This gives

$$\hat{\beta}_{GLS} = (X'T'TX)^{-1} X'T'T\mathbf{Y}, \quad (6.6.10)$$

which is clearly the same as (6.6.7). Cochran and Orcutt (1949) pointed out that if $\{W_t\}$ is an AR(p) process satisfying

$$\phi(B)W_t = Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2),$$

then application of $\phi(B)$ to each side of the regression equations (6.6.1) transforms them into regression equations with uncorrelated, zero-mean, constant-variance errors, so that ordinary least squares can again be used to compute best linear unbiased estimates of the components of β in terms of $Y_t^* = \phi(B)Y_t$, $t = p + 1, \dots, n$. This approach eliminates the need to compute the matrix T but suffers from the drawback that \mathbf{Y}^* does not contain all the information in \mathbf{Y} . Cochran and Orcutt's transformation can be improved, and at the same generalized to ARMA errors, as follows.

Instead of applying the operator $\phi(B)$ to each side of the regression equations (6.6.1), we multiply each side of equation (6.6.2) by the matrix $T(\phi, \theta)$ that maps $\{W_t\}$ into the residuals [see (5.3.1)] of $\{W_t\}$ from the ARMA model (6.6.3). We have already seen how to calculate these residuals using the innovations algorithm in Section 3.3. To see that T is a square root of the matrix V as defined in the previous paragraph, we simply recall that the residuals are uncorrelated with zero mean and variance σ^2 , so that

$$\text{Cov}(T\mathbf{W}) = T\Gamma_n T' = \sigma^2 I,$$

where I is the $n \times n$ identity matrix. Hence

$$T'T = \sigma^2 \Gamma_n^{-1} = V^{-1}.$$

GLS estimation of β can therefore be carried out by multiplying each side of (6.6.2) by T and applying ordinary least squares to the transformed regression model. It remains only to compute $T\mathbf{Y}$ and TX .

Any data vector $\mathbf{d} = (d_1, \dots, d_n)'$ can be left-multiplied by T simply by reading it into ITSM, entering the model (6.6.3), and pressing the green button labeled RES, which plots the residuals. (The calculations are performed using the innovations algorithm as described in Section 3.3.) The GLS estimator $\hat{\beta}_{\text{GLS}}$ is computed as follows. The data vector \mathbf{Y} is left-multiplied by T to generate the transformed data vector \mathbf{Y}^* , and each column of the design matrix X is left-multiplied by T to generate the corresponding column of the transformed design matrix X^* . Then

$$\hat{\beta}_{\text{GLS}} = (X'^* X^*)^{-1} X'^* \mathbf{Y}^*. \quad (6.6.11)$$

The calculations of \mathbf{Y}^* , X^* , and hence of $\hat{\beta}_{\text{GLS}}$, are all carried out by the program ITSM in the option Regression>Estimation>Generalized LS.

6.6.2 ML Estimation

If (as is usually the case) the parameters of the ARMA(p, q) model for the errors are unknown, they can be estimated together with the regression coefficients by **maximizing the Gaussian likelihood**

$$L(\beta, \phi, \theta, \sigma^2) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\beta)' \Gamma_n^{-1} (\mathbf{Y} - X\beta) \right\},$$

where $\Gamma_n(\phi, \theta, \sigma^2)$ is the covariance matrix of $\mathbf{W} = \mathbf{Y} - X\beta$. Since $\{W_t\}$ is an ARMA(p, q) process with parameters (ϕ, θ, σ^2) , the maximum likelihood estimators $\hat{\beta}$, $\hat{\phi}$, and $\hat{\theta}$ are found (as in Section 5.2) by minimizing

$$\ell(\beta, \phi, \theta) = \ln(n^{-1} S(\beta, \phi, \theta)) + n^{-1} \sum_{t=1}^n \ln r_{t-1}, \quad (6.6.12)$$

where

$$S(\beta, \phi, \theta) = \sum_{t=1}^n (W_t - \hat{W}_t)^2 / r_{t-1},$$

\hat{W}_t is the best one-step predictor of W_t , and $r_{t-1}\sigma^2$ is its mean squared error. The function $\ell(\beta, \phi, \theta)$ can be expressed in terms of the observations $\{Y_t\}$ and the parameters β , ϕ , and θ using the innovations algorithm (see Section 3.3) and minimized numerically to give the maximum likelihood estimators, $\hat{\beta}$, $\hat{\phi}$, and $\hat{\theta}$. The maximum likelihood estimator of σ^2 is then given, as in Section 5.2, by $\hat{\sigma}^2 = S(\hat{\beta}, \hat{\phi}, \hat{\theta})/n$.

An extension of an iterative scheme, proposed by Cochran and Orcutt (1949) for the case $q = 0$, simplifies the minimization considerably. It is based on the observation that for fixed ϕ and θ , the value of β that minimizes $\ell(\beta, \phi, \theta)$ is $\hat{\beta}_{\text{GLS}}(\phi, \theta)$, which can be computed algebraically from (6.6.11) instead of by searching numerically for the minimizing value. The scheme is as follows.

- (i) Compute $\hat{\beta}_{\text{OLS}}$ and the estimated residuals $Y_t - \mathbf{x}'_t \hat{\beta}_{\text{OLS}}$, $t = 1, \dots, n$.

- (ii) Fit an ARMA(p, q) model by maximum Gaussian likelihood to the estimated residuals.
- (iii) For the fitted ARMA model compute the corresponding estimator $\hat{\beta}_{\text{GLS}}$ from (6.6.11).
- (iv) Compute the residuals $Y_t - \mathbf{x}'_t \hat{\beta}_{\text{GLS}}$, $t = 1, \dots, n$, and return to (ii), stopping when the estimators have stabilized.

If $\{W_t\}$ is a causal and invertible ARMA process, then under mild conditions on the explanatory variables \mathbf{x}_t , the maximum likelihood estimates are asymptotically multivariate normal (see Fuller 1976). In addition, the estimated regression coefficients are asymptotically independent of the estimated ARMA parameters.

The large-sample covariance matrix of the ARMA parameter estimators, suitably normalized, has a complicated form that involves both the regression variables \mathbf{x}_t and the covariance function of $\{W_t\}$. It is therefore convenient to estimate the covariance matrix as $-H^{-1}$, where H is the Hessian matrix of the observed log-likelihood evaluated at its maximum.

The OLS, GLS, and maximum likelihood estimators of the regression coefficients all have the same asymptotic covariance matrix, so in this sense the dependence does not play a major role. However, the asymptotic covariance of both the OLS and GLS estimators can be very inaccurate if the appropriate covariance matrix Γ_n is not used in the expressions (6.6.5) and (6.6.8). This point is illustrated in the following examples.

Remark 1. The use of the innovations algorithm for GLS and ML estimation extends to regression with ARIMA errors (see Example 6.6.3 below) and FARIMA errors (FARIMA processes are defined in Section 10.5). \square

Example 6.6.1 The Overshort Data

The analysis of the overshoot data in Example 3.2.8 suggested the model

$$Y_t = \beta + W_t,$$

where $-\beta$ is interpreted as the daily leakage from the underground storage tank and $\{W_t\}$ is the MA(1) process

$$W_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

(Here $k = 1$ and $x_{t1} = 1$.) The OLS estimate of β is simply the sample mean $\hat{\beta}_{\text{OLS}} = \bar{Y}_n = -4.035$. Under the assumption that $\{W_t\}$ is iid noise, the estimated variance of the OLS estimator of β is $\hat{\gamma}_y(0)/57 = 59.92$. However, since this estimate of the variance fails to take dependence into account, it is not reliable.

To find maximum Gaussian likelihood estimates of β and the parameters of $\{W_t\}$ using ITSM, open the file OSHORTS.TSM, select the option Regression>Specify and check the box marked Include intercept term only. Then press the blue GLS button and you will see the estimated value of β . (This is in fact the same as the OLS estimator since the default model in ITSM is WN(0,1).) Then select Model>Estimation>Autofit and press Start. The autofit option selects the minimum AICC model for the residuals,

$$W_t = Z_t - 0.818Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 2041), \quad (6.6.13)$$

and displays the estimated MA coefficient $\hat{\theta}_1^{(0)} = -0.818$ and the corresponding GLS estimate $\hat{\beta}_{\text{GLS}}^{(1)} = -4.745$, with a standard error of 1.188, in the Regression estimates window. (If we reestimate the variance of the OLS estimator, using

(6.6.5) with Γ_{57} computed from the model (6.6.13), we obtain the value 2.214, a drastic reduction from the value 59.92 obtained when dependence is ignored. For a positively correlated time series, ignoring the dependence would lead to underestimation of the variance.)

Pressing the blue MLE button will reestimate the MA parameters using the residuals from the updated regression and at the same time reestimate the regression coefficient, printing the new parameters in the Regression estimates window. After this operation has been repeated several times, the parameters will stabilize, as shown in Table 6.2. Estimated 95 % confidence bounds for β using the GLS estimate are $-4.75 \pm 1.96(1.408)^{1/2} = (-7.07, -2.43)$, strongly suggesting that the storage tank has a leak. Such a conclusion would not have been reached without taking into account the dependence in the data. □

Table 6.2 Estimates of β and θ_1 for the overshoot data of Example 6.6.1

Iteration i	$\hat{\theta}^{(i)}$	$\hat{\beta}_1^{(i)}$
0	0	-4.035
1	-0.818	-4.745
2	-0.848	-4.780
3	-0.848	-4.780

Example 6.6.2 The Lake Data

In Examples 5.2.4 and 5.5.2 we found maximum likelihood ARMA(1,1) and AR(2) models for the mean-corrected lake data. Now let us consider fitting a linear trend to the data with AR(2) noise. The choice of an AR(2) model was suggested by an analysis of the residuals obtained after removing a linear trend from the data using OLS. Our model now takes the form

$$Y_t = \beta_0 + \beta_1 t + W_t,$$

where $\{W_t\}$ is the AR(2) process satisfying

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

From Example 1.3.5, we find that the OLS estimate of β is $\hat{\beta}_{\text{OLS}} = (10.202, -0.0242)'$. If we ignore the correlation structure of the noise, the estimated covariance matrix Γ_n of \mathbf{W} is $\hat{\gamma}(0)I$ (where I is the identity matrix). The corresponding estimated covariance matrix of $\hat{\beta}_{\text{OLS}}$ is (from (6.6.5))

$$\hat{\gamma}_Y(0) (X'X)^{-1} = \hat{\gamma}_Y(0) \begin{bmatrix} n & \sum_{t=1}^n t \\ \sum_{t=1}^n t & \sum_{t=1}^n t^2 \end{bmatrix}^{-1} = \begin{bmatrix} 0.07203 & -0.00110 \\ -0.00110 & 0.00002 \end{bmatrix}. \quad (6.6.14)$$

However, the estimated model for the noise process, found by fitting an AR(2) model to the residuals $Y_t - \hat{\beta}'_{\text{OLS}} \mathbf{x}_t$, is

$$W_t = 1.008W_{t-1} - 0.295W_{t-2} + Z_t, \quad \{Z_t\} \sim \text{WN}(0, 0.4571).$$

Table 6.3 Estimates of β and ϕ for the lake data after 3 iterations

Iteration i	$\hat{\phi}_1^{(i)}$	$\hat{\phi}_2^{(i)}$	$\hat{\beta}_1^{(i)}$	$\hat{\beta}_2^{(i)}$
0	0	0	10.20	-0.0242
1	1.008	-0.295	10.09	-0.0216
2	1.005	-0.291	10.09	-0.0216

Assuming that this is the true model for $\{W_t\}$, the GLS estimate is found to be $(10.091, -0.0216)'$, in close agreement with the OLS estimate. The estimated covariance matrices for the OLS and GLS estimates are given by

$$\text{Cov}(\hat{\beta}_{\text{OLS}}) = \begin{bmatrix} 0.22177 & -0.00335 \\ -0.00335 & 0.00007 \end{bmatrix}$$

and

$$\text{Cov}(\hat{\beta}_{\text{GLS}}) = \begin{bmatrix} 0.21392 & -0.00321 \\ -0.00321 & 0.00006 \end{bmatrix}.$$

Notice how the estimated variances of the OLS and GLS estimators are nearly three times the magnitude of the corresponding variance estimates of the OLS calculated under the independence assumption [see (6.6.14)]. Estimated 95% confidence bounds for the slope β_1 using the GLS estimate are $-0.0216 \pm 1.96(0.00006)^{1/2} = -0.0216 \pm .0048$, indicating a significant decreasing trend in the level of Lake Huron during the years 1875–1972.

The iterative procedure described above was used to produce maximum likelihood estimates of the parameters. The calculations using ITSM are analogous to those in Example 6.6.1. The results from each iteration are summarized in Table 6.3. As in Example 6.6.1, the convergence of the estimates is very rapid. □

Example 6.6.3 Seat-Belt Legislation; SBL.TSM

Figure 6-18 shows the numbers of monthly deaths and serious injuries Y_t , $t = 1, \dots, 120$, on UK roads for 10 years beginning in January 1975. They are filed as SBL.TSM. Seat-belt legislation was introduced in February 1983 in the hope of reducing the mean number of monthly “deaths and serious injuries,” (from $t = 99$ onwards). In order to study whether or not there was a drop in mean from that time onwards, we consider the regression,

$$Y_t = a + bf(t) + W_t, \quad t = 1, \dots, 120, \quad (6.6.15)$$

where $f_t = 0$ for $1 \leq t \leq 98$, and $f_t = 1$ for $t \geq 99$. The seat-belt legislation will be considered effective if the estimated value of the regression coefficient b is significantly negative. This problem also falls under the heading of *intervention analysis* (see Section 11.2).

OLS regression based on the model (6.6.15) suggests that the error sequence $\{W_t\}$ is highly correlated with a strong seasonal component of period 12. (To do the regression using ITSM open the file SBL.TSM, select `Regression>Specify`, check only `Include intercept term` and `Include auxiliary variables`, press the `Browse` button, and select the file SBLIN.TSM, which contains the function f_t of (6.6.15) and enter 1 for the number of columns. Then select the

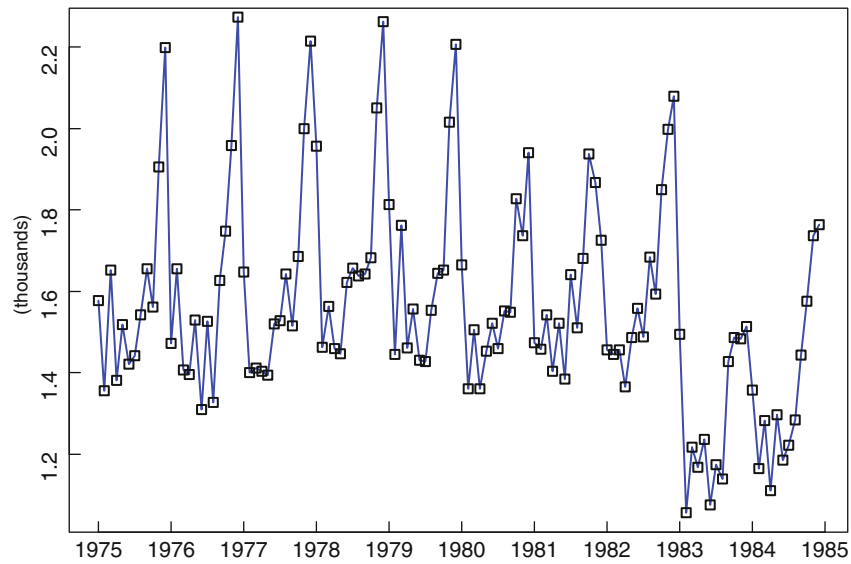


Figure 6-18
Monthly deaths and serious
injuries $\{Y_t\}$ on UK roads,
January 1975–December
1984

option `Regression>Estimation>Generalized LS`. The estimates of the coefficients a and b are displayed in the `Regression estimates` window, and the data become the estimates of the residuals $\{W_t\}$.) The graphs of the data and sample ACF clearly suggest a strong seasonal component with period 12. In order to transform the model (6.6.15) into one with stationary residuals, we therefore consider the differenced data $X_t = Y_t - Y_{t-12}$, which satisfy

$$X_t = bg_t + N_t, \quad t = 13, \dots, 120, \quad (6.6.16)$$

where $g_t = 1$ for $98 \leq t \leq 110$, $g_t = 0$ otherwise, and $\{N_t = W_t - W_{t-12}\}$ is a stationary sequence to be represented by a suitably chosen ARMA model. The series $\{X_t\}$ is contained in the file `SBLD.TSM`, and the function g_t is contained in the file `SBLDIN.TSM`.

The next step is to perform ordinary least squares regression of X_t on g_t following steps analogous to those of the previous paragraph (but this time checking *only* the box marked `Include auxiliary variables` in the `Regression Trend Function` dialog box) and again using the option `Regression>Estimation>Generalized LS` or pressing the blue `GLS` button. The model

$$X_t = -346.92g_t + N_t, \quad (6.6.17)$$

is then displayed in the `Regression estimates` window together with the assumed noise model (white noise in this case). Inspection of the sample ACF of the residuals suggests an MA(13) or AR(13) model for $\{N_t\}$. Fitting AR and MA models of order up to 13 (with no mean-correction) using the option `Model>Estimation>Autofit` gives an MA(12) model as the minimum AICC fit for the residuals. Once this model has been fitted, the model in the `Regression estimates` window is automatically updated to

$$X_t = -328.45g_t + N_t, \quad (6.6.18)$$

with the fitted MA(12) model for the residuals also displayed. After several iterations (each iteration is performed by pressing the `MLE` button) we arrive at the model

$$X_t = -328.45g_t + N_t, \quad (6.6.19)$$

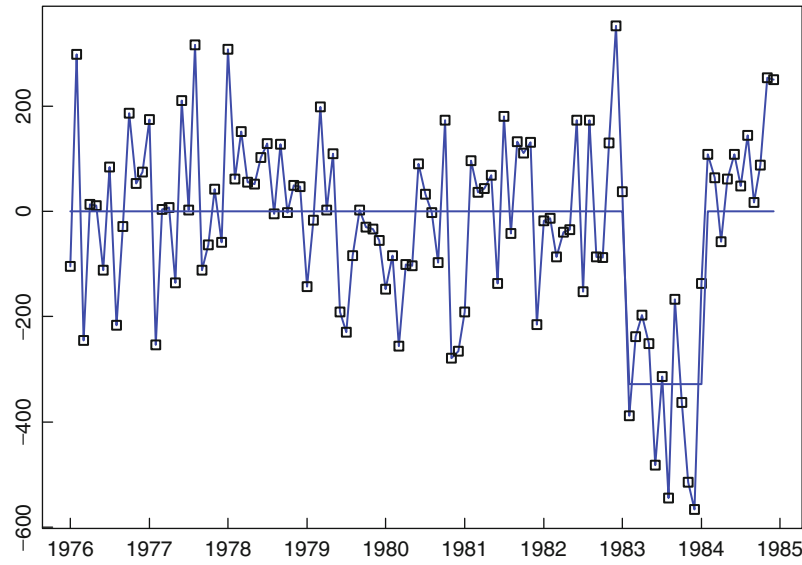


Figure 6-19

The differenced deaths and serious injuries on UK roads $\{X_t = Y_t - Y_{t-12}\}$, showing the fitted GLS regression line

with

$$N_t = Z_t + 0.219Z_{t-1} + 0.098Z_{t-2} + 0.031Z_{t-3} + 0.064Z_{t-4} + 0.069Z_{t-5} + 0.111Z_{t-6} + 0.081Z_{t-7} + 0.057Z_{t-8} + 0.092Z_{t-9} - 0.028Z_{t-10} + 0.183Z_{t-11} - 0.627Z_{t-12},$$

where $\{Z_t\} \sim \text{WN}(0, 12, 581)$. The estimated standard deviation of the regression coefficient estimator is 49.41, so the estimated coefficient, -328.45 , is very significantly negative, indicating the effectiveness of the legislation. The differenced data are shown in Figure 6-19 with the fitted regression function.

□

Problems

6.1 Suppose that $\{X_t\}$ is an $\text{ARIMA}(p, d, q)$ process satisfying the difference equations

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2).$$

Show that these difference equations are also satisfied by the process $W_t = X_t + A_0 + A_1 t + \dots + A_{d-1} t^{d-1}$, where A_0, \dots, A_{d-1} are arbitrary random variables.

6.2 Verify the representation given in (6.3.4).

6.3 Test the data in Example 6.3.1 for the presence of a unit root in an $\text{AR}(2)$ model using the augmented Dickey–Fuller test.

6.4 Apply the augmented Dickey–Fuller test to the levels of Lake Huron data (LAKE.TSM). Perform two analyses assuming $\text{AR}(1)$ and $\text{AR}(2)$ models.

6.5 If $\{Y_t\}$ is a causal ARMA process (with zero mean) and if X_0 is a random variable with finite second moment such that X_0 is uncorrelated with Y_t for each $t = 1, 2, \dots$, show that the best linear predictor of Y_{n+1} in terms of $1, X_0, Y_1, \dots, Y_n$ is the same as the best linear predictor of Y_{n+1} in terms of $1, Y_1, \dots, Y_n$.

6.6 Let $\{X_t\}$ be the ARIMA(2,1,0) process satisfying

$$(1 - 0.8B + 0.25B^2) \nabla X_t = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 1).$$

- (a) Determine the forecast function $g(h) = P_n X_{n+h}$ for $h > 0$.
- (b) Assuming that n is large, compute $\sigma_n^2(h)$ for $h = 1, \dots, 5$.

6.7 Use a text editor to create a new data set ASHORT.TSM that consists of the data in AIRPASS.TSM with the last 12 values deleted. Use ITSM to find an ARIMA model for the logarithms of the data in ASHORT.TSM. Your analysis should include

- (a) a logical explanation of the steps taken to find the chosen model,
- (b) approximate 95 % bounds for the components of ϕ and θ ,
- (c) an examination of the residuals to check for whiteness as described in Section 1.6,
- (d) a graph of the series ASHORT.TSM showing forecasts of the next 12 values and 95 % prediction bounds for the forecasts,
- (e) numerical values for the 12-step ahead forecast and the corresponding 95 % prediction bounds,
- (f) a table of the actual forecast errors, i.e., the true value (deleted from AIRPASS.TSM) minus the forecast value, for each of the 12 forecasts.

Does the last value of AIRPASS.TSM lie within the corresponding 95 % prediction bounds?

6.8 Repeat Problem 6.7, but instead of differencing, apply the classical decomposition method to the logarithms of the data in ASHORT.TSM by deseasonalizing, subtracting a quadratic trend, and then finding an appropriate ARMA model for the residuals. Compare the 12 forecast errors found from this approach with those found in Problem 6.7.

6.9 Repeat Problem 6.7 for the series BEER.TSM, deleting the last 12 values to create a file named BSHORT.TSM.

6.10 Repeat Problem 6.8 for the series BEER.TSM and the shortened series BSHORT.TSM.

6.11 A time series $\{X_t\}$ is differenced at lag 12, then at lag 1 to produce a zero-mean series $\{Y_t\}$ with the following sample ACF:

$$\begin{aligned} \hat{\rho}(12j) &\approx (0.8)^j, & j = 0, \pm 1, \pm 2, \dots, \\ \hat{\rho}(12j \pm 1) &\approx (0.4)(0.8)^j, & j = 0, \pm 1, \pm 2, \dots, \\ \hat{\rho}(h) &\approx 0, & \text{otherwise,} \end{aligned}$$

and $\hat{\gamma}(0) = 25$.

- (a) Suggest a SARIMA model for $\{X_t\}$ specifying all parameters.
- (b) For large n , express the one- and twelve-step linear predictors $P_n X_{n+1}$ and $P_n X_{n+12}$ in terms of X_t , $t = -12, -11, \dots, n$, and $Y_t - \hat{Y}_t$, $t = 1, \dots, n$.
- (c) Find the mean squared errors of the predictors in (b).

6.12 Use ITSM to verify the calculations of Examples 6.6.1–6.6.3.

6.13 The file TUNDRA.TSM contains the average maximum temperature over the month of February for the years 1895–1993 in an area of the USA whose vegetation is characterized as tundra.

-
- (a) Fit a straight line to the data using OLS. Is the slope of the line significantly different from zero?
 - (b) Find an appropriate ARMA model to the residuals from the OLS fit in (a).
 - (c) Calculate the MLE estimates of the intercept and the slope of the line and the ARMA parameters in (a). Is the slope of the line significantly different from zero?
 - (d) Use your model to forecast the average maximum temperature for the years 1994–2004.