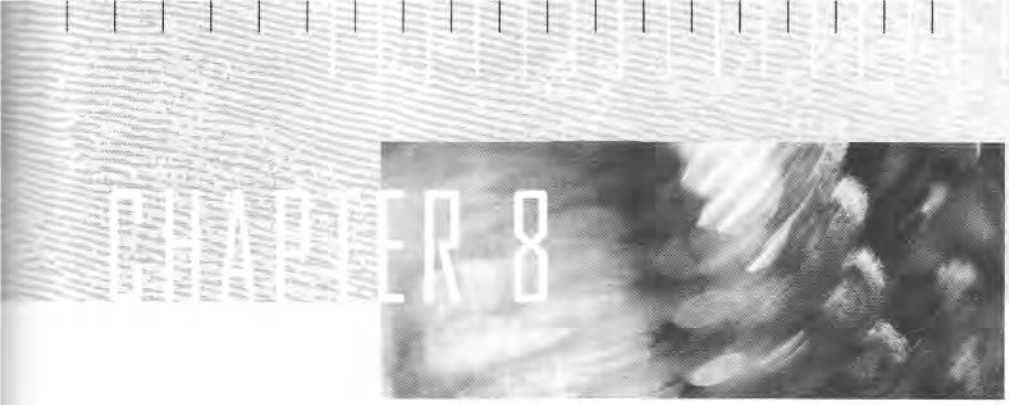



# *Modeling Cycles: MA, AR, and ARMA Models*

When building forecasting models, we don't want to pretend that the model we fit is true. Instead, we want to be aware that we're *approximating* a more complex reality. That's the modern view, and it has important implications for forecasting. In particular, we've seen that the key to successful time series modeling and forecasting is parsimonious, yet accurate, approximation of the Wold representation. In this chapter, we consider three approximations: **moving average (MA) models, autoregressive (AR) models,** and **autoregressive moving average (ARMA) models.** The three models vary in their specifics and have different strengths in capturing different sorts of autocorrelation behavior.

We begin by characterizing the autocorrelation functions and related quantities associated with each model, under the assumption that the model is "true." We do this separately for MA, AR, and ARMA models.[1] These characterizations have nothing to do with data or estimation, but they're crucial for developing a basic understanding of the properties of the models, which is necessary to perform intelligent modeling and forecasting. They enable us to

[1] Sometimes, especially when characterizing population properties under the assumption that the models are correct, we refer to them as *processes*, which is short for **stochastic processes**—hence the terms *moving average process, autoregressive process,* and *ARMA process.*

make statements such as "If the data were really generated by an autoregressive process, then we'd expect its autocorrelation function to have property *x*." Armed with that knowledge, we use the *sample* autocorrelations and partial autocorrelations, in conjunction with the AIC and the SIC, to suggest candidate forecasting models, which we then estimate.

# 1. Moving Average (MA) Models

The finite-order moving average process is a natural and obvious approximation to the Wold representation, which is an infinite-order moving average process. Finite-order moving average processes also have direct motivation. The fact that all variation in time series, one way or another, is driven by shocks of various sorts suggests the possibility of modeling time series directly as distributed lags of current and past shocks—that is, as moving average processes.[2]

## The MA(1) Process

The first-order moving average process, or **MA(1) process,** is

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1} = (1 + \theta L)\varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The defining characteristic of the MA process in general, and the MA(1) in particular, is that the current value of the observed series is expressed as a function of current and lagged unobservable shocks. Think of it as a regression model with nothing but current and lagged disturbances on the right-hand side.

To help develop a feel for the behavior of the MA(1) process, we show two simulated realizations of length 150 in Figure 8.1. The processes are

$$y_t = \varepsilon_t + 0.4\varepsilon_{t-1}$$

and

$$y_t = \varepsilon_t + 0.95\varepsilon_{t-1},$$

where in each case $\varepsilon_t \overset{iid}{\sim} N(0, 1)$. To construct the realizations, we used the same series of underlying white noise shocks; the only difference in the realizations comes from the different coefficients. Past shocks feed *positively* into the current value of the series, with a small weight of $\theta = 0.4$ in one case and a large weight of $\theta = 0.95$ in the other. You might think that $\theta = 0.95$ would induce much more persistence than $\theta = 0.4$, but it doesn't. The structure of the

[2] Economic equilibria, for example, may be disturbed by shocks that take some time to be fully assimilated.

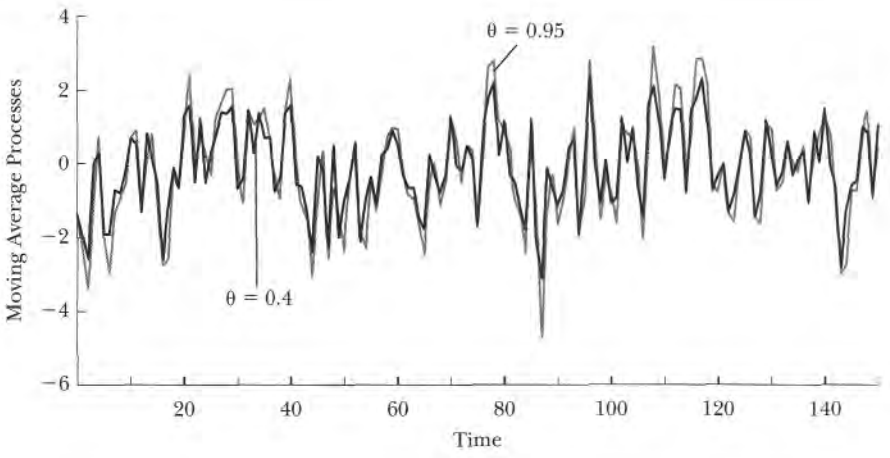



FIGURE 8.1
*Realizations of Two MA(1) Processes*

MA(1) process, in which only the first lag of the shock appears on the right, forces it to have a very short memory, and hence weak dynamics, regardless of the parameter value.

The unconditional mean and variance are

$$E(y_t) = E(\varepsilon_t) + \theta E(\varepsilon_{t-1}) = 0$$

and

$$\text{var}(y_t) = \text{var}(\varepsilon_t) + \theta^2 \text{var}(\varepsilon_{t-1}) = \sigma^2 + \theta^2\sigma^2 = \sigma^2(1 + \theta^2).$$

Note that for a fixed value of $\sigma$, as $\theta$ increases in absolute value, so, too, does the unconditional variance. That's why the MA(1) process with parameter $\theta = 0.95$ varies a bit more than the process with a parameter of $\theta = 0.4$.

The conditional mean and variance of an MA(1), where the conditioning information set is $\Omega_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$, are

$$E(y_t \mid \Omega_{t-1}) = E(\varepsilon_t + \theta\varepsilon_{t-1} \mid \Omega_{t-1}) = E(\varepsilon_t \mid \Omega_{t-1}) + \theta E(\varepsilon_{t-1} \mid \Omega_{t-1}) = \theta\varepsilon_{t-1}$$

and

$$\text{var}(y_t \mid \Omega_{t-1}) = E((y_t - E(y_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = E(\varepsilon_t^2 \mid \Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2.$$

The conditional mean explicitly adapts to the information set, in contrast to the unconditional mean, which is constant. Note, however, that only the first lag of the shock enters the conditional mean—more distant shocks have no effect on the current conditional expectation. This is indicative of the one-period memory of MA(1) processes, which we'll now characterize in terms of the autocorrelation function.

To compute the autocorrelation function for the MA(1) process, we must first compute the autocovariance function. We have

$$\gamma(\tau) = E(y_t y_{t-\tau}) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) = \begin{cases} \theta\sigma^2, & \tau = 1 \\ 0, & \text{otherwise} \end{cases}.$$

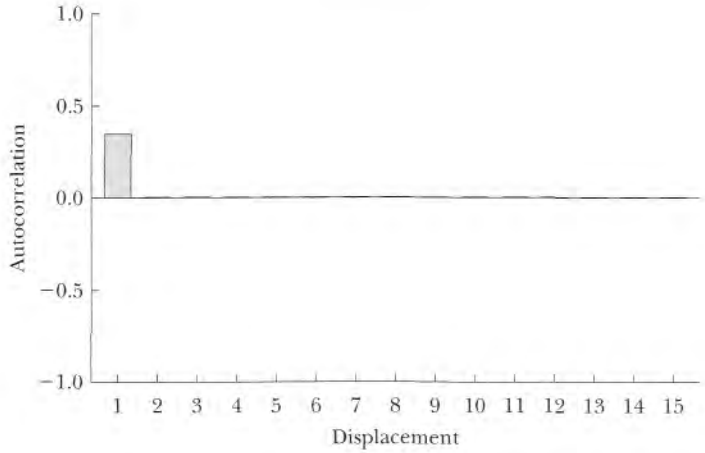



FIGURE 8.2
*Population Autocorrelation Function, MA(1) Process,* $\theta = 0.4$

(The proof is left as a problem.) The autocorrelation function is just the autocovariance function scaled by the variance,

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \begin{cases} \dfrac{\theta}{1+\theta^2}, & \tau = 1 \\ 0, & \text{otherwise} . \end{cases}$$

The key feature here is the sharp **cutoff in the autocorrelation function.** All autocorrelations are 0 beyond displacement 1, the order of the MA process. In Figures 8.2 and 8.3, we show the autocorrelation functions for our two MA(1) processes with parameters $\theta = 0.4$ and $\theta = 0.95$. At displacement 1, the process with parameter $\theta = 0.4$ has a smaller autocorrelation (0.34)

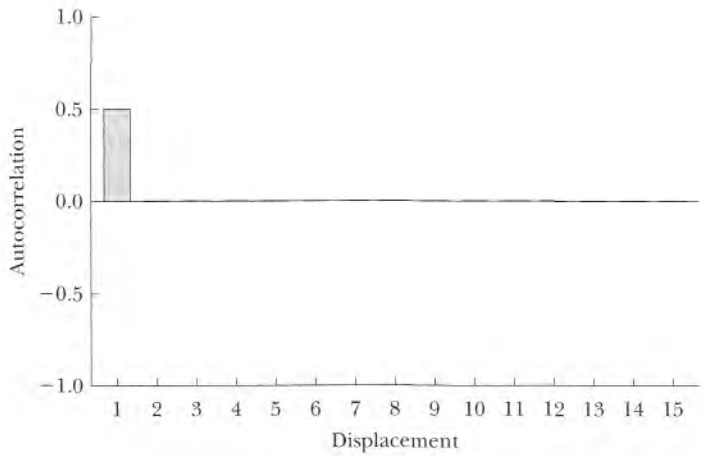



FIGURE 8.3
*Population Autocorrelation Function, MA(1) Process,* $\theta = 0.95$

than the process with parameter $\theta = 0.95$ (0.50), but both drop to 0 beyond displacement 1.

Note that the requirements of covariance stationarity (constant unconditional mean, constant and finite unconditional variance, autocorrelation dependent only on displacement) are met for any MA(1) process, *regardless* of the values of its parameters. If, moreover, $|\theta| < 1$, then we say that the MA(1) process is **invertible.** In that case, we can "invert" the MA(1) process and express the current value of the series not in terms of a current shock and a lagged shock but rather in terms of a current shock *and lagged values of the series.* That's called an **autoregressive representation.** An autoregressive representation has a current shock and lagged observable values of the series on the right, whereas a moving average representation has a current shock and lagged unobservable shocks on the right.

Let's compute the autoregressive representation. The process is

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0, \sigma^2) .$$

Thus, we can solve for the innovation as

$$\varepsilon_t = y_t - \theta\varepsilon_{t-1} .$$

Lagging by successively more periods gives expressions for the innovations at various dates,

$$\varepsilon_{t-1} = y_{t-1} - \theta\varepsilon_{t-2}$$

$$\varepsilon_{t-2} = y_{t-2} - \theta\varepsilon_{t-3}$$

$$\varepsilon_{t-3} = y_{t-3} - \theta\varepsilon_{t-4} ,$$

and so forth. Making use of these expressions for lagged innovations, we can substitute backward in the MA(1) process, yielding

$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - \cdots .$$

In lag operator notation, we write the infinite autoregressive representation as

$$\frac{1}{1 + \theta L} y_t = \varepsilon_t .$$

Note that the back substitution used to obtain the autoregressive representation only makes sense, and in fact a convergent autoregressive representation only exists, if $|\theta| < 1$, because in the back substitution we raise $\theta$ to progressively higher powers.

We can restate the invertibility condition in another way: The inverse of the root of the moving average lag operator polynomial $(1 + \theta L)$ must be less than 1 in absolute value. Recall that a polynomial of degree $m$ has $m$ roots. Thus, the MA(1) lag operator polynomial has one root, which is the solution to

$$1 + \theta L = 0 .$$

The root is $L = -1/\theta$, so its inverse will be less than 1 in absolute value if $|\theta| < 1$, and the two invertibility conditions are equivalent. The "inverse root" way of stating invertibility conditions seems tedious, but it turns out to be of greater applicability than the $|\theta| < 1$ condition, as we'll see shortly.

Autoregressive representations are appealing to forecasters, because one way or another, if a model is to be used for real-world forecasting, it must link the present observables to the past history of observables, so that we can extrapolate to form a forecast of future observables based on present and past observables. Superficially, moving average models don't seem to meet that requirement, because the current value of a series is expressed in terms of current and lagged unobservable shocks, not observable variables. But under the invertibility conditions that we've described, moving average processes have equivalent autoregressive representations. Thus, although we want autoregressive representations for forecasting, we don't have to start with an autoregressive model. However, we typically restrict ourselves to invertible processes, because for forecasting purposes we want to be able to express current observables as functions of past observables.

Finally, let's consider the partial autocorrelation function for the MA(1) process. From the infinite autoregressive representation of the MA(1) process, we see that the partial autocorrelation function will decay gradually to 0. As we discussed in Chapter 7, the partial autocorrelations are just the coefficients on the last included lag in a sequence of progressively higher-order autoregressive approximations. If $\theta > 0$, then the pattern of decay will be one of damped oscillation; otherwise, the decay will be one-sided.

In Figures 8.4 and 8.5 we show the partial autocorrelation functions for our example MA(1) processes. For each process, $|\theta| < 1$, so that an autoregressive representation exists, and $\theta > 0$, so that the coefficients in the autoregressive

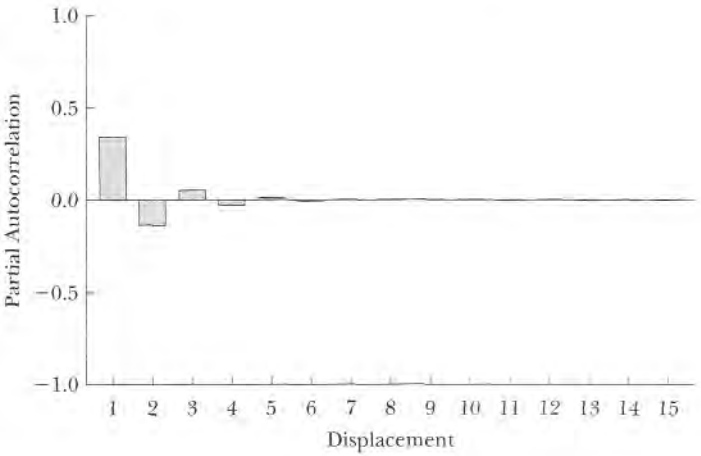



FIGURE 8.4
*Population Partial Autocorrelation Function, MA(1) Process,* $\theta = 0.4$

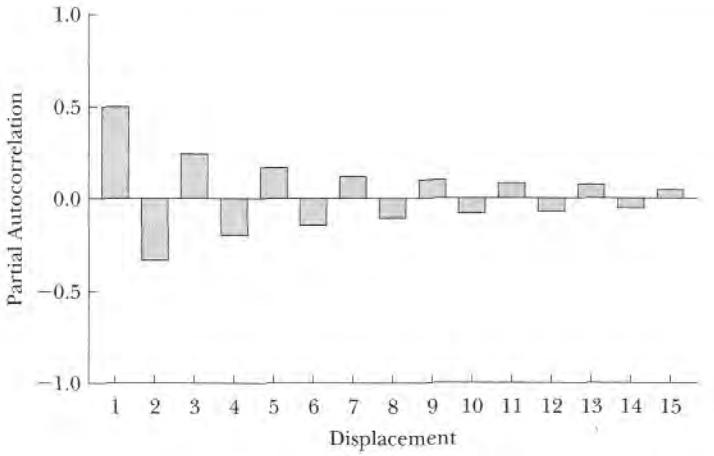



FIGURE 8.5
*Population Partial Autocorrelation Function, MA(1) Process, $\theta = 0.95$*

representations alternate in sign. Specifically, we showed the general autoregressive representation to be

$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - \cdots ,$$

so the autoregressive representation for the process with $\theta = 0.4$ is

$$y_t = \varepsilon_t + 0.4 y_{t-1} - 0.4^2 y_{t-2} + \cdots = \varepsilon_t + 0.4 y_{t-1} - 0.16 y_{t-2} + \cdots ,$$

and the autoregressive representation for the process with $\theta = 0.95$ is

$$y_t = \varepsilon_t + 0.95 y_{t-1} - 0.95^2 y_{t-2} + \cdots = \varepsilon_t + 0.95 y_{t-1} - 0.9025 y_{t-2} + \cdots .$$

The partial autocorrelations display a similar damped oscillation.[3] The decay, however, is slower for the $\theta = 0.95$ case.

## The MA($q$) Process

Now consider the general finite-order moving average process of order $q$, or MA($q$) for short,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} = \Theta(L)\varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2) ,$$

where

$$\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

is a $q$th-order lag operator polynomial. The **MA($q$) process** is a natural generalization of the MA(1). By allowing for more lags of the shock on the right side

[3] Note, however, that the partial autocorrelations are *not* the successive coefficients in the infinite autoregressive representation. Rather, they are the coefficients on the last included lag in sequence of progressively longer autoregressions. The two are related but distinct.

of the equation, the MA($q$) process can capture richer dynamic patterns, which we can potentially exploit for improved forecasting. The MA(1) process is of course a special case of the MA($q$), corresponding to $q = 1$.

The properties of the MA($q$) processes parallel those of the MA(1) process in all respects, so in what follows we'll refrain from grinding through the mathematical derivations. Instead, we'll focus on the key features of practical importance. Just as the MA(1) process was covariance stationary for any value of its parameters, so, too, is the finite-order MA($q$) process. As with the MA(1) process, the MA($q$) process is *invertible* only if a root condition is satisfied. The MA($q$) lag operator polynomial has $q$ roots; when $q > 1$, the possibility of **complex roots** arises. The **condition for invertibility of the MA($q$) process** is that the inverses of all of the roots must be inside the unit circle, in which case we have the convergent autoregressive representation,

$$\frac{1}{\Theta(L)} y_t = \varepsilon_t .$$

The conditional mean of the MA($q$) process evolves with the information set, in contrast to the unconditional moments, which are fixed. In contrast to the MA(1) case, in which the conditional mean depends on only the first lag of the innovation, in the MA($q$) case the conditional mean depends on $q$ lags of the innovation. Thus, the MA($q$) process has the potential for longer memory.

The potentially longer memory of the MA($q$) process emerges clearly in its autocorrelation function. In the MA(1) case, all autocorrelations beyond displacement 1 are 0; in the MA($q$) case, all autocorrelations beyond displacement $q$ are 0. This autocorrelation cutoff is a distinctive property of moving average processes. The partial autocorrelation function of the MA($q$) process, in contrast, decays gradually, in accord with the infinite autoregressive representation, in either an oscillating or a one-sided fashion, depending on the parameters of the process.

In closing this section, let's step back for a moment and consider in greater detail the precise way in which finite-order moving average processes approximate the Wold representation. The Wold representation is

$$y_t = B(L)\varepsilon_t ,$$

where $B(L)$ is of infinite order. The MA(1), in contrast, is simply a first-order moving average, in which a series is expressed as a one-period moving average of current and past innovations. Thus, when we fit an MA(1) model, we're using the first-order polynomial $1 + \theta L$ to approximate the infinite-order polynomial $B(L)$. Note that $1 + \theta L$ is a rational polynomial with numerator polynomial of degree 1 and degenerate denominator polynomial (degree 0).

MA($q$) processes have the potential to deliver better approximations to the Wold representation, at the cost of more parameters to be estimated. The Wold representation involves an infinite moving average; the MA($q$) process approximates the infinite moving average with a *finite-order* moving average,

$$y_t = \Theta(L)\varepsilon_t ,$$

whereas the MA(1) process approximates the infinite moving average with only a *first-order* moving average, which can sometimes be very restrictive.

## 2. Autoregressive (AR) Models

The autoregressive process is also a natural approximation to the Wold representation. We've seen, in fact, that under certain conditions a moving average process has an autoregressive representation, so an autoregressive process is in a sense the same as a moving average process. Like the moving average process, the autoregressive process has direct motivation; it's simply a *stochastic difference equation*, a simple mathematical model in which the current value of a series is linearly related to its past values, plus an additive stochastic shock. Stochastic difference equations are a natural vehicle for discrete-time stochastic dynamic modeling.

### The AR(1) Process

The first-order autoregressive process, AR(1) for short, is

$$y_t = \varphi y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2) .$$

In lag operator form, we write

$$(1 - \varphi L) y_t = \varepsilon_t .$$

In Figure 8.6 we show simulated realizations of length 150 of two AR(1) processes; the first is

$$y_t = 0.4 y_{t-1} + \varepsilon_t ,$$

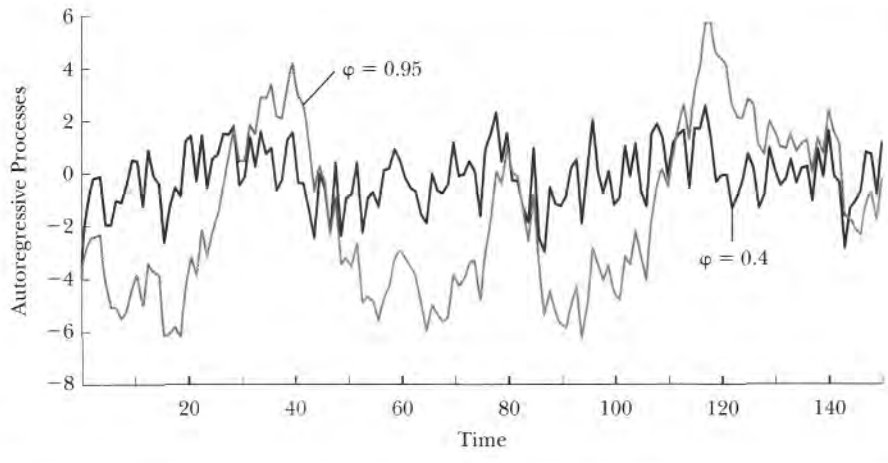



FIGURE 8.6
*Realizations of Two AR(1) Processes*

and the second is

$$y_t = 0.95y_{t-1} + \varepsilon_t,$$

where in each case $\varepsilon_t \overset{iid}{\sim} N(0, 1)$, and the same innovation sequence underlies each realization. The fluctuations in the AR(1) with parameter $\varphi = 0.95$ appear much more persistent that those of the AR(1) with parameter $\varphi = 0.4$. This contrasts sharply with the MA(1) process, which has a very short memory regardless of parameter value. Thus, the AR(1) model is capable of capturing much more persistent dynamics than is the MA(1).

Recall that a finite-order moving average process is always covariance stationary but that certain conditions must be satisfied for invertibility, in which case an autoregressive representation exists. For autoregressive processes, the situation is precisely the reverse. Autoregressive processes are always invertible—in fact, invertibility isn't even an issue, as finite-order autoregressive processes *already are* in autoregressive form—but certain conditions must be satisfied for an autoregressive process to be covariance stationary.

If we begin with the AR(1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t,$$

and substitute backward for lagged $y$'s on the right side, we obtain

$$y_t = \varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots.$$

In lag operator form, we write

$$y_t = \frac{1}{1 - \varphi L}\varepsilon_t.$$

This moving average representation for $y$ is convergent if and only if $|\varphi| < 1$; thus, $|\varphi| < 1$ is the condition for covariance stationarity in the AR(1) case. Equivalently, the condition for covariance stationarity is that the inverse of the root of the autoregressive lag operator polynomial be less than 1 in absolute value.

From the moving average representation of the covariance stationary AR(1) process, we can compute the unconditional mean and variance,

$$\begin{aligned} E(y_t) &= E(\varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots) \\ &= E(\varepsilon_t) + \varphi E(\varepsilon_{t-1}) + \varphi^2 E(\varepsilon_{t-2}) + \cdots \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \text{var}(y_t) &= \text{var}(\varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots) \\ &= \sigma^2 + \varphi^2\sigma^2 + \varphi^4\sigma^2 + \cdots \\ &= \sigma^2 \sum_{i=0}^{\infty} \varphi^{2i} \\ &= \frac{\sigma^2}{1 - \varphi^2}. \end{aligned}$$

The conditional moments, in contrast, are

$$\begin{aligned} E(y_t \mid y_{t-1}) &= E(\varphi y_{t-1} + \varepsilon_t \mid y_{t-1}) \\ &= \varphi E(y_{t-1} \mid y_{t-1}) + E(\varepsilon_t \mid y_{t-1}) \\ &= \varphi y_{t-1} + 0 \\ &= \varphi y_{t-1} \end{aligned}$$

and

$$\begin{aligned} \text{var}(y_t \mid y_{t-1}) &= \text{var}(\varphi y_{t-1} + \varepsilon_t \mid y_{t-1}) \\ &= \varphi^2 \text{var}(y_{t-1} \mid y_{t-1}) + \text{var}(\varepsilon_t \mid y_{t-1}) \\ &= 0 + \sigma^2 \\ &= \sigma^2 . \end{aligned}$$

Note in particular the simple way in which the conditional mean adapts to the changing information set as the process evolves.

To find the autocovariances, we proceed as follows. The process is

$$y_t = \varphi y_{t-1} + \varepsilon_t ,$$

so that, multiplying both sides of the equation by $y_{t-\tau}$ we obtain

$$y_t y_{t-\tau} = \varphi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau} .$$

For $\tau \geq 1$, taking expectations of both sides gives

$$\gamma(\tau) = \varphi \gamma(\tau - 1) .$$

This is called the **Yule-Walker equation.** It is a recursive equation; that is, given $\gamma(\tau)$, for any $\tau$, the Yule-Walker equation immediately tells us how to get $\gamma(\tau + 1)$. If we knew $\gamma(0)$ to start things off (an "initial condition"), we could use the Yule-Walker equation to determine the entire autocovariance sequence. And we *do* know $\gamma(0)$; it's just the variance of the process, which we already showed to be $\gamma(0) = \frac{\sigma^2}{1-\varphi^2}$. Thus, we have

$$\gamma(0) = \frac{\sigma^2}{1 - \varphi^2}$$

$$\gamma(1) = \varphi \frac{\sigma^2}{1 - \varphi^2}$$

$$\gamma(2) = \varphi^2 \frac{\sigma^2}{1 - \varphi^2} ,$$

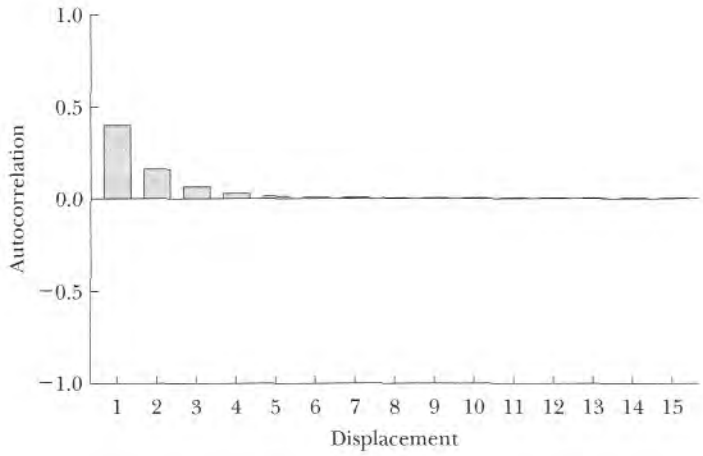
and so on. In general, then,

$$\gamma(\tau) = \varphi^\tau \frac{\sigma^2}{1 - \varphi^2}, \quad \tau = 0, 1, 2, \ldots .$$

Dividing through by $\gamma(0)$ gives the autocorrelations,

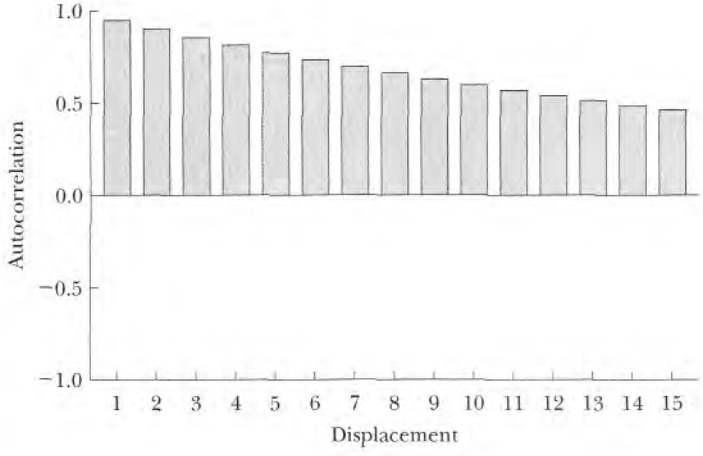
$$\rho(\tau) = \varphi^\tau, \quad \tau = 0, 1, 2, \ldots .$$

FIGURE 8.7
*Population Autocorrelation Function, AR(1) Process, $\varphi = 0.4$*




Note the gradual autocorrelation decay, which is typical of autoregressive processes. The autocorrelations approach 0, but only in the limit as the displacement approaches infinity. In particular, they don't cut off to 0, as is the case for moving average processes. If $\varphi$ is positive, the autocorrelation decay is one-sided. If $\varphi$ is negative, the decay involves back-and-forth oscillations. The relevant case in business and economics is $\varphi > 0$, but either way, the autocorrelations damp gradually, not abruptly. In Figures 8.7 and 8.8, we show the autocorrelation functions for AR(1) processes with parameters $\varphi = 0.4$ and $\varphi = 0.95$. The persistence is much stronger when $\varphi = 0.95$, in contrast to the MA(1) case, in which the persistence was weak regardless of the parameter.

FIGURE 8.8
*Population Autocorrelation Function, AR(1) Process, $\varphi = 0.95$*



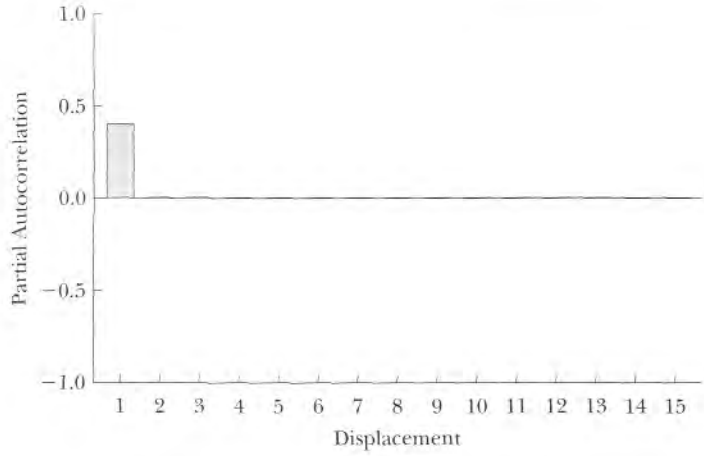



FIGURE 8.9
*Population Partial Autocorrelation Function, AR(1) Process, $\varphi = 0.4$*

Finally, the partial autocorrelation function for the AR(1) process cuts off abruptly; specifically,

$$p(\tau) = \begin{cases} \varphi, & \tau = 1 \\ 0, & \tau > 1 \end{cases}$$

It's easy to see why. The partial autocorrelations are just the last coefficients in a sequence of successively longer population autoregressions. If the true process is in fact an AR(1), the first partial autocorrelation is just the autoregressive coefficient, and coefficients on all longer lags are 0.

In Figures 8.9 and 8.10 we show the partial autocorrelation functions for our two AR(1) processes. At displacement 1, the partial autocorrelations are simply the parameters of the process (0.4 and 0.95, respectively), and at longer displacements, the partial autocorrelations are 0.

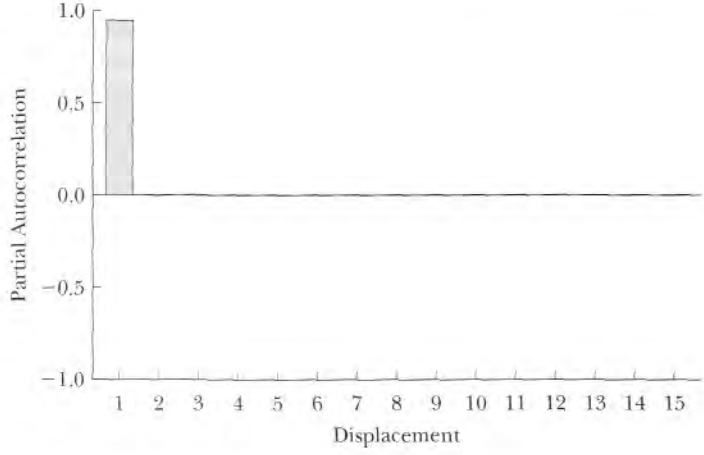



FIGURE 8.10
*Population Partial Autocorrelation Function, AR(1) Process, $\varphi = 0.95$*

## THE AR($p$) PROCESS

The general $p$th order autoregressive process, or AR($p$) for short, is

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2) .$$

In lag operator form, we write

$$\Phi(L) y_t = (1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p) y_t = \varepsilon_t .$$

As with our discussion of the MA($q$) process, in our discussion of the **AR($p$) process,** we dispense here with mathematical derivations and instead rely on parallels with the AR(1) case to establish intuition for its key properties.

An AR($p$) process is covariance stationary if and only if the inverses of all roots of the autoregressive lag operator polynomial $\Phi(L)$ are inside the unit circle.[4] In the covariance stationary case, we can write the process in the convergent infinite moving average form

$$y_t = \frac{1}{\Phi(L)} \varepsilon_t .$$

The autocorrelation function for the general AR($p$) process, as with that of the AR(1) process, decays gradually with displacement. Finally, the AR($p$) partial autocorrelation function has a sharp cutoff at displacement $p$, for the same reason that the AR(1) partial autocorrelation function has a sharp cutoff at displacement 1.

Let's discuss the AR($p$) autocorrelation function in a bit greater depth. The key insight is that, in spite of the fact that its qualitative behavior (gradual damping) matches that of the AR(1) autocorrelation function, it can nevertheless display a richer variety of patterns, depending on the order and parameters of the process. It can, for example, have damped monotonic decay, as in the AR(1) case with a positive coefficient, but it can also have damped oscillation in ways that AR(1) can't have. In the AR(1) case, the only possible oscillation occurs when the coefficient is negative, in which case the autocorrelations switch signs at each successively longer displacement. In higher-order autoregressive models, however, the autocorrelations can oscillate with much richer patterns reminiscent of cycles in the more traditional sense. This occurs when some roots of the autoregressive lag operator polynomial are complex.[5]

Consider, for example, the AR(2) process,

$$y_t = 1.5 y_{t-1} - 0.9 y_{t-2} + \varepsilon_t .$$

The corresponding lag operator polynomial is $1 - 1.5L + 0.9L^2$, with two complex conjugate roots, $0.83 \pm 0.65i$. The inverse roots are $0.75 \pm 0.58i$, both of

[4] A necessary **condition for covariance stationarity,** which is often useful as a quick check, is $\sum_{i=1}^{p} \varphi_i < 1$. If the condition is satisfied, the process may or may not be stationary; but if the condition is violated, the process can't be stationary.

[5] Note that complex roots can't occur in the AR(1) case.

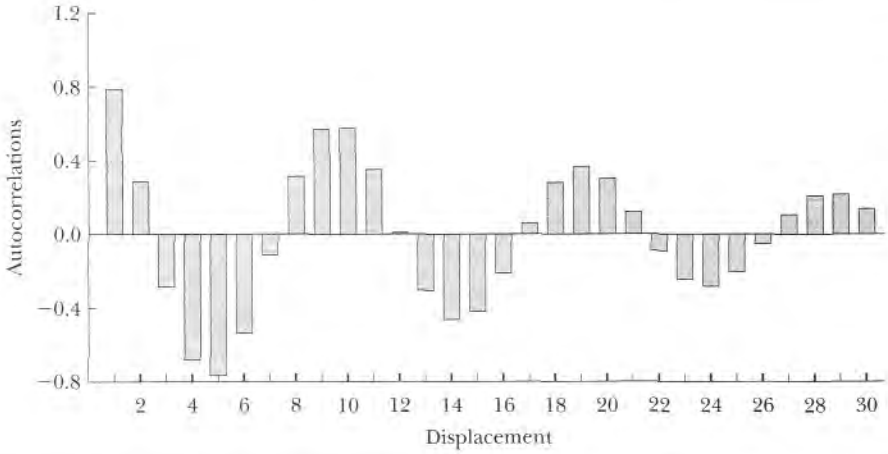



FIGURE 8.11
*Population Autocorrelation Function, AR(2) Process with Complex Roots*

which are close to, but inside, the unit circle; thus, the process is covariance stationary. It can be shown that the autocorrelation function for an AR(2) process is

$$\rho(0) = 1$$
$$\rho(1) = \frac{\varphi_1}{1 - \varphi_2}$$
$$\rho(\tau) = \varphi_1 \rho(\tau - 1) + \varphi_2 \rho(\tau - 2), \quad \tau = 2, 3, \ldots .$$

Using this formula, we can evaluate the autocorrelation function for the process at hand; we plot it in Figure 8.11. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

Finally, let's step back once again to consider in greater detail the precise way that finite-order autoregressive processes approximate the Wold representation. As always, the Wold representation is

$$y_t = B(L)\varepsilon_t ,$$

where $B(L)$ is of infinite order. The AR(1), as compared to the MA(1), is simply a different approximation to the Wold representation. The moving average representation associated with the AR(1) process is

$$y_t = \frac{1}{1 - \varphi L}\varepsilon_t .$$

Thus, when we fit an AR(1) model, we're using $\frac{1}{1-\varphi L}$, a rational polynomial with degenerate numerator polynomial (degree 0) and denominator polynomial of degree 1, to approximate $B(L)$. The moving average representation associated with the AR(1) process is of infinite order, as is the Wold representation, but it does not have infinitely many free coefficients. In fact, only one parameter, $\varphi$, underlies it.

The AR($p$) is an obvious generalization of the AR(1) strategy for approximating the Wold representation. The moving average representation associated with the AR($p$) process is

$$y_t = \frac{1}{\Phi(L)}\varepsilon_t .$$

When we fit an AR($p$) model to approximate the Wold representation we're still using a rational polynomial with degenerate numerator polynomial (degree 0), but the denominator polynomial is of higher degree.

## 3. Autoregressive Moving Average (ARMA) Models

Autoregressive and moving average models are often combined in attempts to obtain better and more parsimonious approximations to the Wold representation, yielding the autoregressive moving average process, **ARMA($p$, $q$) process** for short. As with moving average and autoregressive processes, ARMA processes also have direct motivation.[6] First, if the random shock that drives an autoregressive process is itself a moving average process, then it can be shown that we obtain an ARMA process. Second, ARMA processes can arise from aggregation. For example, sums of AR processes, or sums of AR and MA processes, can be shown to be ARMA processes. Finally, AR processes observed subject to measurement error also turn out to be ARMA processes.

The simplest ARMA process that's not a pure autoregression or pure moving average is the ARMA(1, 1), given by

$$y_t = \varphi y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$$
$$\varepsilon_t \sim WN(0, \sigma^2) ,$$

or, in lag operator form,

$$(1 - \varphi L)y_t = (1 + \theta L)\varepsilon_t ,$$

where $|\theta| < 1$ is required for stationarity and $|\theta| < 1$ is required for invertibility.[7] If the covariance stationarity condition is satisfied, then we have the moving average representation

$$y_t = \frac{(1 + \theta L)}{(1 - \varphi L)}\varepsilon_t ,$$

which is an infinite distributed lag of current and past innovations. Similarly, if the invertibility condition is satisfied, then we have the infinite autoregressive representation,

$$\frac{(1 - \varphi L)}{(1 + \theta L)} y_t = \varepsilon_t .$$

[6] For more extensive discussion, see Granger and Newbold (1986).

[7] Both stationarity and invertibility need to be checked in the ARMA case, because both autoregressive and moving average components are present.

The ARMA($p$, $q$) process is a natural generalization of the ARMA(1, 1) that allows for multiple moving average and autoregressive lags. We write

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$
$$\varepsilon_t \sim WN(0, \sigma^2) ,$$

or

$$\Phi(L) y_t = \Theta(L) \varepsilon_t ,$$

where

$$\Phi(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p$$

and

$$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q .$$

If the inverses of all roots of $\Phi(L)$ are inside the unit circle, then the process is covariance stationary and has convergent infinite moving average representation

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t .$$

If the inverses of all roots of $\Theta(L)$ are inside the unit circle, then the process is invertible and has convergent infinite autoregressive representation

$$\frac{\Phi(L)}{\Theta(L)} y_t = \varepsilon_t .$$

As with autoregressions and moving averages, ARMA processes have a fixed unconditional mean but a time-varying conditional mean. In contrast to pure moving average or pure autoregressive processes, however, neither the autocorrelation nor partial autocorrelation functions of ARMA processes cut off at any particular displacement. Instead, each damps gradually, with the precise pattern depending on the process.

ARMA models approximate the Wold representation by a ratio of two finite-order lag operator polynomials, neither of which is degenerate. Thus, ARMA models use ratios of full-fledged polynomials in the lag operator to approximate the Wold representation,

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t .$$

ARMA models, by allowing for both moving average and autoregressive components, often provide accurate approximations to the Wold representation that nevertheless have just a few parameters. That is, ARMA models are often both highly accurate and highly parsimonious. In a particular situation, for example, it might take an AR(5) to get the same approximation accuracy as could be obtained with an ARMA(2, 1), but the AR(5) has five parameters to be estimated, whereas the ARMA(2, 1) has only three.

## 4. Application: Specifying and Estimating Models for Employment Forecasting

In Chapter 7, we examined the correlogram for the Canadian employment series, and we saw that the sample autocorrelations damp slowly and the sample partial autocorrelations cut off, just the opposite of what's expected for a moving average. Thus, the correlogram indicates that a finite-order moving average process would not provide a good approximation to employment dynamics. Nevertheless, nothing stops us from fitting moving average models, so let's fit them and use the AIC and the SIC to guide model selection.

Moving average models are nonlinear in the parameters; thus, estimation proceeds by nonlinear least squares (numerical minimization). The idea is the same as when we encountered nonlinear least squares in our study of nonlinear trends—pick the parameters to minimize the sum of squared residuals—but finding an expression for the residual is a little bit trickier. To understand why moving average models are nonlinear in the parameters, and to get a feel for how they're estimated, consider an invertible MA(1) model, with a nonzero mean explicitly included for added realism,

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \ .$$

Substitute backward $m$ times to obtain the autoregressive approximation

$$y_t \approx \frac{\mu}{1+\theta} + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots + (-1)^{m+1}\theta^m y_{t-m} + \varepsilon_t \ .$$

Thus, an invertible moving average can be approximated as a finite-order autoregression. The larger is $m$, the better the approximation. This lets us (approximately) express the residual in terms of observed data, after which we can use a computer to solve for the parameters that minimize the sum of squared residuals,

$$\hat{\mu}, \hat{\theta} = \underset{\mu,\theta}{\operatorname{argmin}} \sum_{t=1}^{T} \left( y_t - \left( \frac{\mu}{1+\theta} + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots + (-1)^{m+1}\theta^m y_{t-m} \right) \right)^2$$

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} \left( y_t - \left( \frac{\hat{\mu}}{1+\hat{\theta}} + \hat{\theta} y_{t-1} - \hat{\theta}^2 y_{t-2} + \cdots + (-1)^{m+1}\hat{\theta}^m y_{t-m} \right) \right)^2 \ .$$

The parameter estimates must be found using numerical optimization methods, because the parameters of the autoregressive approximation are restricted. The coefficient of the second lag of $y$ is the square of the coefficient on the first lag of $y$, and so on. The parameter restrictions must be imposed in estimation, which is why we can't simply run an ordinary least-squares regression of $y$ on lags of itself.

The next step would be to estimate MA($q$) models, $q = 1, 2, 3, 4$. Both the AIC and the SIC suggest that the MA(4) is best. To save space, we report only

TABLE 8.1
*Employment MA(4) Model*

LS // Dependent variable is CANEMP.
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 49 iterations

| Variable | Coefficient | Std. Error | *t*-Statistic | Prob. |
|---|---|---|---|---|
| *C* | 100.5438 | 0.843322 | 119.2234 | 0.0000 |
| MA(1) | 1.587641 | 0.063908 | 24.84246 | 0.0000 |
| MA(2) | 0.994369 | 0.089995 | 11.04917 | 0.0000 |
| MA(3) | −0.020305 | 0.046550 | −0.436189 | 0.6635 |
| MA(4) | −0.298387 | 0.020489 | −14.56311 | 0.0000 |

| | | | |
|---|---|---|---|
| $R^2$ | 0.849951 | Mean dependent var. | 101.0176 |
| Adjusted $R^2$ | 0.845071 | SD dependent var. | 7.499163 |
| SE of regression | 2.951747 | Akaike info criterion | 2.203073 |
| Sum squared resid. | 1071.676 | Schwarz criterion | 2.314481 |
| Log likelihood | −317.6208 | *F*-statistic | 174.1826 |
| Durbin-Watson stat. | 1.246600 | Prob(*F*-statistic) | 0.000000 |

| | | | | |
|---|---|---|---|---|
| Inverted MA roots | .41 | −.56 + .72i | −.56 − .72i | −.87 |

the results of MA(4) estimation in Table 8.1. The results of the MA(4) estimation, although better than lower-order MAs, are nevertheless poor. The $R^2$ of 0.84 is rather low, for example, and the Durbin-Watson statistic indicates that the MA(4) model fails to account for all the serial correlation in employment. The residual plot, which we show in Figure 8.12, clearly indicates a neglected cycle, an impression confirmed by the residual correlogram (Table 8.2 and Figure 8.13).

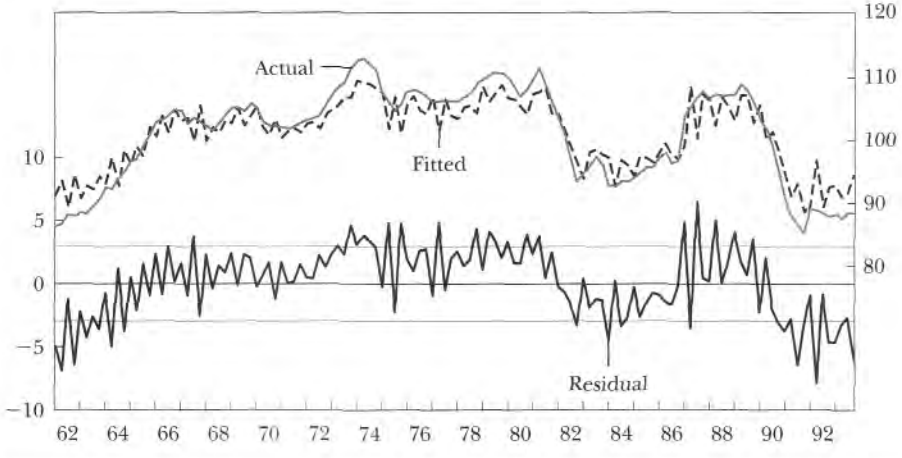



FIGURE 8.12
*Employment MA(4) Model, Residual Plot*

TABLE 8.2 *Employment MA(4) Model, Residual Correlogram*

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 4 ARMA term(s)

| | Acorr. | P. Acorr. | Std. Error | Ljung-Box | *p*-value |
|---|---|---|---|---|---|
| 1 | 0.345 | 0.345 | .088 | 15.614 | |
| 2 | 0.660 | 0.614 | .088 | 73.089 | |
| 3 | 0.534 | 0.426 | .088 | 111.01 | |
| 4 | 0.427 | −0.042 | .088 | 135.49 | |
| 5 | 0.347 | −0.398 | .088 | 151.79 | 0.000 |
| 6 | 0.484 | 0.145 | .088 | 183.70 | 0.000 |
| 7 | 0.121 | −0.118 | .088 | 185.71 | 0.000 |
| 8 | 0.348 | −0.048 | .088 | 202.46 | 0.000 |
| 9 | 0.148 | −0.019 | .088 | 205.50 | 0.000 |
| 10 | 0.102 | −0.066 | .088 | 206.96 | 0.000 |
| 11 | 0.081 | −0.098 | .088 | 207.89 | 0.000 |
| 12 | 0.029 | −0.113 | .088 | 208.01 | 0.000 |

If we insist on using a moving average model, we'd want to explore orders greater than 4, but all the results thus far indicate that moving average processes don't provide good approximations to employment dynamics. Thus, let's consider alternative approximations, such as autoregressions. Autoregressions can be conveniently estimated by ordinary least-squares regression. Consider, for example, the AR(1) model,

$$(y_t - \mu) = \varphi(y_{t-1} - \mu) + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

We can write it as

$$y_t = c + \varphi y_{t-1} + \varepsilon_t,$$

where $c = \mu(1 - \varphi)$. The least-squares estimators are

$$\hat{c}, \hat{\varphi} = \underset{c,\varphi}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - c - \varphi y_{t-1})^2$$

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{c} - \hat{\varphi} y_{t-1})^2.$$

The implied estimate of $\mu$ is $\hat{\mu} = \hat{c}/(1 - \hat{\varphi})$. Unlike the moving average case, for which the sum-of-squares function is nonlinear in the parameters,

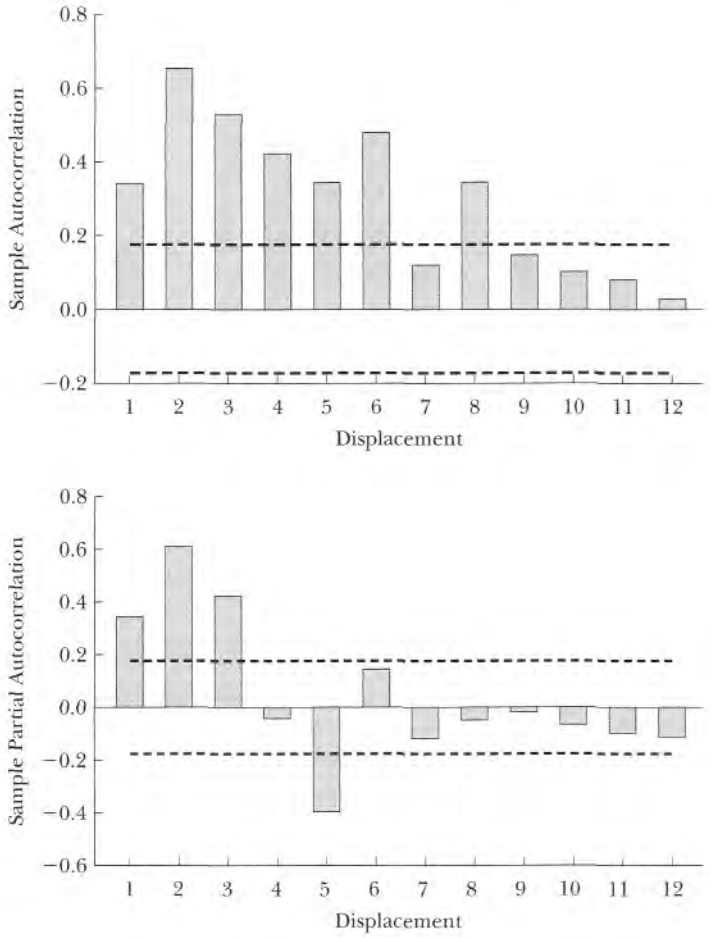



FIGURE 8.13
*Employment MA(4) Model: Residual Sample Autocorrelation and Partial Autocorrelation Functions, with Plus or Minus Two-Standard-Error Bands*

requiring the use of numerical minimization methods, the sum of squares function for autoregressive processes is linear in the parameters, so that estimation is particularly stable and easy. In the AR(1) case, we simply run an ordinary least-squares regression of $y$ on one lag of $y$; in the AR($p$) case, we regress y on $p$ lags of $y$.

We estimate AR($p$) models, $p = 1, 2, 3, 4$. Both the AIC and the SIC suggest that the AR(2) is best. To save space, we report only the results of AR(2) estimation in Table 8.3. The estimation results look good, and the residuals (Figure 8.14) look like white noise. The residual correlogram (Table 8.4 and Figure 8.15) supports that conclusion.

Finally, we consider ARMA($p$, $q$) approximations to the Wold representation. ARMA models are estimated in a fashion similar to moving average

TABLE 8.3
*Employment AR(2) Model*

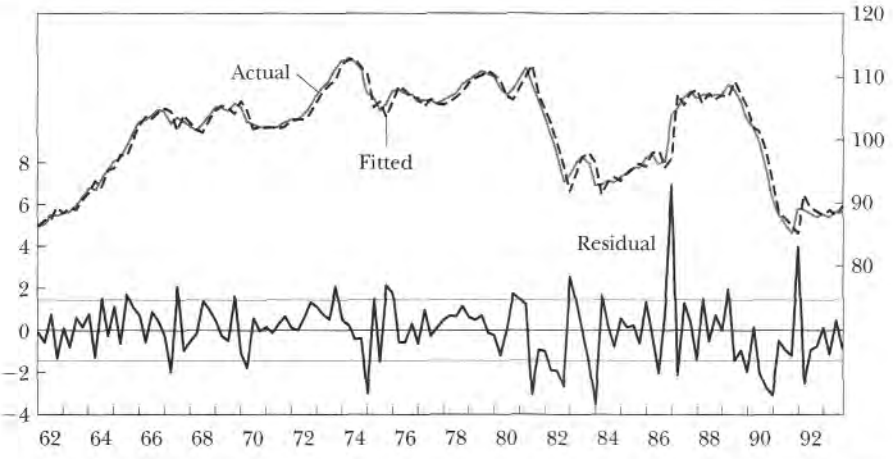
LS // Dependent variable is CANEMP.
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 3 iterations

| Variable | Coefficient | Std. Error | *t*-Statistic | Prob. |
|---|---|---|---|---|
| $C$ | 101.2413 | 3.399620 | 29.78017 | 0.0000 |
| AR(1) | 1.438810 | 0.078487 | 18.33188 | 0.0000 |
| AR(2) | −0.476451 | 0.077902 | −6.116042 | 0.0000 |

| | | | |
|---|---|---|---|
| $R^2$ | 0.963372 | Mean dependent var. | 101.0176 |
| Adjusted $R^2$ | 0.962786 | SD dependent var. | 7.499163 |
| SE of regression | 1.446663 | Akaike info criterion | 0.761677 |
| Sum squared resid. | 261.6041 | Schwarz criterion | 0.828522 |
| Log likelihood | −227.3715 | *F*-statistic | 1643.837 |
| Durbin-Watson stat. | 2.067024 | Prob(*F*-statistic) | 0.000000 |
| Inverted AR roots | .92 | .52 | |

models; they have autoregressive approximations with nonlinear restrictions on the parameters, which we impose when doing a numerical sum of squares minimization. We examine all ARMA($p$, $q$) models with $p$ and $q$ less than or equal to 4; the SIC and AIC values appear in Tables 8.5 and 8.6. The SIC selects the AR(2) (an ARMA(2, 0)), which we've already discussed. The AIC, which penalizes degrees of freedom less harshly, selects an ARMA(3, 1) model.

FIGURE 8.14
*Employment AR(2) Model, Residual Plot*



Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 2 ARMA term(s)

TABLE 8.4
*Employment AR(2) Model, Residual Correlogram*

| | Acorr. | P. Acorr. | Std. Error | Ljung-Box | *p*-value |
|---|---|---|---|---|---|
| 1 | −0.035 | −0.035 | .088 | 0.1606 | |
| 2 | 0.044 | 0.042 | .088 | 0.4115 | |
| 3 | 0.011 | 0.014 | .088 | 0.4291 | 0.512 |
| 4 | 0.051 | 0.050 | .088 | 0.7786 | 0.678 |
| 5 | 0.002 | 0.004 | .088 | 0.7790 | 0.854 |
| 6 | 0.019 | 0.015 | .088 | 0.8272 | 0.935 |
| 7 | −0.024 | −0.024 | .088 | 0.9036 | 0.970 |
| 8 | 0.078 | 0.072 | .088 | 1.7382 | 0.942 |
| 9 | 0.080 | 0.087 | .088 | 2.6236 | 0.918 |
| 10 | 0.050 | 0.050 | .088 | 2.9727 | 0.936 |
| 11 | −0.023 | −0.027 | .088 | 3.0504 | 0.962 |
| 12 | −0.129 | −0.148 | .088 | 5.4385 | 0.860 |

TABLE 8.5
*Employment AIC Values, Various ARMA Models*

| | | MA Order | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| | 0 | | 2.86 | 2.32 | 2.47 | 2.20 |
| | 1 | 1.01 | 0.83 | 0.79 | 0.80 | 0.81 |
| **AR Order** | 2 | 0.762 | 0.77 | 0.78 | 0.80 | 0.80 |
| | 3 | 0.77 | 0.761 | 0.77 | 0.78 | 0.79 |
| | 4 | 0.79 | 0.79 | 0.77 | 0.79 | 0.80 |

TABLE 8.6
*Employment SIC Values, Various ARMA Models*

| | | MA Order | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| | 0 | | 2.91 | 2.38 | 2.56 | 2.31 |
| | 1 | 1.05 | 0.90 | 0.88 | 0.91 | 0.94 |
| **AR Order** | 2 | 0.83 | 0.86 | 0.89 | 0.92 | 0.96 |
| | 3 | 0.86 | 0.87 | 0.90 | 0.94 | 0.96 |
| | 4 | 0.90 | 0.92 | 0.93 | 0.97 | 1.00 |

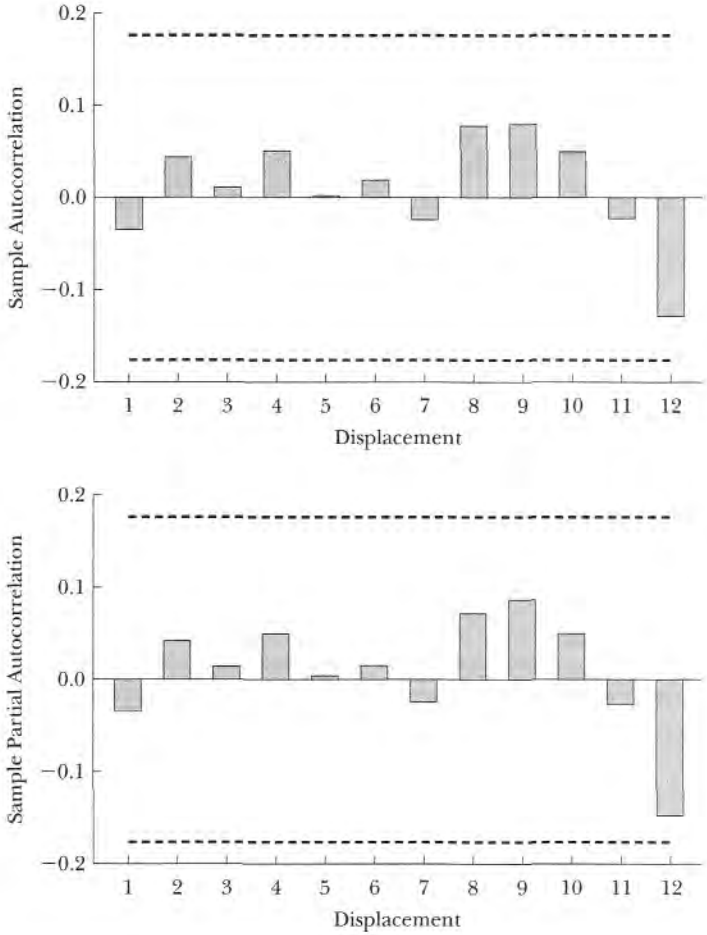



FIGURE 8.15
*Employment AR(2) Model: Residual Sample Autocorrelation and Partial Autocorrelation Functions, with Plus or Minus Two-Standard-Error Bands*

The ARMA(3, 1) model looks good; the estimation results appear in Table 8.7, the residual plot in Figure 8.16, and the residual correlogram in Table 8.8 and Figure 8.17.

Although the ARMA(3, 1) looks good, apart from its lower AIC, it looks no better than the AR(2), which basically seemed perfect. In fact, there are at least three reasons to prefer the AR(2). First, for the reasons discussed in Chapter 5, when the AIC and the SIC disagree, we recommend using the more parsimonious model selected by the SIC. Second, if we consider a model selection strategy involving examination of not just the AIC and SIC but also autocorrelations and partial autocorrelations, which we advocate, we're led to the AR(2). Finally, and importantly, the impression that the ARMA(3, 1) provides a richer approximation to employment dynamics is likely spurious in this case. The

TABLE 8.7
*Employment ARMA(3, 1) Model*

LS // Dependent variable is CANEMP.
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 17 iterations

| Variable | Coefficient | Std. Error | *t*-Statistic | Prob. |
|---|---|---|---|---|
| *C* | 101.1378 | 3.538602 | 28.58130 | 0.0000 |
| AR(1) | 0.500493 | 0.087503 | 5.719732 | 0.0000 |
| AR(2) | 0.872194 | 0.067096 | 12.99917 | 0.0000 |
| AR(3) | −0.443355 | 0.080970 | −5.475560 | 0.0000 |
| MA(1) | 0.970952 | 0.035015 | 27.72924 | 0.0000 |

| | | | |
|---|---|---|---|
| $R^2$ | 0.964535 | Mean dependent var. | 101.0176 |
| Adjusted $R^2$ | 0.963381 | SD dependent var. | 7.499163 |
| SE of regression | 1.435043 | Akaike info criterion | 0.760668 |
| Sum squared resid. | 253.2997 | Schwarz criterion | 0.872076 |
| Log likelihood | −225.3069 | *F*-statistic | 836.2912 |
| Durbin-Watson stat. | 2.057302 | Prob(*F*-statistic) | 0.000000 |

| | | | |
|---|---|---|---|
| Inverted AR roots | .93 | .51 | −.94 |
| Inverted MA roots | −.97 | | |

ARMA(3, 1) has a inverse autoregressive root of −0.94 and an inverse moving average root of −0.97. Those roots are of course just *estimates*, subject to sampling uncertainty, and are likely to be statistically indistinguishable from one another, in which case we can *cancel* them, which brings us down to an ARMA(2, 0), or AR(2), model with roots virtually indistinguishable from those

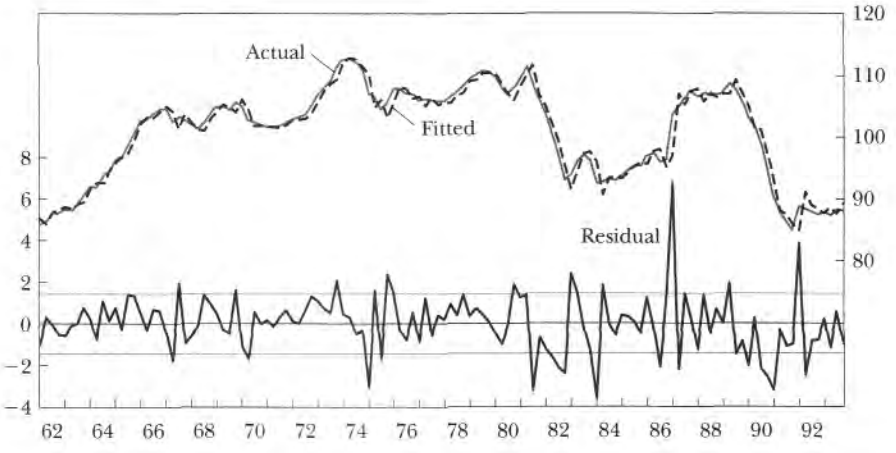


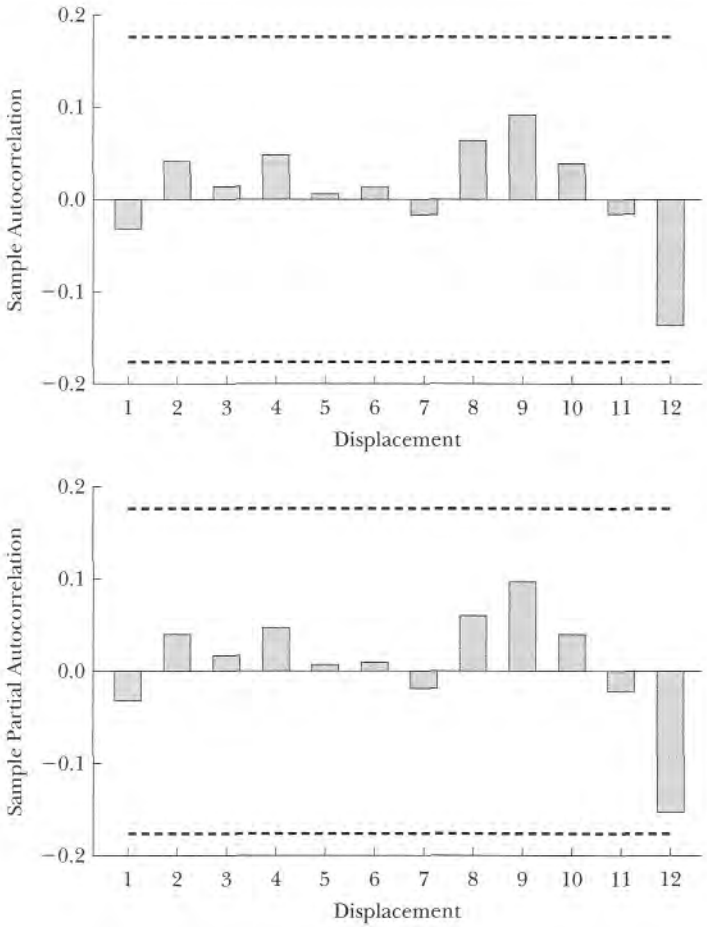

FIGURE 8.16
*Employment ARMA(3, 1) Model, Residual Plot*

TABLE 8.8
*Employment ARMA(3, 1) Model, Residual Correlogram*

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for four ARMA term(s)

| | Acorr. | P. Acorr. | Std. Error | Ljung-Box | *p*-value |
|---|---|---|---|---|---|
| 1 | −0.032 | −0.032 | .09 | 0.1376 | |
| 2 | 0.041 | 0.040 | .09 | 0.3643 | |
| 3 | 0.014 | 0.017 | .09 | 0.3904 | |
| 4 | 0.048 | 0.047 | .09 | 0.6970 | |
| 5 | 0.006 | 0.007 | .09 | 0.7013 | 0.402 |
| 6 | 0.013 | 0.009 | .09 | 0.7246 | 0.696 |
| 7 | −0.017 | −0.019 | .09 | 0.7650 | 0.858 |
| 8 | 0.064 | 0.060 | .09 | 1.3384 | 0.855 |
| 9 | 0.092 | 0.097 | .09 | 2.5182 | 0.774 |
| 10 | 0.039 | 0.040 | .09 | 2.7276 | 0.842 |
| 11 | −0.016 | −0.022 | .09 | 2.7659 | 0.906 |
| 12 | −0.137 | −0.153 | .09 | 5.4415 | 0.710 |

FIGURE 8.17
*Employment ARMA(3, 1) Model: Residual Sample Autocorrelation and Partial Autocorrelation Functions, with Plus or Minus Two-Standard-Error Bands*



of our earlier-estimated AR(2) process! We refer to this situation as one of **common factors** in an ARMA model. Be on the lookout for such situations, which arise frequently and can lead to substantial model simplification.

Thus, we arrive at an AR(2) model for employment. In the next chapter, we'll learn how to use it to produce point and interval forecasts.

## Exercises, Problems, and Complements

1. (ARMA lag inclusion) Review Table 8.1. Why is the MA(3) term included even though the *p*-value indicates that it is not significant? What would be the costs and benefits of dropping the insignificant MA(3) term?

2. (Shapes of correlograms) Given the following ARMA processes, sketch the expected forms of the autocorrelation and partial autocorrelation functions. (*Hint:* Examine the roots of the various autoregressive and moving average lag operator polynomials.)

   a. $y_t = \left(\frac{1}{1 - 1.05L - 0.09L^2}\right)\varepsilon_t$

   b. $y_t = (1 - 0.4L)\varepsilon_t$

   c. $y_t = \left(\frac{1}{1 - 0.7L}\right)\varepsilon_t$.

3. (The autocovariance function of the MA(1) process, revisited) In the text, we wrote

$$\gamma(\tau) = E(y_t\, y_{t-\tau}) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) = \begin{cases} \theta\sigma^2, & \tau = 1 \\ 0, & \text{otherwise} \end{cases}.$$

   Fill in the missing steps by evaluating explicitly the expectation $E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1}))$.

4. (ARMA algebra) Derive expressions for the autocovariance function, autocorrelation function, conditional mean, unconditional mean, conditional variance, and unconditional variance of the following processes:

   a. $y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$

   b. $y_t = \varphi y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$

5. (Diagnostic checking of model residuals) If a forecasting model has extracted all the systematic information from the data, then what's left—the residual—should be white noise. More precisely, the true innovations are white noise, and if a model is a good approximation to the Wold representation, then its 1-step-ahead forecast errors should be approximately white noise. The model residuals are the in-sample analog of out-of-sample 1-step-ahead forecast errors—hence the usefulness of various tests of the hypothesis that residuals are white noise.

   The Durbin-Watson test is the most popular. Recall the Durbin-Watson test statistic, discussed in Chapter 2,

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} .$$

Note that

$$\sum_{t=2}^{T}(e_t - e_{t-1})^2 \approx 2\sum_{t=2}^{T} e_t^2 - 2\sum_{t=2}^{T} e_t e_{t-1} .$$

Thus,

$$DW \approx 2(1 - \hat{\rho}(1)) ,$$

so that the Durbin-Watson test is effectively based only on the first sample autocorrelation and really only tests whether the first autocorrelation is 0. We say therefore that the Durbin-Watson is a test for **first-order serial correlation.** In addition, the Durbin-Watson test is not valid in the presence of lagged dependent variables.[8] On both counts, we'd like a more general and flexible framework for diagnosing serial correlation. The residual correlogram, comprised of the residual sample autocorrelations, the sample partial autocorrelations, and the associated Q-statistics, delivers the goods.

a. When we discussed the correlogram in the text, we focused on the case of an observed time series, in which case we showed that the Q-statistics are distributed as $\chi^2_m$. Now, however, we want to assess whether unobserved model disturbances are white noise. To do so, we use the model residuals, which are estimates of the unobserved disturbances. Because we fit a model to get the residuals, we need to account for the degrees of freedom used. The upshot is that the distribution of the Q-statistics under the white noise hypothesis is better approximated by a $\chi^2_{m-k}$ random variable, where $k$ is the number of parameters estimated. That's why, for example, we don't report (and in fact the software doesn't compute) the $p$-values for the Q-statistics associated with the residual correlogram of our employment forecasting model until $m > k$.
b. **Durbin's *h*-test** is an alternative to the Durbin-Watson test. As with the Durbin-Watson test, it's designed to detect first-order serial correlation, but it's valid in the presence of lagged dependent variables. Do some background reading on Durbin's $h$-test, and report what you learned.
c. The **Breusch-Godfrey test** is another alternative to the Durbin-Watson test. It's designed to detect $p$th-order serial correlation, where $p$ is selected by the user, and is also valid in the presence of lagged dependent variables. Do some background reading on the Breusch-Godfrey procedure, and report what you learned.
d. Which do you think is likely to be most useful to you in assessing the properties of residuals from forecasting models: the residual correlogram, Durbin's $h$-test, or the Breusch-Godfrey test? Why?

[8] Following standard, if not strictly appropriate, practice, in this book we often report and examine the Durbin-Watson statistic even when lagged dependent variables are included. We always supplement the Durbin-Watson statistic, however, with other diagnostics such as the residual correlogram, which remain valid in the presence of lagged dependent variables and which almost always produce the same inference as the Durbin-Watson statistic.

6. (Mechanics of fitting ARMA models) The book's web page presents data for daily transfers over BankWire, a financial wire transfer system in a country responsible for much of the world's finance, over a recent span of 200 business days.
   a. Is trend or seasonality operative? Defend your answer.
   b. Using the methods developed in Chapters 7 and 8, find a parsimonious ARMA($p$, $q$) model that fits well, and defend its adequacy.

7. (Modeling cyclical dynamics) As a research analyst at the U.S. Department of Energy, you have been asked to model nonseasonally adjusted U.S. imports of crude oil.
   a. Find a suitable time series on the web.
   b. Create a model that captures the trend in the series.
   c. Adding to the model from part *b*, create a model with trend and a full set of seasonal dummy variables.
   d. Observe the residuals of the model from part *b* and their correlogram. Is there evidence of neglected dynamics? If so, what to do?

8. (**Aggregation** and **disaggregation: top-down forecasting model** vs. **bottom-up forecasting model**) Related to the issue of methods and complexity discussed in Chapter 3 is the question of aggregation. Often we want to forecast an aggregate, such as total sales of a manufacturing firm, but we can take either an aggregated or disaggregated approach.

   Suppose, for example, that total sales is composed of sales of three products. The aggregated, or top-down or macro, approach is simply to model and forecast total sales. The disaggregated, or bottom-up or micro, approach is to model and forecast separately the sales of the individual products and then to add them together.

   Perhaps surprisingly, it's impossible to know in advance whether the aggregated or disaggregated approach is better. It all depends on the specifics of the situation; the only way to tell is to try both approaches and compare the forecasting results.

   However, in real-world situations characterized by likely model misspecification and parameter estimation uncertainty, there are reasons to suspect that the aggregated approach may be preferable. First, standard (e.g., linear) models fit to aggregated series may be less prone to specification error, because aggregation can produce approximately linear relationships even when the underlying disaggregated relationships are not linear. Second, if the disaggregated series depends in part on a common factor (e.g., general business conditions), then it will emerge more clearly in the aggregate data. Finally, modeling and forecasting of one aggregated series, as opposed to many disaggregated series, rely on far fewer parameter estimates.

   Of course, if our interest centers on the disaggregated components, then we have no choice but to take a disaggregated approach.

   It is possible that an aggregate forecast may be useful in forecasting disaggregated series. Why? (*Hint:* See Fildes and Stekler, 2000.)

9. (Nonlinear forecasting models: regime switching) In this chapter, we've studied dynamic **linear models,** which are tremendously important in practice. They're called *linear* because $y_t$ is a simple linear function of past y's or past $\varepsilon$'s. In some forecasting situations, however, good statistical characterization of dynamics may require some notion of regime switching, as between "good" and "bad" states, which is a type of **nonlinear model.**

Models incorporating **regime switching** have a long tradition in business cycle analysis, in which expansion is the good state, and contraction (recession) is the bad state. This idea is also manifest in the great interest in the popular press; for example, in identifying and forecasting turning points in economic activity. It is only within a regime-switching framework that the concept of a turning point has intrinsic meaning; turning points are naturally and immediately defined as the times separating expansions and contractions.

**Threshold models** are squarely in line with the regime-switching tradition. The following threshold model, for example, has three regimes, two thresholds, and a $d$-period delay regulating the switches:

$$y_t = \begin{cases} c^{(u)} + \varphi^{(u)} y_{t-1} + \varepsilon_t^{(u)}, & \theta^{(u)} < y_{t-d} \\ c^{(m)} + \varphi^{(m)} y_{t-1} + \varepsilon_t^{(m)}, & \theta^{(l)} < y_{t-d} < \theta^{(u)} \\ c^{(l)} + \varphi^{(l)} y_{t-1} + \varepsilon_t^{(l)}, & \theta^{(l)} > y_{t-d} \end{cases}.$$

The superscripts indicate "upper," "middle," and "lower" regimes, and the regime operative at any time $t$ depends on the observable past history of $y$—in particular, on the value of $y_{t-d}$.

Although observable threshold models are of interest, models with *latent* (or unobservable) states as opposed to observed states may be more appropriate in many business, economic, and financial contexts. In such a setup, time series dynamics are governed by a finite-dimensional parameter vector that switches (potentially each period) depending on which of two unobservable states is realized, with state transitions governed by a first-order Markov process (meaning that the state at any time $t$ depends only on the state at time $t-1$, not at time $t-2$, $t-3$, etc.).

To make matters concrete, let's take a simple example. Let $\{s_t\}_{t=1}^T$ be the (latent) sample path of a two-state first-order autoregressive process, taking just the two values 0 or 1, with the transition probability matrix given by

$$M = \begin{pmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{pmatrix}.$$

The $ij$th element of $M$ gives the probability of moving from state $i$ (at time $t-1$) to state $j$ (at time $t$). Note that there are only two free parameters, the staying probabilities, $p_{00}$ and $p_{11}$. Let $\{y_t\}_{t=1}^T$ be the sample path of an observed time series that depends on $\{s_t\}_{t=1}^T$ such that the density of $y_t$ conditional on $s_t$ is

$$f(y_t \mid s_t; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_t - \mu_{s_t})^2}{2\sigma^2}\right).$$

Thus, $y_t$ is Gaussian white noise with a potentially switching mean. The two means around which $y_t$ moves are of particular interest and may, for example, correspond to episodes of differing growth rates ("booms" and "recessions," "bull" and "bear" markets, etc.).

10. (Difficulties with nonlinear optimization) Nonlinear optimization is a tricky business, fraught with problems. Some eye-opening reading includes Newbold, Agiakloglou, and Miller (1994) and McCullough and Vinod (1999).

Some problems are generic. It's relatively easy to find a local optimum, for example, but much harder to be confident that the local optimum is global. Simple checks such as trying a variety of startup values and checking the optimum to which convergence occurs are used routinely, but the problem nevertheless remains. Other problems may be software specific. For example, some software may use highly accurate analytic derivatives, whereas other software uses approximate numerical derivatives. Even the same software package may change algorithms or details of implementation across versions, leading to different results. Software for ARMA model estimation is unavoidably exposed to all such problems, because estimation of any model involving MA terms requires numerical optimization of a likelihood or sum-of-squares function.

## Bibliographical and Computational Notes

Characterization of time series by means of autoregressive, moving average, or ARMA models was suggested, more or less simultaneously, by the Russian statistician and economist E. Slutsky and the British statistician G. U. Yule. Slutsky (1927) remains a classic. The Slutsky-Yule framework was modernized, extended, and made part of an innovative and operational modeling and forecasting paradigm in a more recent classic, a 1970 book by Box and Jenkins, the latest edition of which is Box, Jenkins, and Reinsel (1994). In fact, ARMA and related models are often called **Box-Jenkins models.**

Granger and Newbold (1986) contains more detailed discussion of a number of topics that arose in this chapter, including the idea of moving average processes as describing economic equilibrium disturbed by transient shocks, the Yule-Walker equation, and the insight that aggregation and measurement error lead naturally to ARMA processes.

The sample autocorrelations and partial autocorrelations, together with related diagnostics, provide graphical aids to model selection that complement the Akaike and Schwarz information criteria introduced earlier. Not long ago, the sample autocorrelation and partial autocorrelation functions were often used *alone* to guide forecast model selection, a tricky business that was more art than science. Use of the Akaike and Schwarz criteria results in more systematic and replicable model selection, but the sample autocorrelation and partial autocorrelation functions nevertheless remain important as basic graphical summaries of dynamics in time series data. The two approaches are complements, not substitutes.

Our discussion of estimation was a bit fragmented; we discussed estimation of moving average and ARMA models using nonlinear least squares, whereas we discussed estimation of autoregressive models using ordinary least squares. A more unified approach proceeds by writing each model as a regression on an intercept, with a serially correlated disturbance. Thus, the moving average model is

$$y_t = \mu + \varepsilon_t$$

$$\varepsilon_t = \Theta(L)v_t$$

$$v_t \sim WN(0, \sigma^2) ,$$

the autoregressive model is

$$y_t = \mu + \varepsilon_t$$
$$\Phi(L)\varepsilon_t = v_t$$
$$v_t \sim WN(0, \sigma^2),$$

and the ARMA model is

$$y_t = \mu + \varepsilon_t$$
$$\Phi(L)\varepsilon_t = \Theta(L)v_t$$
$$v_t \sim WN(0, \sigma^2).$$

We can estimate each model in identical fashion using nonlinear least squares. Eviews and other forecasting packages proceed in precisely that way.[9]

This framework—regression on a constant with serially correlated disturbances—has a number of attractive features. First, the mean of the process is the regression constant term.[10] Second, it leads us naturally toward regression on more than just a constant, as other right-hand-side variables can be added as desired. Finally, it exploits the fact that because autoregressive and moving average models are special cases of the ARMA model, their estimation is also a special case of estimation of the ARMA model.

Our description of estimating ARMA models—compute the autoregressive representation, truncate it, and estimate the resulting approximate model by nonlinear least squares—is conceptually correct but intentionally simplified. The actual estimation methods implemented in modern software are more sophisticated, and the precise implementations vary across software packages. Beneath it all, however, all estimation methods are closely related to our discussion, whether implicitly or explicitly. You should consult your software manual for details. (Hopefully they're provided!)

Pesaran, Pierse, and Kumar (1989) and Granger (1990) study the question of top-down versus bottom-up forecasting. For a comparative analysis in the context of forecasting Euro-area macroeconomic activity, see Stock and Watson (2003).

Our discussion of regime-switching models draws heavily on Diebold and Rudebusch (1996). Tong (1983) is a key reference on observable-state threshold models, as is Hamilton (1989) for latent-state threshold models. There are a number of extensions of those basic regime-switching models of potential interest for forecasters, such as allowing for smooth as opposed to abrupt transitions in threshold models with observed states (Granger and Teräsvirta, 1993) and allowing for time-varying transition probabilities in threshold models with latent states (Diebold, Lee, and Weinbach, 1994).

## Concepts for Review

Moving Average (MA) model
Autoregressive (AR) model
Autoregressive moving average (ARMA) model
Stochastic process
MA(1) process
Cutoff in the autocorrelation function
Invertibility

[9] That's why, for example, information on the number of iterations required for convergence is presented even for estimation of the autoregressive model.

[10] Hence the notation "$\mu$" for the intercept.

Autoregressive representation
MA($q$) process
Complex roots
Condition for invertibility of the MA($q$)
Yule-Walker equation
AR($p$) process
Condition for covariance stationarity
ARMA($p$, $q$) process
Common factors
First-order serial correlation
Durbin's $h$-test
Breusch-Godfrey test
Aggregation
Disaggregation
Top-down forecasting model
Bottom-up forecasting model
Linear model
Nonlinear model
Regime switching
Threshold model
Box-Jenkins model

## References and Additional Readings


Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31, 307–327.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). "ARCH Modeling in Finance: A Selective Review of the Theory and Empirical Evidence." *Journal of Econometrics*, 52, 5–59.

Box, G. E. P., Jenkins, G. W., and Reinsel, G. (1994). *Time Series Analysis, Forecasting and Control.* 3rd ed. Englewood Cliffs, N. J.: Prentice Hall.

Burns, A. F., and Mitchell, W. C. (1946). *Measuring Business Cycles.* New York: National Bureau of Economic Research.

Diebold, F. X., Lee, J.-H., and Weinbach, G. (1994). "Regime Switching with Time-Varying Transition Probabilities." In C. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration.* Oxford: Oxford University Press, 283–302. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Lopez, J. (1995). "Modeling Volatility Dynamics." In Kevin Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects.* Boston: Kluwer Academic Press, 427–472.

Diebold, F. X., and Rudebusch, G. D. (1996). "Measuring Business Cycles: A Modern Perspective." *Review of Economics and Statistics*, 78, 67–77. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1999). *Business Cycles: Durations, Dynamics, and Forecasting.* Princeton, N.J.: Princeton University Press.

Engle, R. F. (1982). "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation." *Econometrica*, 50, 987–1008.

Fildes, R., and Steckler, H. (2000). "The State of Macroeconomic Forecasting." Manuscript.

Granger, C. W. J. (1990). "Aggregation of Time Series Variables: A Survey." In T. Barker and M. H. Pesaran (eds.), *Disaggregation in Econometric Modelling.* London: Routledge.

Granger, C. W. J., and Newbold, P. (1986). *Forecasting Economic Time Series.* 2nd ed. Orlando, Fla.: Academic Press.

Granger, C. W. J., and Teräsvirta, Y. (1993). *Modelling Nonlinear Economic Relationships.* Oxford: Oxford University Press.

Hamilton, J. D. (1989). "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, 57, 357–384.

McCullough, B. D., and Vinod, H. D. (1999). "The Numerical Reliability of Econometric Software." *Journal of Economic Literature*, 37, 633–665.

Newbold, P., Agiakloglou, C., and Miller, J. P. (1994). "Adventures with ARIMA Software." *International Journal of Forecasting*, 10, 573–581.

Pesaran, M. H., Pierse, R. G., and Kumar, M. S. (1989). "Econometric Analysis of Aggregation in the Context of Linear Prediction Models." *Econometrica*, 57, 861–888.

Slutsky, E. (1927). "The Summation of Random Causes as the Source of Cyclic Processes." *Econometrica*, 5, 105–146.

Stock, J. H., and Watson, M. W. (2003). "Macroeconomic Forecasting in the Euro Area: Country-Specific versus Area-Wide Information." *European Economic Review*, 47, 1–18.

Taylor, S. (1996). *Modeling Financial Time Series*. 2nd ed. New York: Wiley.

Tong, H. (1990). *Non-linear Time Series*. Oxford: Clarendon Press.