# CHAPTER 11

# TIME SERIES REGRESSION MODELS

In this chapter, we introduce several useful ideas that incorporate external information into time series modeling. We start with models that include the effects of interventions on time series' normal behavior. We also consider models that assimilate the effects of outliers—observations, either in the observed series or in the error terms, that are highly unusual relative to normal behavior. Lastly, we develop methods to look for and deal with spurious correlation—correlation between series that is artificial and will not help model or understand the time series of interest. We will see that prewhitening of series helps us find meaningful relationships.

## 11.1 Intervention Analysis

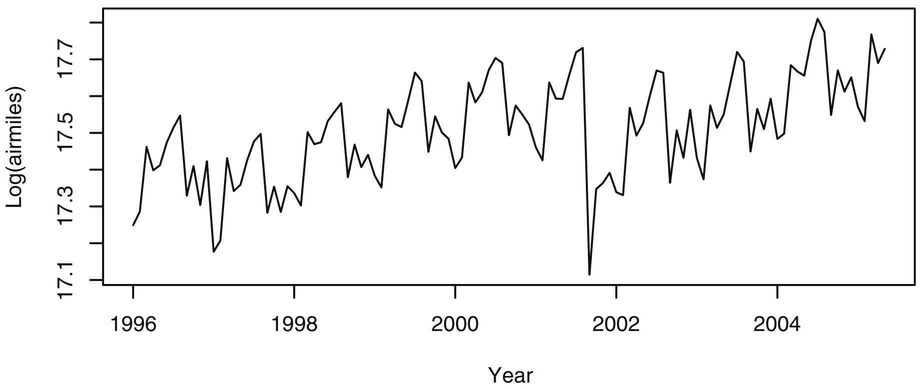
Exhibit 11.1 shows the time plot of the logarithms of monthly airline passenger-miles in the United States from January 1996 through May 2005. The time series is highly seasonal, displaying the fact that air traffic is generally higher during the summer months and the December holidays and lower in the winter months.[†] Also, air traffic was increasing somewhat linearly overall until it had a sudden drop in September 2001. The sudden drop in the number of air passengers in September 2001 and several months thereafter was triggered by the terrorist acts on September 11, 2001, when four planes were hijacked, three of which were crashed into the twin towers of the World Trade Center and the Pentagon and the fourth into a rural field in Pennsylvania. The terrorist attacks of September 2001 deeply depressed air traffic around that period, but air traffic gradually regained the losses as time went on. This is an example of an intervention that results in a change in the trend of a time series.

**Intervention analysis**, introduced by Box and Tiao (1975), provides a framework for assessing the effect of an intervention on a time series under study. It is assumed that the intervention affects the process by changing the mean function or trend of a time series. Interventions can be natural or man-made. For example, some animal population levels crashed to a very low level in a particular year because of extreme climate in that year. The postcrash annual population level may then be expected to be different from that in the precrash period. Another example is the increase of the speed limit from 65 miles per hour to 70 miles per hour on an interstate highway. This may make driving on

[†] In the exercises, we ask you to display the time series plot using seasonal plotting symbols on a full-screen graph, where the seasonality is quite easy to see.

the highway more dangerous. On the other hand, drivers may stay on the highway for a shorter length of time because of the faster speed, so the net effect of the increased speed limit change is unclear. The effect of the increase in speed limit may be studied by analyzing the mean function of some accident time series data; for example, the quarterly number of fatal car accidents on some segment of an interstate highway. (Note that the autocovariance function of the time series might also be changed by the intervention, but this possibility will not be pursued here.)

**Exhibit 11.1 Monthly U.S. Airline Miles: January 1996 through May 2005**



```
> win.graph(width=4.875,height=2.5,pointsize=8)
> data(airmiles)
> plot(log(airmiles),ylab='Log(airmiles)',xlab='Year')
```

We first consider the simple case of a single intervention. The general model for the time series $\{Y_t\}$, perhaps after suitable transformation, is given by

$$Y_t = m_t + N_t \tag{11.1.1}$$

where $m_t$ is the change in the mean function and $N_t$ is modeled as some ARIMA process, possibly seasonal. The process $\{N_t\}$ represents the underlying time series were there no intervention. It is referred to as the natural or unperturbed process, and it may be stationary or nonstationary, seasonal or nonseasonal. Suppose the time series is subject to an intervention that takes place at time $T$. Before $T$, $m_t$ is assumed to be identically zero. The time series $\{Y_t, t < T\}$ is referred to as the **preintervention data** and can be used to specify the model for the unperturbed process $N_t$.

Based on subject matter considerations, the effect of the intervention on the mean function can often be specified up to some parameters. A useful function in this specification is the **step function**

$$S_t^{(T)} = \begin{cases} 1, \text{ if } t \ge T \\ 0, \text{ otherwise} \end{cases} \tag{11.1.2}$$

that is 0 during the preintervention period and 1 throughout the postintervention period. The **pulse function**

$$P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)} \tag{11.1.3}$$

equals 1 at $t = T$ and 0 otherwise. That is, $P_t^{(T)}$ is the indicator or dummy variable flagging the time that the intervention takes place. If the intervention results in an immediate and permanent shift in the mean function, the shift can be modeled as

$$m_t = \omega S_t^{(T)} \tag{11.1.4}$$

where ω is the unknown permanent change in the mean due to the intervention. Testing whether $\omega = 0$ or not is similar to testing whether the population means are the same with data in the form of two independent random samples from the two populations. However, the major difference here is that the pre- and postintervention data cannot generally be assumed to be independent and identically distributed. The inherent serial correlation in the data makes the problem more interesting but at the same time more difficult. If there is a **delay** of $d$ time units before the intervention takes effect and $d$ is known, then we can specify

$$m_t = \omega S_{t-d}^{(T)} \tag{11.1.5}$$

In practice, the intervention may affect the mean function gradually, with its full force reflected only in the long run. This can be modeled by specifying $m_t$ as an AR(1)-type model with the error term replaced by a multiple of the lag 1 of $S_t^{(T)}$:

$$m_t = \delta m_{t-1} + \omega S_{t-1}^{(T)} \tag{11.1.6}$$

with the initial condition $m_0 = 0$. After some algebra, it can be shown that

$$m_t = \begin{cases} \omega \dfrac{1-\delta^{t-T}}{1-\delta}, \text{ for } t > T \\ 0, \text{ otherwise} \end{cases} \tag{11.1.7}$$

Often δ is selected in the range $1 > \delta > 0$. In that case, $m_t$ approaches $\omega/(1-\delta)$ for large $t$, which is the ultimate change (gain or loss) for the mean function. Half of the ultimate change is attained when $1 - \delta^{t-T} = 0.5$; that is, when $t = T + \log(0.5)/\log(\delta)$. The duration $\log(0.5)/\log(\delta)$ is called the **half-life** of the intervention effect, and the shorter it is, the quicker the ultimate change is felt by the system. Exhibit 11.2 displays the half-life as a function of δ, which shows that the half-life increases with δ. Indeed, the half-life becomes infinitely large when δ approaches 1.

**Exhibit 11.2 Half-life based on an AR(1) Process with Step Function Input**

| δ | 0.2 | 0.4 | 0.6 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|
| Half-life | 0.43 | 0.76 | 1.46 | 3.11 | 6.58 | ∞ |

It is interesting to note the limiting case when $\delta = 1$. Then $m_t = \omega(T - t)$ for $t \geq T$ and 0 otherwise. The time sequence plot of $m_t$ displays the shape of a ramp with slope $\omega$. This specification implies that the intervention changes the mean function linearly in the postintervention period. This ramp effect (with a one time unit delay) is shown in Exhibit 11.3 (c).

Short-lived intervention effects may be specified using the pulse dummy variable

$$P_t^{(T)} = \begin{cases} 1, \text{ if } t = T \\ 0, \text{ otherwise} \end{cases} \tag{11.1.8}$$

For example, if the intervention impacts the mean function only at $t = T$, then

$$m_t = \omega P_t^{(T)} \tag{11.1.9}$$

Intervention effects that die out gradually may be specified via the AR(1)-type specification

$$m_t = \delta m_{t-1} + \omega P_t^{(T)} \tag{11.1.10}$$

That is, $m_t = \omega\delta^{T-t}$ for $t \geq T$ so that the mean changes immediately by an amount $\omega$ and subsequently the change in the mean decreases geometrically by the common factor of $\delta$; see Exhibit 11.4 (a). Delayed changes can be incorporated by lagging the pulse function. For example, if the change in the mean takes place after a delay of one time unit and the effect dies out gradually, we can specify

$$m_t = \delta m_{t-1} + \omega P_{t-1}^{(T)} \tag{11.1.11}$$

Again, we assume the initial condition $m_0 = 0$.

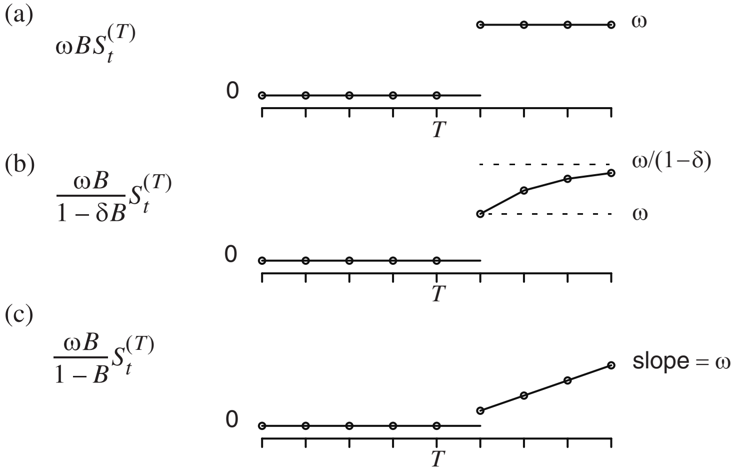
It is useful to write[†] the preceding model in terms of the backshift operator $B$, where $Bm_t = m_{t-1}$ and $BP_t^{(T)} = P_{t-1}^{(T)}$. Then $(1 - \delta B)m_t = \omega BP_t^{(T)}$. Or, we can write

$$m_t = \frac{\omega B}{1 - \delta B} P_t^{(T)} \tag{11.1.12}$$

Recall $(1 - B)S_t^{(T)} = P_t^{(T)}$, which can be rewritten as $S_t^{(T)} = \frac{1}{1 - B} P_t^{(T)}$.

[†] The remainder of this chapter makes use of the backshift operator introduced in Appendix D on page 106. You may want to review that appendix before proceeding further.

**Exhibit 11.3 Some Common Models for Step Response Interventions (All are shown with a delay of 1 time unit)**



Several specifications can be combined to model more sophisticated intervention effects.

For example,

$$m_t = \frac{\omega_1 B}{1-\delta B} P_t^{(T)} + \frac{\omega_2 B}{1-B} P_t^{(T)} \tag{11.1.13}$$

depicts the situation displayed in Exhibit 11.4 (b) where $\omega_1$ and $\omega_2$ are both greater than zero, and

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1 B}{1-\delta B} P_t^{(T)} + \frac{\omega_2 B}{1-B} P_t^{(T)} \tag{11.1.14}$$

may model situations like Exhibit 11.4 (c) with $\omega_1$ and $\omega_2$ both negative. This last case may model the interesting situation where a special sale may cause strong rush buying, initially so much so that the sale is followed by depressed demand. More generally, we can model the change in the mean function by an ARMA-type specification

$$m_t = \frac{\omega(B)}{\delta(B)} P_t^{(T)} \tag{11.1.15}$$

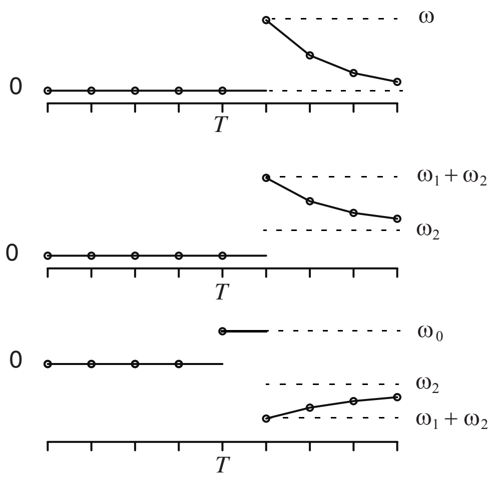
where $\omega(B)$ and $\delta(B)$ are some polynomials in $B$. Because $(1-B)S_t^{(T)} = P_t^{(T)}$, the model for $m_t$ can be specified in terms of either the pulse or step dummy variable.

**Exhibit 11.4 Some Common Models for Pulse Response Interventions (All are shown with a delay of 1 time unit)**

(a) $\dfrac{\omega B}{1-\delta B}P_t^{(T)}$

(b) $\left[\dfrac{\omega_1 B}{1-\delta B}+\dfrac{\omega_2 B}{1-B}\right]P_t^{(T)}$

(c) $\left[\omega_0+\dfrac{\omega_1 B}{1-\delta B}+\dfrac{\omega_2 B}{1-B}\right]P_t^{(T)}$



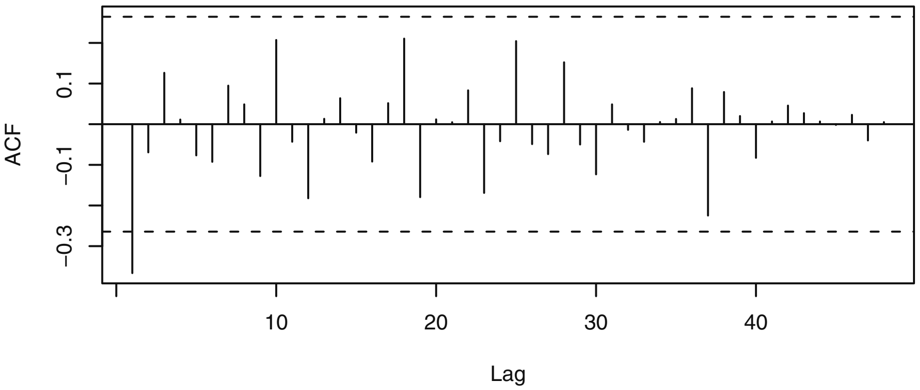
Estimation of the parameters of an intervention model may be carried out by the method of maximum likelihood estimation. Indeed, $Y_t - m_t$ is a seasonal ARIMA process so that the likelihood function equals the joint pdf of $Y_t - m_t$, $t = 1, 2,\dots, n$, which can be computed by methods studied in Chapter 7 or else by the state space modeling methods of Appendix H on page 222.

We now revisit the monthly passenger-airmiles data. Recall that the terrorist acts in September 2001 had lingering depressing effects on air traffic. The intervention may be specified as an AR(1) process with the pulse input at September 2001. But the unexpected turn of events in September 2001 had a strong instantaneous chilling effect on air traffic. Thus, we model the intervention effect (the 9/11 effect) as

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B}P_t^{(T)}$$

where $T$ denotes September 2001. In this specification, $\omega_0 + \omega_1$ represents the instantaneous 9/11 effect, and, for $k \geq 1$, $\omega_1(\omega_2)^k$ gives the 9/11 effect $k$ months afterward. It remains to specify the seasonal ARIMA structure of the underlying unperturbed process. Based on the preintervention data, an ARIMA$(0,1,1)\times(0,1,0)_{12}$ model was tentatively specified for the unperturbed process; see Exhibit 11.5.

**Exhibit 11.5 Sample ACF for (1−B)(1−B$^{12}$) Log(Air Passenger Miles) Over the Preintervention Period**



```
> acf(as.vector(diff(diff(window(log(airmiles),end=c(2001,8)),
   12))),lag.max=48)
```

Model diagnostics of the fitted model suggested that a seasonal MA(1) coefficient was needed and the existence of some *additive* outliers occurring in December 1996, January 1997, and December 2002. (Outliers will be discussed in more detail later; here additive outliers may be regarded as interventions of unknown nature that have a pulse response function.) Hence, the model is specified as an ARIMA(0,1,1)×(0,1,1)$_{12}$ plus the 9/11 intervention and three additive outliers. The fitted model is summarized in Exhibit 11.6.

**Exhibit 11.6 Estimation of Intervention Model for Logarithms of Air Miles (Standard errors are shown below the estimates)**

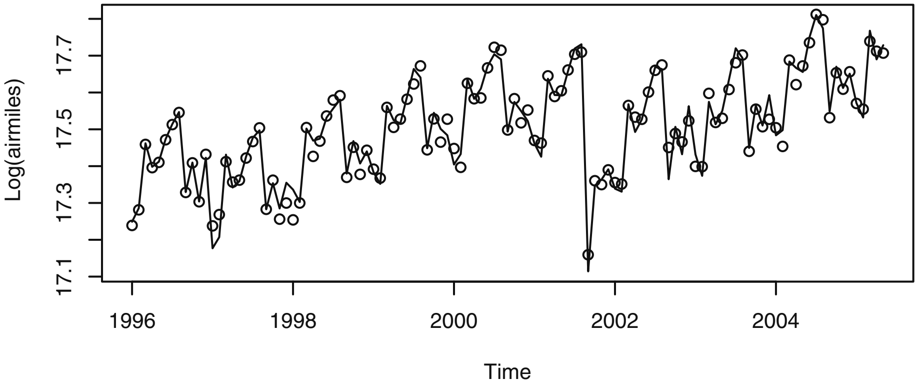
| $\theta$ | $\Theta$ | Dec96 | Jan97 | Dec02 | $\omega_0$ | $\omega_1$ | $\omega_2$ |
|---|---|---|---|---|---|---|---|
| 0.383 | 0.650 | 0.099 | −0.069 | 0.081 | −0.095 | −0.27 | 0.814 |
| (0.093) | (0.119) | (0.023) | (0.022) | (0.020) | (0.046) | (0.044) | (0.098) |

$\sigma^2$ estimated as 0.000672: log-likelihood = 219.99, AIC= −423.98

```
> air.m1=arimax(log(airmiles),order=c(0,1,1),
   seasonal=list(order=c(0,1,1),period=12),
   xtransf=data.frame(I911=1*(seq(airmiles)==69),
   I911=1*(seq(airmiles)==69)),transfer=list(c(0,0),c(1,0)),
   xreg=data.frame(Dec96=1*(seq(airmiles)==12),
   Jan97=1*(seq(airmiles)==13),Dec02=1*(seq(airmiles)==84)),
   method='ML')
> air.m1
```

Model diagnostics suggested that the fitted model above provides a good fit to the data. The open circles in the time series plot shown in Exhibit 11.7 represent the fitted values from the final estimated model. They indicate generally good agreement between the model and the data.

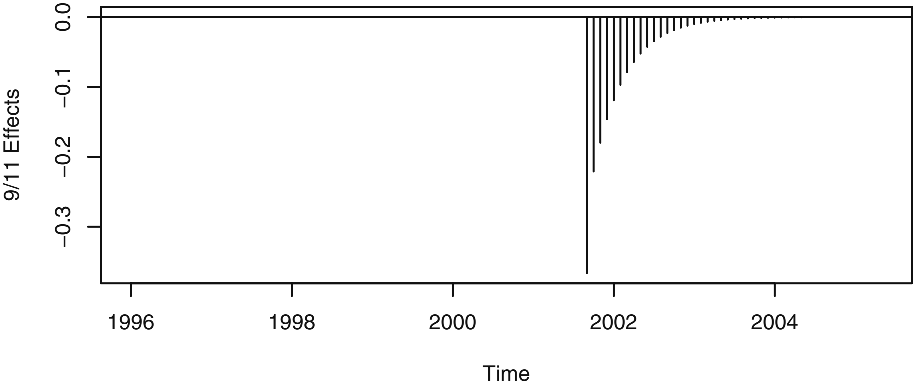
**Exhibit 11.7 Logs of Air Passenger Miles and Fitted Values**



```
> plot(log(airmiles),ylab='Log(airmiles)')
> points(fitted(air.m1))
```

The fitted model estimates that the 9/11 intervention reduced air traffic by 31% = {1 – exp(–0.0949–0.2715)} × 100% in September 2001, and air traffic $k$ months later was lowered by $\{1 - \exp(-0.2715 \times 0.8139^k)\} \times 100\%$. Exhibit 11.8 graphs the estimated 9/11 effects on air traffic, which indicate that air traffic regained its losses toward the end of 2003.

**Exhibit 11.8 The Estimated 9/11 Effects for the Air Passenger Series**



```
> Nine11p=1*(seq(airmiles)==69)
> plot(ts(Nine11p*(-0.0949)+
filter(Nine11p,filter=.8139,method='recursive', side=1)*
(-0.2715),frequency=12,start=1996),ylab='9/11 Effects',
type='h'); abline(h=0)
```

## 11.2 Outliers

Outliers refer to atypical observations that may arise because of measurement and/or copying errors or because of abrupt, short-term changes in the underlying process. For time series, two kinds of outliers can be distinguished, namely **additive** outliers and **innovative** outliers. These two kinds of outliers are often abbreviated as AO and IO, respectively. An additive outlier occurs at time $T$ if the underlying process is perturbed additively at time $T$ so that the data equal

$$Y'_t = Y_t + \omega_A P_t^{(T)} \tag{11.2.1}$$

where $\{Y_t\}$ is the unperturbed process. Henceforth in this section, $Y'$ denotes the observed process that may be affected by some outliers and $Y$ the unperturbed process should there be no outliers. Thus, $Y'_T = Y_T + \omega_A$ but $Y'_t = Y_t$ otherwise, so the time series is only affected at time $T$ if it has an additive outlier at $T$. An additive outlier can also be treated as an intervention that has a pulse response at $T$ so that $m_t = \omega_A P_t^{(T)}$.

On the other hand, an innovative outlier occurs at time $t$ if the error (also known as an innovation) at time $t$ is perturbed (that is, the errors equal $e'_t = e_t + \omega_I P_t^{(T)}$, where $e_t$ is a zero-mean white noise process). So, $e'_T = e_T + \omega_I$ but $e'_t = e_t$ otherwise. Suppose that the unperturbed process is stationary and admits an MA(∞) representation

$$Y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots$$

Consequently, the perturbed process can be written

$$\begin{aligned} Y'_t &= e'_t + \psi_1 e'_{t-1} + \psi_2 e'_{t-2} + \cdots \\ &= [e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \cdots] + \psi_{t-T}\omega_I \end{aligned}$$

or

$$Y'_t = Y_t + \psi_{t-T}\omega_I \tag{11.2.2}$$

where $\psi_0 = 1$ and $\psi_j = 0$ for negative $j$. Thus, an innovative outlier at $T$ perturbs all observations on and after $T$, although with diminishing effect, as the observation is further away from the origin of the outlier.

To detect whether an observation is an AO or IO, we use the AR(∞) representation of the unperturbed process to define the residuals:

$$a_t = Y'_t - \pi_1 Y'_{t-1} - \pi_2 Y'_{t-2} - \cdots \tag{11.2.3}$$

For simplicity, we assume the process has zero mean and that the parameters are known. In practice, the unknown parameter values are replaced by their estimates from the possibly perturbed data. Under the null hypothesis of no outliers and for large samples, this has a negligible effect on the properties of the test procedures described below. If the series has exactly one IO at time $T$, then the residual $a_T = \omega_I + e_T$ but $a_t = e_t$ otherwise. So $\omega_I$ can be estimated by $\tilde{\omega}_I = a_T$ with variance equal to $\sigma^2$. Thus, a test statistic for testing for an IO at $T$ is

$$\lambda_{1,T} = \frac{a_T}{\sigma} \tag{11.2.4}$$

which has (approximately) a standard normal distribution under the null hypothesis that there are no outliers in the time series. When $T$ is known beforehand, the observation in question is declared an outlier if the corresponding standardized residual exceeds 1.96 in magnitude at the 5% significance level. In practice, there is often no prior knowledge about $T$, and the test is applied to all observations. In addition, $\sigma$ will need to be estimated. A simple conservative procedure is to use the Bonferroni rule for controlling the overall error rate of multiple tests. Let

$$\lambda_1 = \max_{1 \le t \le n} |\lambda_{1,t}| \tag{11.2.5}$$

be attained at $t = T$. Then the $T$th observation is deemed an IO if $\lambda_1$ exceeds the upper $0.025/n \times 100$ percentile of the standard normal distribution. This procedure guarantees that there is at most a 5% probability of a false detection of an IO. Note that an outlier will inflate the maximum likelihood estimate of $\sigma$, so if there is no adjustment for outliers, the power of most tests is usually reduced. A robust estimate of the noise standard deviation may be used in lieu of the maximum likelihood estimate to increase the power of the test. For example, $\sigma$ can be more robustly estimated by the mean absolute residual times $\sqrt{2/\pi}$.

The detection of an AO is more complex. Suppose that the process admits an AO at $T$ and is otherwise free of outliers. Then it can be shown that

$$a_t = -\omega_A \pi_{t-T} + e_t \tag{11.2.6}$$

where $\pi_0 = -1$ and $\pi_j = 0$ for negative $j$. Hence, $a_t = e_t$ for $t < T$, $a_T = \omega_A + e_T$, $a_{T+1} = -\omega_A\pi_1 + e_{T+1}$, $a_{T+2} = -\omega_A\pi_2 + e_{T+2}$, and so forth. A least squares estimator of $\omega_A$ is

$$\tilde{\omega}_{T,A} = -\rho^2 \sum_{t=1}^{n} \pi_{t-T} a_t \tag{11.2.7}$$

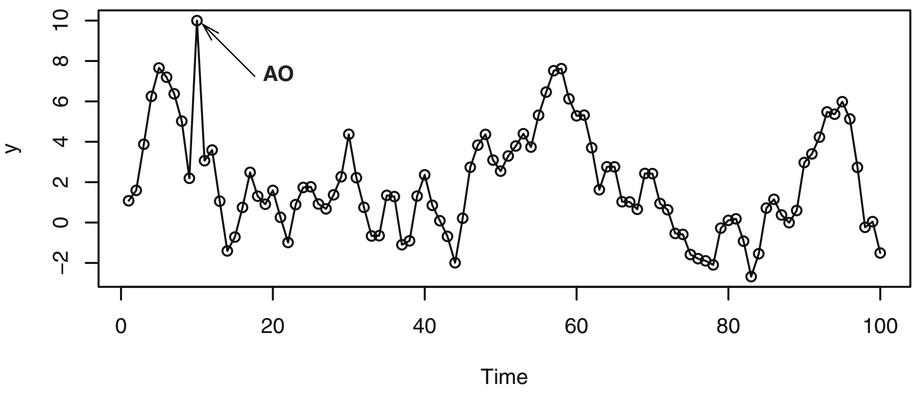
where $\rho^2 = (1 + \pi_1^2 + \pi_2^2 + \cdots + \pi_{n-T}^2)^{-1}$, with the variance of the estimate being equal to $\rho^2\sigma^2$. We can then define

$$\lambda_{2,T} = \frac{\tilde{\omega}_{T,A}}{\rho\sigma} \tag{11.2.8}$$

as the test statistic for testing the null hypothesis that the time series has no outliers versus the alternative hypothesis of an AO at $T$. As before, $\rho$ and $\sigma$ will need to be estimated. The test statistic $\lambda_{2,T}$ is approximately distributed as $N(0,1)$ under the null hypothesis. Again, $T$ is often unknown, and the test is applied repeatedly to each time point. The Bonferroni rule may again be applied to control the overall error rate. Furthermore, the nature of an outlier is not known beforehand. In the case where an outlier is detected at $T$, it may be classified to be an IO if $|\lambda_{1,T}| > |\lambda_{2,T}|$ and an AO otherwise. See Chang et al. (1988) for another approach to classifying the nature of an outlier. When an outlier is found, it can be incorporated into the model, and the outlier-detection procedure can then be repeated with the refined model until no more outliers are found.

As a first example, we simulated a time series of length $n = 100$ from the ARIMA(1,0,1) model with $\phi = 0.8$ and $\theta = -0.5$. We then changed the 10th observation from −2.13 to 10 (that is, $\omega_A = 12.13$); see Exhibit 11.9. Based on the sample ACF, PACF and EACF, an AR(1) model was tentatively identified. Based on the Bonferroni rule, the 9th, 10th, and 11th observations were found to be possible additive outliers with the corresponding robustified test statistics being −3.54, 9.55, and −5.20. The test for IO revealed that the 10th and 11th observations may be IO, with the corresponding robustified test statistics being 7.11 and −6.64. Because among the tests for AO and IO the largest magnitude occurs for the test for AO at $T = 10$, the 10th observation was tentatively marked as an AO. Note that the nonrobustified test statistic for AO at $T = 10$ equals 7.49, which is substantially less than the more robust test value of 9.55, showing that robustifying the estimate of the noise standard deviation does increase the power of the test. After incorporating the AO in the model, no more outliers were found. However, the lag 1 residual ACF was significant, suggesting the need for an MA(1) component. Hence, an ARIMA(1,0,1) + AO at $T = 10$ model was fitted to the data. This model was found to have no additional outliers and passed all model diagnostic checks.

**Exhibit 11.9 Simulated ARIMA(1,0,1) Process with an Additive Outlier**



```
> The extensive R code for the simulation and analysis of this
  example may be found in the R code script file for Chapter 11.
```

For a real example, we return to the seasonal ARIMA(0,1,1)×(0,1,1)$_{12}$ model that we fitted to the carbon dioxide time series in Chapter 10. The time series plot of the standardized residuals from this model, shown in Exhibit 10.11 on page 238, showed a suspiciously large standardized residual in September 1998. Calculation shows that there is no evidence of an additive outlier, as $\lambda_{2,\,t}$ is not significantly large for any $t$. However, the robustified $\lambda_1 = \max_{1 \le t \le n}|\lambda_{1,\,t}| = 3.7527$, which is attained at $t = 57$, corresponding to September 1998. The Bonferroni critical value with $\alpha = 5\%$ and $n = 132$ is 3.5544. So our observed $\lambda_1$ is large enough to claim significance for an innovation outlier in September 1998. Exhibit 11.10 shows the results of fitting the ARIMA(0,1,1) $\times(0,1,1)_{12}$ model with an IO at $t = 57$ to the $CO_2$ time series. These results should be compared with the earlier results shown in Exhibit 10.10 on page 237, where the outlier was not taken into account. Notice that the estimates of $\theta$ and $\Theta$ have not changed very much, the AIC is better (that is, smaller), and the IO effect is highly significant. Diagnostics based on this model turn out to be excellent, no further outliers are detected, and we have a very adequate model for this seasonal time series.

**Exhibit 11.10 ARIMA(0,1,1)×(0,1,1)$_{12}$ Model with IO at *t* = 57 for CO$_2$ Series**

| Coefficient | $\theta$ | $\Theta$ | IO-57 |
|---|---|---|---|
| Estimate | 0.5925 | 0.8274 | 2.6770 |
| Standard Error | 0.0775 | 0.1016 | 0.7246 |

$\hat{\sigma}_e^2 = 0.4869$: log-likelihood = −133.08, AIC = 272.16

```
> m1.co2=arima(co2,order=c(0,1,1),seasonal=list(order=c(0,1,1),
   period=12)); m1.co2
> detectAO(m1.co2); detectIO(m1.co2)
> m4.co2=arimax(co2,order=c(0,1,1),seasonal=list(order=c(0,1,1),
   period=12),io=c(57)); m4.co2
```

## 11.3 Spurious Correlation

A main purpose of building a time series model is for forecasting, and the ARIMA model does this by exploiting the autocorrelation pattern in the data. Often, the time series under study may be related to, or led by, some other covariate time series. For example, Stige et al. (2006) found that pasture production in Africa is generally related to some climatic indices. In such cases, better understanding of the underlying process and/or more accurate forecasts may be achieved by incorporating relevant covariates into the time series model.

Let $Y = \{Y_t\}$ be the time series of the response variable and $X = \{X_t\}$ be a covariate time series that we hope will help explain or forecast $Y$. To explore the correlation structure between $X$ and $Y$ and their lead-led relationship, we define the cross-covariance function $\gamma_{t,s}(X,Y) = Cov(X_t,Y_s)$ for each pair of integers $t$ and $s$. Stationarity of a univariate time series can be easily extended to the case of multivariate time series. For example, $X$ and $Y$ are jointly (weakly) stationary if their means are constant and the covariance $\gamma_{t,s}(X,Y)$ is a function of the time difference $t - s$. For jointly stationary processes, the **cross-correlation function** between $X$ and $Y$ at lag $k$ can then be defined by $\rho_k(X,Y) = Corr(X_t,Y_{t-k}) = Corr(X_{t+k},Y_t)$. Note that if $Y = X$, the cross-correlation becomes the autocorrelation of $Y$ at lag $k$. The coefficient $\rho_0(Y,X)$ measures the contemporaneous linear association between $X$ and $Y$, whereas $\rho_k(X,Y)$ measures the linear association between $X_t$ and that of $Y_{t-k}$. Recall that the autocorrelation function is an even function, that is, $\rho_k(Y,Y) = \rho_{-k}(Y,Y)$. (This is because $Corr(Y_t,Y_{t-k}) = Corr(Y_{t-k},Y_t) = Corr(Y_t,Y_{t+k})$, by stationarity.) However, the cross-correlation function is generally not an even function since $Corr(X_t,Y_{t-k})$ need not equal $Corr(X_t,Y_{t+k})$.

As an illustration, consider the regression model

$$Y_t = \beta_0 + \beta_1 X_{t-d} + e_t \tag{11.3.1}$$

where the $X$'s are independent, identically distributed random variables with variance $\sigma_X^2$ and the $e$'s are also white noise with variance $\sigma_e^2$ and are independent of the $X$'s. It can be checked that the cross-correlation function (CCF) $\rho_k(X,Y)$ is identically zero except for lag $k = -d$, where

$$\rho_{-d}(X, Y) = \frac{\beta_1\sigma_X}{\sqrt{\beta_1^2\sigma_X^2 + \sigma_e^2}} \tag{11.3.2}$$

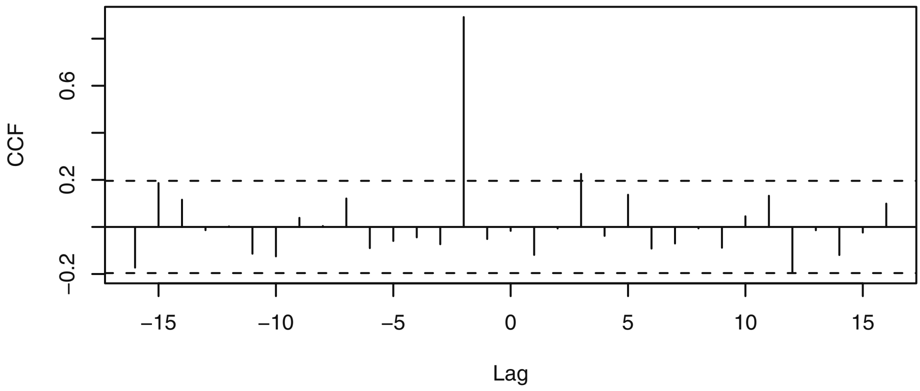
In this case, the theoretical CCF is nonzero only at lag $-d$, reflecting the fact that $X$ is "leading" $Y$ by $d$ units of time. The CCF can be estimated by the **sample cross-correlation function** (sample CCF) defined by

$$r_k(X, Y) = \frac{\sum(X_t - \bar{X})(Y_{t-k} - \bar{Y})}{\sqrt{\sum(X_t - \bar{X})^2}\sqrt{\sum(Y_t - \bar{Y})^2}} \tag{11.3.3}$$

where the summations are done over all data where the summands are available. The sample CCF becomes the sample ACF when $Y = X$. The covariate $X$ is independent of $Y$ if and only if $\beta_1 = 0$, in which case the sample autocorrelation $r_k(X,Y)$ is approximately normally distributed with zero mean and variance $1/n$, where $n$ is the sample size—the number of pairs of $(X_t,Y_t)$ available. Sample cross-correlations that are larger than $1.96/\sqrt{n}$ in magnitude are then deemed significantly different from zero.

We have simulated 100 pairs of $(X_t,Y_t)$ from the model of Equation (11.3.1) with $d = 2$, $\beta_0 = 0$, and $\beta_1 = 1$. The $X$'s and $e$'s are generated as normal random variables distributed as $N(0,1)$ and $N(0,0.25)$, respectively. Theoretically, the CCF should then be zero except at lag $-2$, where it equals $\rho_{-2}(X, Y) = 1/\sqrt{1 + 0.25} = 0.8944$. Exhibit 11.11 shows the sample CCF of the simulated data, which is significant at lags $-2$ and 3. But the sample CCF at lag 3 is quite small and only marginally significant. Such a false alarm is not unexpected as the exhibit displays a total of 33 sample CCF values out of which we may expect $33\times0.05 = 1.65$ false alarms on average.

**Exhibit 11.11 Sample Cross-Correlation from Equation (11.3.1) with *d* = 2**



```
> win.graph(width=4.875,height=2.5,pointsize=8)
> set.seed(12345); X=rnorm(105); Y=zlag(X,2)+.5*rnorm(105)
> X=ts(X[-(1:5)],start=1,freq=1); Y=ts(Y[-(1:5)],start=1,freq=1)
> ccf(X,Y,ylab='CCF')
```

Even though $X_{t-2}$ correlates with $Y_t$, the regression model considered above is rather restrictive, as $X$ and $Y$ are each white noise series. For stationary time series, the response variable and the covariate are each generally autocorrelated, and the error term of the regression model is also generally autocorrelated. Hence a more useful regression model is given by

$$Y_t = \beta_0 + \beta_1 X_{t-d} + Z_t \tag{11.3.4}$$

where $Z_t$ may follow some ARIMA($p,d,q$) model. Even if the processes $X$ and $Y$ are independent of each other ($\beta_1 = 0$), the autocorrelations in $Y$ and $X$ have the unfortunate consequence of implying that the sample CCF is no longer approximately $N(0,1/n)$. Under the assumption that both $X$ and $Y$ are stationary and that they are independent of each other, it turns out that the sample variance tends to be different from $1/n$. Indeed, it may be shown that the variance of $\sqrt{n}r_k(X, Y)$ is approximately

$$1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y) \tag{11.3.5}$$

where $\rho_k(X)$ is the autocorrelation of $X$ at lag $k$ and $\rho_k(Y)$ is similarly defined for the $Y$-process. For refinement of this asymptotic result, see Box et al. (1994, p. 413). Suppose $X$ and $Y$ are both AR(1) processes with AR(1) coefficients $\phi_X$ and $\phi_Y$, respectively. Then $r_k(X,Y)$ is approximately normally distributed with zero mean, but the variance is now approximately equal to

$$\frac{1 + \phi_X\phi_Y}{n(1 - \phi_X\phi_Y)} \tag{11.3.6}$$

When both AR(1) coefficients are close to 1, the ratio of the sampling variance of $r_k(X,Y)$ to the nominal value of $1/n$ approaches infinity. Thus, the unquestioned use of the $1/n$ rule in deciding the significance of the sample CCF may lead to many more false positives than the nominal 5% error rate, even though the response and covariate time series are independent of each other. Exhibit 11.12 shows some numerical results for the case where $\phi_X = \phi_Y = \phi$.

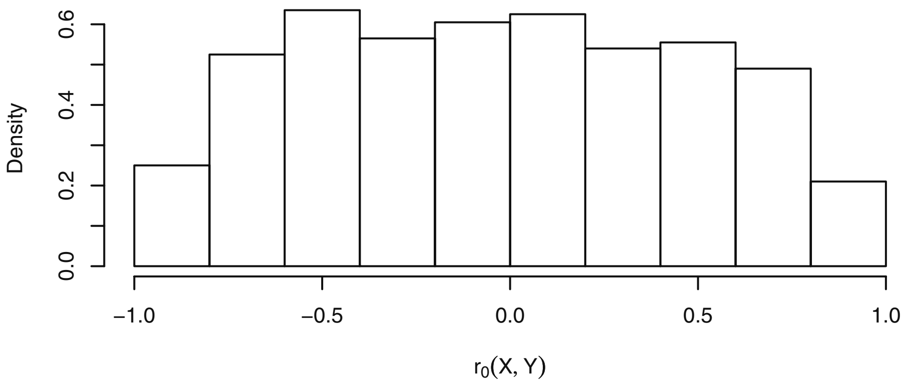
**Exhibit 11.12 Asymptotic Error Rates of a Nominal 5% Test of Independence for a Pair of AR(1) Processes**

| $\phi = \phi_X = \phi_Y$ | 0.00 | 0.15 | 0.30 | 0.45 | 0.60 | 0.75 | 0.90 |
|---|---|---|---|---|---|---|---|
| Error Rate | 5% | 6% | 7% | 11% | 18% | 30% | 53% |

```
> phi=seq(0,.95,.15)
> rejection=2*(1-pnorm(1.96*sqrt((1-phi^2)/(1+phi^2))))
> M=signif(rbind(phi,rejection),2)
> rownames(M)=c('phi', 'Error Rate')
> M
```

The problem of inflated variance of the sample cross-correlation coefficients becomes more acute for nonstationary data. In fact, the sample cross-correlation coefficients may no longer be approximately normally distributed even with a large sample size. Exhibit 11.13 displays the histogram of 1000 simulated lag zero cross-correlations between two independent IMA(1,1) processes each of size 500. An MA(1) coefficient of $\theta = 0.8$ was used for both simulated processes. Note that the distribution of $r_0(X,Y)$ is far from normal and widely dispersed between –1 and 1. See Phillips (1998) for a relevant theoretical discussion.

**Exhibit 11.13 Histogram of 1000 Sample Lag Zero Cross-Correlations of Two Independent IMA(1,1) Processes Each of Size 500**

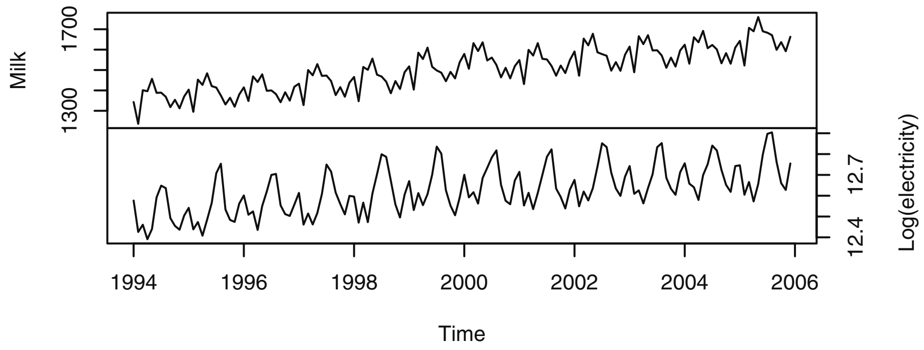


```
> set.seed(23457)
> correlation.v=NULL; B=1000; n=500
> for (i in 1:B) {x=cumsum(arima.sim(model=list(ma=.8),n=n))
> y=cumsum(arima.sim(model=list(ma=.8),n=n))
> correlation.v=c(correlation.v,ccf(x,y,lag.max=1,
   plot=F)$acf[2])}
> hist(correlation.v,prob=T,xlab=expression(r[0](X,Y)))
```

These results provide insight into why we sometimes obtain nonsense (spurious) correlation between time series variables. The phenomenon of spurious correlation was first studied systematically by Yule (1926).

As an example, the monthly milk production and the logarithms of monthly electricity production in the United States from January 1994 to December 2005 are shown in Exhibit 11.14. Both series have an upward trend and are highly seasonal.

**Exhibit 11.14 Monthly Milk Production and Logarithms of Monthly Electricity Production in the U.S.**

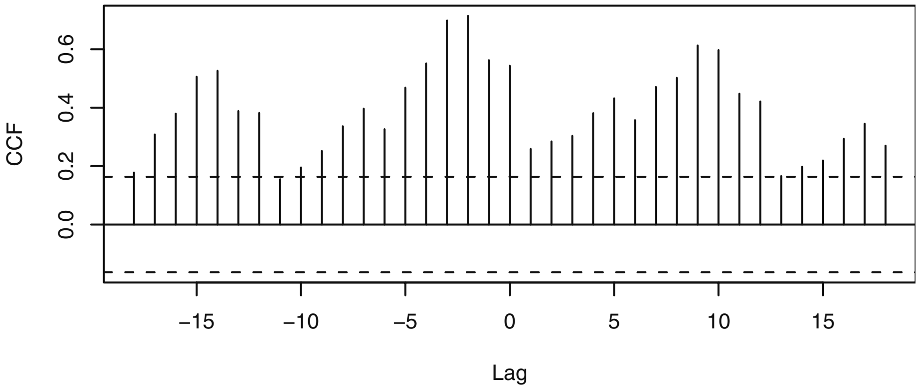


```
> data(milk); data(electricity)
> milk.electricity=ts.intersect(milk,log(electricity))
> plot(milk.electricity,yax.flip=T)
```

Calculation shows that these series have a cross-correlation coefficient at lag zero of 0.54, which is “statistically significantly different from zero” as judged against the standard error criterion of $1.96/\sqrt{n} = 0.16$. Exhibit 11.15 displays the strong cross-correlations between these two variables at a large number of lags.

Needless to say, it is difficult to come up with a plausible reason for the relationship between monthly electricity production and monthly milk production. The nonstationarity in the milk production series and in the electricity series is more likely the cause of the spurious correlations found between the two series. The following section contains further discussion of this example.

**Exhibit 11.15 Sample Cross-Correlation Between Monthly Milk Production and Logarithm of Monthly Electricity Production in the U.S.**



```
> ccf(as.vector(milk.electricity[,1]),
   as.vector(milk.electricity[,2]),ylab='CCF')
```

## 11.4 Prewhitening and Stochastic Regression

In the preceding section, we found that with strongly autocorrelated data it is difficult to assess the dependence between the two processes. Thus, it is pertinent to disentangle the linear association between $X$ and $Y$, say, from their autocorrelation. A useful device for doing this is prewhitening. Recall that, for the case of stationary $X$ and $Y$ that are independent of each other, the variance of $r_k(X, Y)$ is approximately

$$\frac{1}{n}\left[1 + 2\sum_{k=1}^{\infty} \rho_k(X)\rho_k(Y)\right] \qquad (11.4.1)$$

An examination of this formula reveals that the approximate variance is $1/n$ if either one (or both) of $X$ or $Y$ is a white noise process. In practice, the data may be nonstationary, but they may be transformed to approximately white noise by replacing the data by the residuals from a fitted ARIMA model. For example, if $X$ follows an ARIMA(1,1,0) model with no intercept term, then

$$\tilde{X}_t = X_t - X_{t-1} - \phi(X_{t-1} - X_{t-2}) = 1 - (1 + \phi B) + \phi B^2]X_t \qquad (11.4.2)$$

is white noise. More generally, if $X_t$ follows some invertible ARIMA($p$,$d$,$q$) model, then it admits an AR($\infty$) representation

$$\tilde{X}_t = (1 - \pi_1 B - \pi_2 B^2 - \cdots)X_t = \pi(B)X_t$$

where the $\tilde{X}$'s are white noise. The process of transforming the $X$'s to the $\tilde{X}$'s via the *filter* $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$ is known as **whitening** or **prewhitening**. We now can study the CCF between $X$ and $Y$ by prewhitening the $Y$ and $X$ using the *same* filter based on the $X$ process and then computing the CCF of $\tilde{Y}$ and $\tilde{X}$; that is, the prewhitened $Y$ and $X$. Since prewhitening is a *linear* operation, any linear relationships between the original series will be preserved after prewhitening. Note that we have abused the terminology, as $\tilde{Y}$ need not be white noise because the filter $\pi(B)$ is tailor-made only to transform $X$ to a white noise process—not $Y$. We assume, furthermore, that $\tilde{Y}$ is stationary. This approach has two advantages: (i) the statistical significance of the sample CCF of the prewhitened data can be assessed using the cutoff $1.96/\sqrt{n}$, and (ii) the theoretical counterpart of the CCF so estimated is proportional to certain regression coefficients.

To see (ii), consider a more general regression model relating $X$ to $Y$ and, without loss of generality, assume both processes have zero mean:

$$Y_t = \sum_{j=-\infty}^{\infty} \beta_j X_{t-j} + Z_t \tag{11.4.3}$$

where $X$ is independent of $Z$ and the coefficients $\beta$ are such that the process is well-defined. In this model, the coefficients $\beta_k$ could be nonzero for any integer $k$. However, in real applications, the doubly infinite sum is often a finite sum so that the model simplifies to

$$Y_t = \sum_{j=m_1}^{m_2} \beta_j X_{t-j} + Z_t, \tag{11.4.4}$$

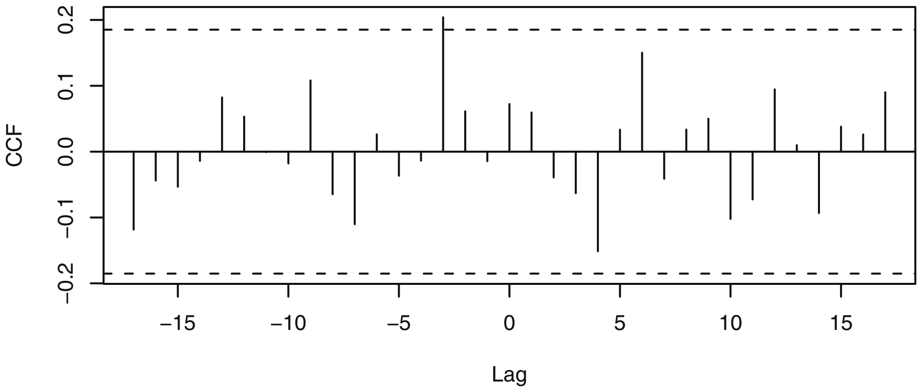
which will be assumed below even though we retain the doubly infinite summation notation for ease of exposition. If the summation ranges only over a finite set of *positive* indices, then $X$ *leads* $Y$ and the covariate $X$ serves as a useful **leading indicator** for future $Y$'s. Applying the filter $\pi(B)$ to both sides of this model, we get

$$\tilde{Y}_t = \sum_{k=-\infty}^{\infty} \beta_k \tilde{X}_{t-k} + \tilde{Z}_t \tag{11.4.5}$$

where $\tilde{Z}_t = Z_t - \pi_1 Z_{t-1} - \pi_2 Z_{t-2} - \cdots$ .The prewhitening procedure thus orthogonalizes the various lags of $X$ in the original regression model. Because $\tilde{X}$ is a white noise sequence and $\tilde{X}$ is independent of $\tilde{Z}$, the theoretical cross-correlation coefficient between $\tilde{X}$ and $\tilde{Y}$ at lag $k$ equals $\beta_{-k}(\sigma_{\tilde{X}}/\sigma_{\tilde{Y}})$. In other words, the theoretical cross-correlation of the prewhitened processes at lag $k$ is proportional to the regression coefficient $\beta_{-k}$.

For a quick preliminary analysis, an approximate prewhitening can be done easily by first differencing the data (if needed) and then fitting an approximate AR model with the order determined by minimizing the AIC. For example, for the milk production and electricity consumption data, both are highly seasonal and contain trends. Consequently, they can be differenced with both regular differencing and seasonal differencing, and then the prewhitening can be carried out by filtering both differenced series by an AR model fitted to the differenced milk data. Exhibit 11.16 shows the sample CCF between the prewhitened series. None of the cross-correlations are now significant except for lag –3, which is just marginally significant. The lone significant cross-correlation is likely a false alarm since we expect about 1.75 false alarms out of the 35 sample cross-correlations examined. Thus, it seems that milk production and electricity consumption are in fact largely uncorrelated, and the strong cross-correlation pattern found between the raw data series is indeed spurious.

**Exhibit 11.16 Sample CCF of Prewhitened Milk and Electricity Production**

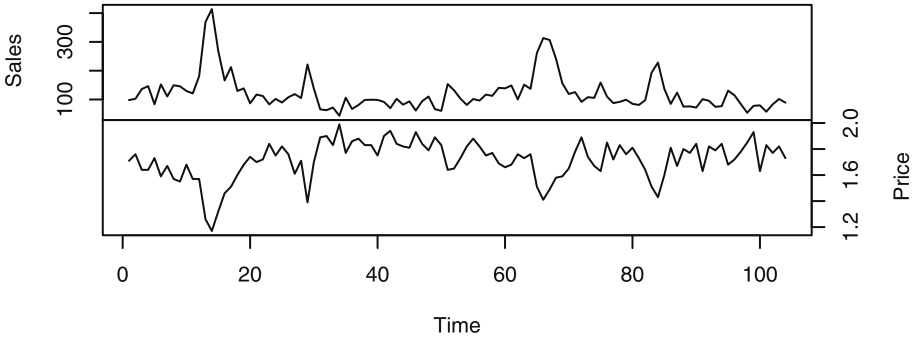


```
> me.dif=ts.intersect(diff(diff(milk,12)),
   diff(diff(log(electricity),12)))
> prewhiten(as.vector(me.dif[,1]),as.vector(me.dif[,2]),
   ylab='CCF')
```

The model defined by Equation (11.3.4) on page 262 is known variously as the **transfer-function model**, the distributed-lag model, or the dynamic regression model. The specification of which lags of the covariate enter into the model is often done by inspecting the sample cross-correlation function based on the prewhitened data. When the model appears to require a fair number of lags of the covariate, the regression coefficients may be parsimoniously specified via an ARMA specification similar to the case of intervention analysis; see Box et al. (1994, Chapter 11) for some details. We illustrate the method below with two examples where only one lag of the covariate appears to be needed. The specification of the stochastic noise process $Z_t$ can be done by examining the residuals from an ordinary least squares (OLS) fit of $Y$ on $X$ using the techniques learned in earlier chapters.

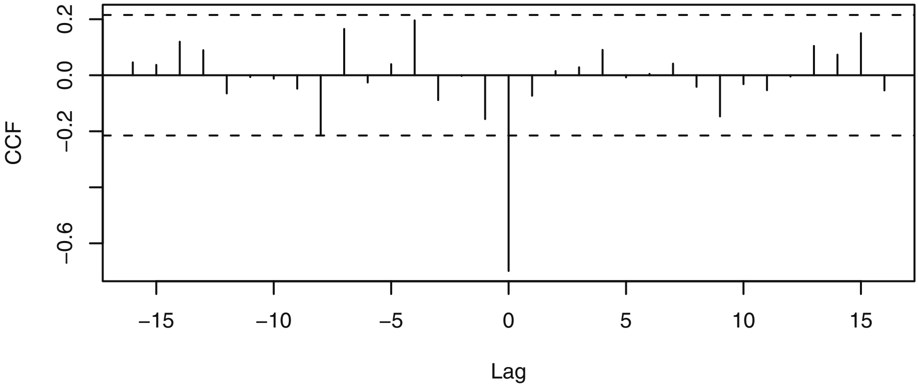
Our first example of this section is a sales and price dataset of a certain potato chip from Bluebird Foods Ltd., New Zealand. The data consist of the log-transformed weekly unit sales of large packages of standard potato chips sold and the weekly average price over a period of 104 weeks from September 20, 1998 through September 10, 2000; see Exhibit 11.17. The logarithmic transformation is needed because the sales data are highly skewed to the right. These data are clearly nonstationary. Exhibit 11.18 shows that, after differencing and using prewhitened data, the CCF is significant only at lag 0, suggesting a strong contemporaneous negative relationship between lag 1 of price and sales. Higher prices are associated with lower sales.

**Exhibit 11.17 Weekly Log(Sales) and Price for Bluebird Potato Chips**



```
> data(bluebird)
> plot(bluebird,yax.flip=T)
```

**Exhibit 11.18 Sample Cross Correlation Between Prewhitened Differenced Log(Sales) and Price of Bluebird Potato Chips**



```
> prewhiten(y=diff(bluebird)[,1],x=diff(bluebird)[,2],ylab='CCF')
```

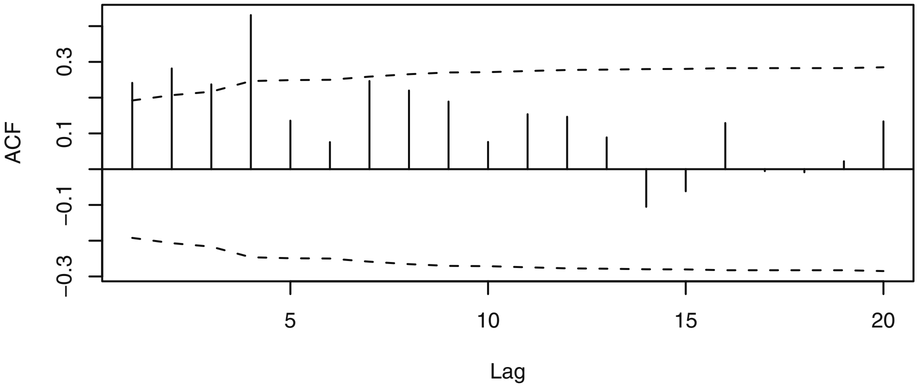
Exhibit 11.19 reports the estimates from the OLS regression of log(sales) on price. The residuals are, however, autocorrelated, as can be seen from their sample ACF and PACF displayed in Exhibits 11.20 and 11.21, respectively. Indeed, the sample autocorrelations of the residuals are significant for the first four lags, whereas the sample partial autocorrelations are significant at lags 1, 2, 4, and 14.

**Exhibit 11.19 OLS Regression Estimates of Log(Sales) on Price**

| | Estimate | Std. Error | *t* value | *Pr*(>) |
|---|---|---|---|---|
| Intercept | 15.90 | 0.2170 | 73.22 | < 0.0001 |
| Price | –2.489 | 0.1260 | –19.75 | < 0.0001 |

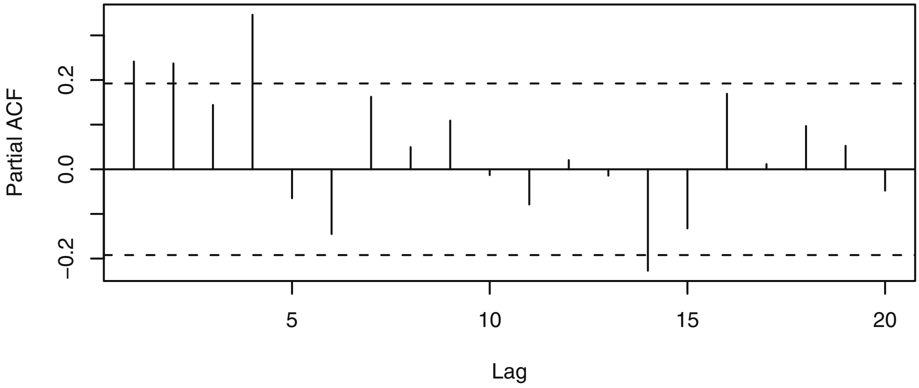
```
> sales=bluebird[,1]; price=bluebird[,2]
> chip.m1=lm(sales~price,data=bluebird)
> summary(chip.m1)
```

**Exhibit 11.20 Sample ACF of Residuals from OLS Regression of Log(Sales) on Price**



```
> acf(residuals(chip.m1),ci.type='ma')
```

**Exhibit 11.21 Sample PACF of Residuals from OLS Regression of Log(Sales) on Price**



```
> pacf(residuals(chip.m1))
```

The sample EACF of the residuals, shown in Exhibit 11.22, contains a triangle of zeros with a vertex at (1,4), thereby suggesting an ARMA(1,4) model. Hence, we fit a regression model of log(sales) on price with an ARMA(1,4) error.

**Exhibit 11.22 The Sample EACF of the Residuals from the OLS Regression of Log(Sales) on Price**

| AR/MA | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | x | x | x | x | 0 | 0 | x | x | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | x | 0 | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | x | x | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | x | x | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | x | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **5** | x | x | x | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **6** | x | x | 0 | x | x | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **7** | x | 0 | x | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

```
> eacf(residuals(chip.m1))
```

It turns out that the estimates of the AR(1) coefficient and the MA coefficients $\theta_1$ and $\theta_3$ are not significant, and hence a model fixing these coefficients to be zero was subsequently fitted and reported in Exhibit 11.23.

**Exhibit 11.23 Maximum Likelihood Estimates of a Regression Model of Log(sales) on Price with a Subset MA(4) for the Errors**

| Parameter | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | Intercept | Price |
|---|---|---|---|---|---|---|
| Estimate | 0 | –0.2884 | 0 | –0.5416 | 15.86 | –2.468 |
| Standard Error | 0 | 0.0794 | 0 | 0 0.1167 | 0.1909 | 0.1100 |

$\sigma^2$ estimated as 0.02623: log likelihood = 41.02, AIC = –70.05

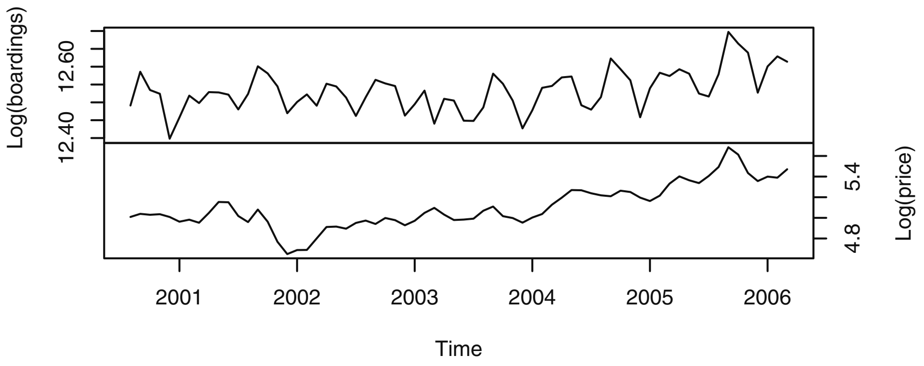
```
> chip.m2=arima(sales,order=c(1,0,4),xreg=data.frame(price))
> chip.m2
> chip.m3=arima(sales,order=c(1,0,4),xreg=data.frame(price),
    fixed=c(NA,0,NA,0,NA,NA,NA)); chip.m3
> chip.m4=arima(sales,order=c(0,0,4),xreg=data.frame(price),
    fixed=c(0,NA,0,NA,NA,NA)); chip.m4
```

Note that the regression coefficient estimate on Price is similar to that from the OLS regression fit earlier, but the standard error of the estimate is about 10% lower than that from the simple OLS regression. This illustrates the general result that the simple OLS estimator is consistent but the associated standard error is generally not trustworthy.

The residuals from this fitted model by and large pass various model diagnostic tests except that the residual ACF is significant at lag 14. As a result, some Box-Ljung test statistics have *p*-values bordering on 0.05 when 14 or more lags of the residual autocorrelations are included in the test. Even though the significant ACF at lag 14 may suggest a quarterly effect, we do not report a more complex model including lag 14 because (1) 14 weeks do not exactly make a quarter and (2) adding a seasonal MA(1) component of period 14 only results in marginal improvement in terms of model diagnostics.

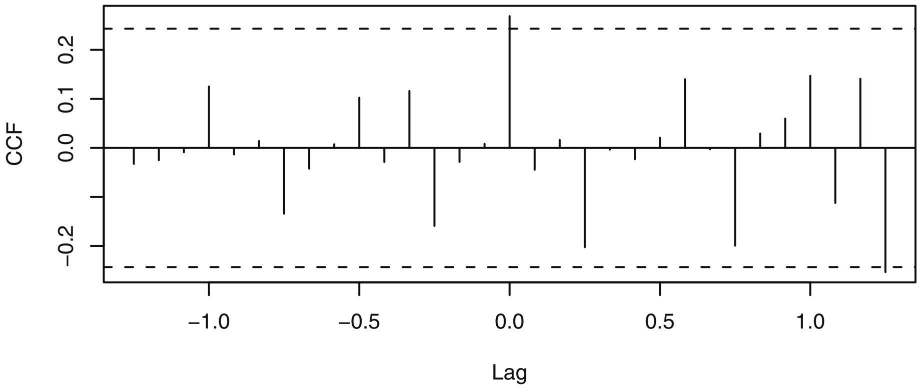
For a second example, we study the impact of higher gasoline price on public transportation usage. The dataset consists of the monthly number of boardings on public transportation in the Denver, Colorado, region together with the average monthly gasoline prices in Denver from August 2000 through March 2006. Both variables are skewed to the right and hence are log-transformed. As we shall see below, the logarithmic transformation also makes the final fitted model more interpretable. The time series plots, shown in Exhibit 11.24, display the increasing trends for both variables and the seasonal fluctuation in the number of boardings. Based on the sample ACF and PACF, an ARIMA(2,1,0) model was fitted to the gasoline price data. This fitted model was then used to filter the boardings data before computing their sample CCF which is shown in Exhibit 11.25. The sample CCF is significant at lags 0 and 15, suggesting positive contemporaneous correlation between gasoline price and public transportation usage. The significant CCF at lag 15, however, is unlikely to be real, as it is hard to imagine why the number of boardings might lead the gasoline price with a lag of 15 months. In this case, the quick preliminary approach of prewhitening the series by fitting a long AR model, however, showed that none of the CCFs are significant. It turns out that even after differencing the data, the AIC selects an AR(16) model. The higher order selected coupled with the relatively short time span may substantially weaken the power to detect correlations between the two variables. Incidentally, this example warns against simply relying on the AIC to select a high-order AR model to do prewhitening, especially with relatively short time series data.

**Exhibit 11.24 Logarithms of Monthly Public Transit Boardings and Gasoline Prices in Denver, August 2000 through March 2006**



```
> data(boardings)
> plot(boardings,yax.flip=T)
```

**Exhibit 11.25 Sample CCF of Prewhitened Log(Boardings) and Log(Price)**



```
> m1=arima(boardings[,2],order=c(2,1,0))
> prewhiten(x=boardings[,2],y=boardings[,1],x.model=m1)
```

Based on the sample ACF, PACF, and EACF of the residuals from a linear model of boardings on gasoline price, a seasonal ARIMA(2,0,0)×$(1,0,0)_{12}$ model was tentatively specified for the error process in the regression model. However, the $\phi_2$ coefficient estimate was not significant, and hence the AR order was reduced to $p = 1$. Using the outlier detection techniques discussed in Section 11.2, we found an additive outlier for March 2003 and an innovative outlier for March 2004. Because the test statistic for the additive outlier had a larger magnitude than that of the innovative outlier (–4.09 vs. 3.65), we incorporated the additive outlier in the model.† Diagnostics of the subsequent fitted model reveals that the residual ACF was significant at lag 3, which suggests the error process is a seasonal ARIMA(1,0,3)×$(1,0,0)_{12}$ + outlier process. As the estimates of the coefficients $\theta_1$ and $\theta_2$ were found to be insignificant, they were suppressed from the final fitted model that is reported in Exhibit 11.26.

Diagnostics of the final fitted model suggest a good fit to the data. Also, no further outliers were detected. A 95% confidence interval for the regression coefficient on Log(Price) is (0.0249, 0.139). Note the interpretation of the fitted model: a 100% increase in the price of gasoline will lead to about an 8.2% increase in public transportation usage.

† Subsequent investigation revealed that a 30 inch snowstorm in March 2003 completely shut down Denver for one full day. It remained partially shut down for a few more days.

**Exhibit 11.26 Maximum Likelihood Estimates of the Regression Model of Log(Boardings) on Log(Price) with ARMA Errors**

| Parameter | $\phi_1$ | $\theta_3$ | $\Phi_1$ | Intercept | Log(Price) | Outlier |
|---|---|---|---|---|---|---|
| Estimate | 0.8782 | 0.3836 | 0.8987 | 12.12 | 0.0819 | –0.0643 |
| Standard Error | 0.0645 | 0.1475 | 0.0395 | 0.1638 | 0.0291 | 0.0109 |

$\sigma^2$ estimated as 0.0004094: log-likelihood = 158.02, AIC = –304.05

```
> log.boardings=boardings[,1]
> log.price=boardings[,2]
> boardings.m1=arima(log.boardings,order=c(1,0,0),
    seasonal=list(order=c(1,0,0),period=12),
    xreg=data.frame(log.price))
> boardings.m1
> detectAO(boardings.m1); detectIO(boardings.m1)
> boardings.m2=arima(log.boardings,order=c(1,0,3),
    seasonal=list(order=c(1,0,0),period=12),
    xreg=data.frame(log.price,outlier=c(rep(0,31),1,rep(0,36))),
    fixed=c(NA,0,0,rep(NA,5)))
> boardings.m2
> detectAO(boardings.m2); detectIO(boardings.m2)
> tsdiag(boardings.m2,tol=.15,gof.lag=24)
```

It is also of interest to note that dropping the outlier term from the model results in a new regression estimate on Log(Price) of 0.0619 with a standard error of 0.0372. Thus, when the outlier is not properly modeled, the regression coefficient ceases to be significant at the 5% level. As demonstrated by this example, the presence of an outlier can adversely affect inference in time series modeling.

## 11.5 Summary

In this chapter, we used information from other events or other time series to help model the time series of main interest. We began with the so-called intervention models, which attempt to incorporate known external events that we believe have a significant effect on the time series of interest. Various simple but useful ways of modeling the effects of these interventions were discussed. Outliers are observations that deviate rather substantially from the general pattern of the data. Models were developed to detect and incorporate outliers in time series. The material in the section on spurious correlation illustrates how difficult it is to assess relationships between two time series, but methods involving prewhitening were shown to help in this regard. Several substantial examples were used to illustrate the methods and techniques discussed.

## Exercises

**11.1** Produce a time series plot of the air passenger miles over the period January 1996 through May 2005 using seasonal plotting symbols. Display the graph full-screen and discuss the seasonality that is displayed. The data are in the file named airmiles.

**11.2** Show that the expression given for $m_t$ in Equation (11.1.7) on page 251 satisfies the "AR(1)" recursion given in Equation (11.1.6) with the initial condition $m_0 = 0$.

**11.3** Find the "half-life" for the intervention effect specified in Equation (11.1.6) on page 251 when $\delta = 0.7$.

**11.4** Show that the "half-life" for the intervention effect specified in Equation (11.1.6) on page 251 increases without bound as $\delta$ increases to 1.

**11.5** Show that for the intervention effect specified by Equation (11.1.6) on page 251

$$\lim_{\delta \to 1} m_t = \begin{cases} \omega(T-t), \text{ for } t \geq T \\ 0, \text{ otherwise} \end{cases}$$

**11.6** Consider the intervention effect displayed in Exhibit 11.3, (b), page 253.

**(a)** Show that the jump at time $T + 1$ is of height $\omega$ as displayed.

**(b)** Show that, as displayed, the intervention effect tends to $\omega/(1 - \delta)$ as $t$ increases without bound.

**11.7** Consider the intervention effect displayed in Exhibit 11.3, (c), page 253. Show that the effect increases linearly starting at time $T + 1$ with slope $\omega$ as displayed.

**11.8** Consider the intervention effect displayed in Exhibit 11.4, (a), page 254.

**(a)** Show that the jump at time $T + 1$ is of height $\omega$ as displayed.

**(b)** Show that, as displayed, the intervention effect tends to go back to 0 as $t$ increases without bound.

**11.9** Consider the intervention effect displayed in Exhibit 11.4, (b), page 254.

**(a)** Show that the jump at time $T + 1$ is of height $\omega_1 + \omega_2$ as displayed.

**(b)** Show that, as displayed, the intervention effect tends to $\omega_2$ as $t$ increases without bound.

**11.10** Consider the intervention effect displayed in Exhibit 11.4, (c), page 254.

**(a)** Show that the jump at time $T$ is of height $\omega_0$ as displayed.

**(a)** Show that the jump at time $T + 1$ is of height $\omega_1 + \omega_2$ as displayed.

**(b)** Show that, as displayed, the intervention effect tends to $\omega_2$ as $t$ increases without bound.

**11.11** Simulate 100 pairs of $(X_t, Y_t)$ from the model of Equation (11.3.1) on page 261 with $d = 3$, $\beta_0 = 0$, and $\beta_1 = 1$. Use $\sigma_X = 2$ and $\sigma_e = 1$. Display and interpret the sample CCF between these two series.

**11.12** Show that when the $X$ and $Y$ are independent AR(1) time series with parameters $\phi_X$ and $\phi_Y$, respectively, Equation (11.3.5) on page 262 reduces to give Equation (11.3.6).

**11.13** Show that for the process defined by Equation (11.4.5) on page 266, the cross-correlation between $\tilde{X}$ and $\tilde{Y}$ at lag $k$ is given by $\beta_{-k}(\sigma_{\tilde{X}}/\sigma_{\tilde{Y}})$.

**11.14** Simulate an AR time series with $\phi = 0.7$, $\mu = 0$, $\sigma_e = 1$, and of length $n = 48$. Plot the time series, and inspect the sample ACF and PACF of the series.

**(a)** Now add a step function response of $\omega = 1$ unit height at time $t = 36$ to the simulated series. The series now has a theoretical mean of zero from $t = 1$ to 35 and a mean of 1 from $t = 36$ on. Plot the new time series and calculate the sample ACF and PACF for the new series. Compare these with the results for the original series.

**(b)** Repeat part (a) but with an impulse response at time $t = 36$ of unit height, $\omega = 1$. Plot the new time series, and calculate the sample ACF and PACF for the new series. Compare these with the results for the original series. See if you can detect the additive outlier at time $t = 36$ assuming that you do not know where the outlier might occur.

**11.15** Consider the air passenger miles time series discussed in this chapter. The file is named airmiles. Use only the *preintervention* data (that is, data prior to September 2001) for this exercise.

**(a)** Verify that the sample ACF for the twice differenced series of the logarithms of the preintervention data is as shown in Exhibit 11.5 on page 255.

**(b)** The plot created in part (a) suggests an ARIMA$(0,1,1)\times(0,1,0)_{12}$. Fit this model and assess its adequacy. In particular, verify that additive outliers are detected in December 1996, January 1997, and December 2002.

**(c)** Now fit an ARIMA$(0,1,1)\times(0,1,0)_{12}$ + three outliers model and assess its adequacy.

**(d)** Finally, fit an ARIMA$(0,1,1)\times(0,1,1)_{12}$ + three outliers model and assess its adequacy.

**11.16** Use the logarithms of the Denver region public transportation boardings and Denver gasoline price series. The data are in the file named boardings.

**(a)** Display the time series plot of the monthly boardings using seasonal plotting symbols. Interpret the plot.

**(b)** Display the time series plot of the monthly average gasoline prices using seasonal plotting symbols. Interpret the plot.

**11.17** The data file named deere1 contains 82 consecutive values for the amount of deviation (in 0.000025 inch units) from a specified target value that an industrial machining process at Deere & Co. produced under certain specified operating conditions. These data were first used in Exercise 6.33, page 146, where we observed an obvious outlier at time $t = 27$.

**(a)** Fit an AR(2) model using the original data including the outlier.

**(b)** Test the fitted AR(2) model of part (a) for both AO and IO outliers.

**(c)** Now fit the AR(2) model incorporating a term in the model for the outlier.

**(d)** Assess the fit of the model in part (c) using all of our diagnostic tools. In particular, compare the properties of this model with the one obtained in part (a).

**11.18** The data file named days contains accounting data from the Winegard Co. of Burlington, Iowa. The data are the number of days until Winegard receives payment for 130 consecutive orders from a particular distributor of Winegard products. (The name of the distributor must remain anonymous for confidentiality reasons.) These data were first investigated in Exercise 6.39, page 147, but several outliers were observed. When the observed outliers were replaced by more typical values, an MA(2) model was suggested.

**(a)** Fit an MA(2) model to the original data, and test the fitted model for both AO and IO outliers.

**(b)** Now fit the MA(2) model incorporating the outliers into the model.

**(c)** Assess the fit of the model obtained in part (b). In particular, are any more outliers indicated?

**(d)** Fit another MA(2) model incorporating any additional outliers found in part (c), and assess the fit of this model.

**11.19** The data file named bluebirdlite contains weekly sales and price data for Bluebird Lite potato chips. Carry out an analysis similar to that for Bluebird Standard potato chips that was begun on page 267.

**11.20** The file named units contains annual unit sales of a certain product from a widely known international company over the years 1983 through 2005. (The name of the company must remain anonymous for proprietary reasons.)

**(a)** Plot the time series of units and describe the general features of the plot.

**(b)** Use ordinary least squares regression to fit a straight line in time to the series.

**(c)** Display the sample PACF of the residuals from this model, and specify an ARIMA model for the residuals.

**(d)** Now fit the model unit sales = AR(2) + time. Interpret the output. In particular, compare the estimated regression coefficient on the time variable obtained here with the one you obtained in part (b).

**(e)** Perform a thorough analysis of the residuals from this last model.

**(f)** Repeat parts (d) and (e) using the logarithms of unit sales as the response variable. Compare these results witjh those obtained in parts (d) and (e).

**11.21** In Chapters 5–8, we investigated an IMA(1,1) model for the logarithms of monthly oil prices. Exhibit 8.3 on page 178 suggested that there may be several outliers in this series. Investigate the IMA(1,1) model for this series for outliers using the techniques developed in this chapter. Be sure to compare your results with those obtained earlier that ignored the outliers. The data are in the file named oil.