Chapter 8

Time Series Modeling

This chapter presents several popular probability models for describing a time series, along with the associated statistical methods. Analogous to using the univariate normal distribution to model a quantitative variable which has a bell-shaped probability distribution, no time series model will provide a perfect fit to the data. The goal is to identify a probability model which provides a reasonable approximation to the time series, fit the model to an observed time series, and then use the fitted model for statistical inference, which is often forecasting.

8.1 Probability Models

A suite of probability models for time series known as *linear models* are introduced in this section. The unifying characteristic of these models is that they express the current value of the time series as a linear function of (a) the current noise term, (b) previous noise terms, and (c) previous values of the time series. We begin by taking a birds-eye view of these linear time series models by introducing *general linear models* (often abbreviated glm) and some of their properties. This is followed by a section that introduces a suite of time series models that are special cases of general linear models that are known as *ARMA* (autoregressive moving average) models. ARMA models are *parsimonious* in the sense that they are able to specify a wide variety of underlying probability models that govern a stationary time series with only a few parameters. With both general linear models and ARMA models, you will see a great deal of symmetry and some mathematics that works out beautifully on the road to developing time series models that can be implemented in real-world applications.

8.1.1 General Linear Models

General linear models provide an important way of thinking about how to define a time series model in a simple and general manner. Working with general linear models also provides some practice with using the backshift operator B, which was defined in Section 7.3.1. We also consider the causal and invertible form of general linear models. The causal form is important for establishing stationarity. The invertible form is important for ensuring a one-to-one relationship between parameter values and the associated population autocorrelation function.

The concepts of white noise from Definition 7.1 and linear filters from Section 7.3.1 are tied together in this section to define general linear models. White noise is a time series of mutually independent random variables denoted by $\{Z_t\}$. Each element in the white noise time series has common population mean 0 and common population variance σ_Z^2. Time series analysts often refer to the Z_t values as *random shocks* whose purpose is to inject randomness into a time series model. Without these shocks, the time series model would be purely deterministic. Linear filters are a general way of expressing one time series as a linear combination of the values in another time series. White noise and linear filters are the key concepts in the definition of general linear models. As you will see in the next paragraph, there are two distinctly different ways of defining general linear models.

More specifically, one way to describe a general linear model is to define the current value in the time series X_t as the current white noise term Z_t plus a linear combination of the previous white noise terms:

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots ,$$

where the coefficients $\psi_1, \psi_2, \ldots$ in the infinite series are real-valued constants. This time series model is stationary when appropriate restrictions are placed on the $\psi_1, \psi_2, \ldots$ values. Since this description of a general linear model is valid at time t, it is also valid at other time values, for example,

$$X_{t-1} = Z_{t-1} + \psi_1 Z_{t-2} + \psi_2 Z_{t-3} + \cdots ,$$

or

$$X_{t-2} = Z_{t-2} + \psi_1 Z_{t-3} + \psi_2 Z_{t-4} + \cdots .$$

Solving these equations for the current white noise value and sequentially substituting into the first formulation of the general linear model, you can see that there is a second way to formulate a general linear model:

$$X_t = Z_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots ,$$

where the coefficients $\pi_1, \pi_2, \ldots$ are real-valued constants and appropriate restrictions are placed on the $\pi_1, \pi_2, \ldots$ values in order to achieve stationarity. In this second formulation of a general linear model, the current value of the time series is a linear combination of the previous values of the time series plus the current white noise term. This formulation is analogous to that of a multiple linear regression model with an infinite number of predictor variables.

A reasonable question to ask at this point is why there is no coefficient associated with Z_t in both formulations of the general linear model. Although some authors associate a coefficient ψ_0 with Z_t, we avoid this practice and simply assume that $\psi_0 = 1$. Including a ψ_0 parameter is redundant because a nonzero constant multiplied by a white noise term is still a white noise term. The population variance of the white noise σ_Z^2 is essentially absorbed into the ψ_0 parameter. Also, some authors use a $-$ rather than a $+$ between terms on the right-hand side of the second formulation of the general linear model.

The two formulations for the general linear model involve a random variable on the left-hand side of the model and random variables on the right-hand side of the model. In some settings, this might be viewed as a transformation of random variables, but this is not the correct interpretation of the model in the time series setting. The general linear model formulations define a hypothesized relationship between the random variable on the left-hand side of the model and the random variables on the right-hand side of the model. In the first formulation of the general linear model, the current value of the time series X_t is hypothesized to be a linear combination of the current and previous noise values. In the second formulation of the general linear model, the current value of the time series X_t is hypothesized to be a linear combination of the previous values in the time series plus a noise term. This probability model is hypothesized to govern the process over time. The goal in constructing a time series model is to write a formula for a model which adequately captures the probabilistic relationship that governs the time series. Estimation of the model parameters will be followed by assessments to see if the model holds in an empirical sense.

The coefficients in the two formulations of a general linear model are related. To make these two formulations of the general linear model more concrete, we will now look at a specific instance.

Example 8.1 Consider the special case of the first formulation of the general linear model

$$X_t = Z_t + \psi_1 Z_{t-1}.$$

This model only has one coefficient ψ_1. The subsequent coefficients are $\psi_j = 0$ for $j = 2, 3, \ldots$. Find the equivalent form of the general linear model using the second formulation.

Recall from Section 7.3.1 that the backshift operator B shifts a time series value back one unit in time, for example,

$$BX_t = X_{t-1}.$$

When the backshift operator includes a superscript, the superscript accounts for multiple backshifts, for example,

$$B^4 Z_t = Z_{t-4}.$$

The special case of the general linear model considered here can be converted from its original form,

$$X_t = Z_t + \psi_1 Z_{t-1},$$

to a form using the backshift operator,

$$X_t = Z_t + \psi_1 B Z_t$$

or

$$X_t = (1 + \psi_1 B) Z_t.$$

Although it might seem like an unusual operation involving B, both sides of this equation can be divided by $1 + \psi_1 B$, which gives

$$\frac{X_t}{1 + \psi_1 B} = Z_t.$$

For ψ_1 values on the interval $-1 < \psi_1 < 1$, this can be expanded as a geometric series with common ratio $-\psi_1 B$:

$$\left(1 - \psi_1 B + \psi_1^2 B^2 - \cdots\right) X_t = Z_t$$

or

$$X_t - \psi_1 X_{t-1} + \psi_1^2 X_{t-2} - \cdots = Z_t$$

or

$$X_t = Z_t + \psi_1 X_{t-1} - \psi_1^2 X_{t-2} + \cdots .$$

This is the second formulation of the general linear model with coefficients $\pi_j = (-1)^{j-1} \psi_1^j$ for $j = 1, 2, \ldots$ and $-1 < \psi < 1$.

A sleight of hand has occurred in the previous example with respect to the use of the backshift operator B, first as an operator and then as a variable. This paragraph concerns that dual use. When B is used as an operator, it has a domain or input, for instance, X_t, and a range or output, for instance, $BX_t = X_{t-1}$. In this case, the effect of the operator B on a time series value is to go back in the time series one unit of time. The input to B is the value of the time series at time t, and the output from B is the value of the time series at time $t-1$. The full domain of the operator B is the entire sequence of time series values. Why is it acceptable to take an operator like the backshift operator B and use it as a variable? It can be demonstrated that the backshift operator B functions like a linear map in the sense of allowing the standard multiplication and addition operations in its domain. In addition to the standard operations such addition, multiplication, and inversion, we may thus treat polynomials in B as polynomials in real variables.

For the particular case of the general linear model considered in the previous example, there was a relationship between the coefficients in the two formulations of the general linear model. We now consider whether there is a relationship between the coefficients $\psi_1, \psi_2, \ldots$ and $\pi_1, \pi_2, \ldots$ in the general setting. We continue with our use of the backshift operator B. The first formulation of the general linear model is

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots,$$

which can be rewritten using the backshift operator as

$$X_t = Z_t + \psi_1 B Z_t + \psi_2 B^2 Z_t + \cdots$$

or

$$X_t = \left(1 + \psi_1 B + \psi_2 B^2 + \cdots\right) Z_t.$$

The polynomial in B in this formulation of the model is denoted by $\psi(B)$, so the first formulation of the general linear model can be written compactly as

$$X_t = \psi(B) Z_t,$$

where $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$.

Now consider the second formulation of the general linear model:

$$X_t = Z_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots.$$

Separating the time series terms on the left-hand side of the equation and the white noise term on the right-hand side of the equation results in

$$X_t - \pi_1 X_{t-1} - \pi_2 X_{t-2} - \cdots = Z_t,$$

which can be rewritten using the backshift operator as

$$X_t - \pi_1 B X_t - \pi_2 B^2 X_t - \cdots = Z_t$$

or

$$\left(1 - \pi_1 B - \pi_2 B^2 - \cdots\right) X_t = Z_t.$$

The polynomial in B in this formulation of the model is denoted by $\pi(B)$, so the second formulation of the general linear model can be written compactly as

$$\pi(B) X_t = Z_t,$$

where $\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \cdots$.

Definition 8.1 gives the two formulations of the general linear model expressed in purely algebraic form and in terms of polynomials in the backshift operator.

Definition 8.1 A time series $\{X_t\}$ can be expressed as a *general linear model* as

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots,$$

where $\psi_1, \psi_2, \ldots$ are real-valued constants and $Z_t \sim WN\left(0, \sigma_Z^2\right)$, or, equivalently, as

$$X_t = \left(1 + \psi_1 B + \psi_2 B^2 + \cdots\right) Z_t = \psi(B) Z_t.$$

Alternatively, the general linear model for a time series can be written as

$$X_t = Z_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots$$

for certain values of the real-valued constants $\pi_1, \pi_2, \ldots$, or, equivalently, as

$$\left(1 - \pi_1 B - \pi_2 B^2 - \cdots\right) X_t = \pi(B) X_t = Z_t.$$

In the previous example, we were able to perform algebraic steps to determine the relationship between the coefficients in the first formulation of the general linear model (that is, $\psi_1, \psi_2, \ldots$) and the coefficients in the second formulation (that is, $\pi_1, \pi_2, \ldots$). This can also be done in the more general setting. The equations that define the two formulations of the general linear model in Definition 8.1 written in terms of the backshift operator are

$$X_t = \psi(B) Z_t \qquad \text{and} \qquad \pi(B) X_t = Z_t.$$

Applying the $\psi(B)$ polynomial to both sides of the second equation gives

$$\psi(B)\pi(B) X_t = \psi(B) Z_t$$

or

$$\psi(B)\pi(B) X_t = X_t$$

or

$$\psi(B)\pi(B) = 1$$

for nonzero X_t. Since the product of the polynomials $\psi(B)$ and $\pi(B)$ is one, they are inverses. For suitable values of the coefficients, this allows us to calculate the coefficients $\psi_1, \psi_2, \ldots$ from the coefficients $\pi_1, \pi_2, \ldots$ and vice versa. The inverse relationship between $\psi(B)$ and $\pi(B)$ will now be confirmed for the polynomials identified in the previous example.

Example 8.2 Verify that $\psi(B)\pi(B) = 1$ for the time series model for $\{X_t\}$ from the previous example:

$$X_t = Z_t + \psi_1 Z_{t-1},$$

where $-1 < \psi_1 < 1$ and $\{Z_t\}$ is a time series of white noise.

From Example 8.1, the polynomials in the backshift operator are

$$\psi(B) = 1 + \psi_1 B$$

and

$$\pi(B) = 1 - \psi_1 B + \psi_1^2 B^2 - \cdots.$$

The product of $\psi(B)$ and $\pi(B)$ is

$$\begin{aligned}\psi(B)\pi(B) &= (1+\psi_1 B)\left(1-\psi_1 B+\psi_1^2 B^2-\cdots\right)\\ &= \left(1-\psi_1 B+\psi_1^2 B^2-\cdots\right)+\left(\psi_1 B-\psi_1^2 B^2+\psi_1^3 B^3-\cdots\right)\\ &= 1\end{aligned}$$

as expected.

The previous discussion constitutes a proof of the following theorem concerning writing the two forms of the general linear model in terms of polynomials in the backshift operator and the relationship between the two polynomials $\psi(B)$ and $\pi(B)$.

Theorem 8.1 The two formulations of the general linear model from Definition 8.1 associated with the two polynomials $\psi(B)$ and $\pi(B)$ are equivalent time series models and are related by

$$\psi(B)\pi(B) = 1$$

for certain values of the coefficients.

We will toggle between the purely algebraic formulations of the general linear model and the associated formulations using the backshift operator B based on which is more convenient and effective for the mathematics in a particular setting. Definition 8.1 gives two different ways of writing a general linear model, but is vague concerning any constraints placed on the coefficients. Some constraints on the coefficients that give the general linear model certain important characteristics are outlined next. Stationarity will play a central role in these constraints. The stationarity property implies that the time series is stable over time; this stability allows us to predict how the time series will behave in the future.

Causality and Invertibility

The general linear model is formulated in two different fashions in Definition 8.1. But we have not yet defined any general constraints on the coefficients in the two different formulations of the general linear model. We begin the consideration of appropriate constraints on the coefficients with some calculations on the first formulation of the general linear model.

The first formulation of the general linear model from Definition 8.1 using the purely algebraic form is

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots .$$

We would like to determine constraints on the coefficients $\psi_1, \psi_2, \ldots$ that will result in a stationary model and also find expressions for quantities associated with the stationary version of this model, such as $E[X_t]$, $V[X_t]$, $\gamma(k)$, and $\rho(k)$. Taking the expected value of both sides of the defining formula results in

$$\begin{aligned}E[X_t] &= E[Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots]\\ &= E[Z_t] + E[\psi_1 Z_{t-1}] + E[\psi_2 Z_{t-2}] + \cdots\\ &= E[Z_t] + \psi_1 E[Z_{t-1}] + \psi_2 E[Z_{t-2}] + \cdots\\ &= 0\end{aligned}$$

because each of the white noise terms has expected value 0. This is a promising first step toward achieving stationarity. So far, no constraints are needed on the coefficients $\psi_1, \psi_2, \ldots$. That will change when we compute the population variance of X_t. Taking the population variance of both sides of the defining formula results in

$$\begin{aligned} V[X_t] &= V[Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots] \\ &= V[Z_t] + V[\psi_1 Z_{t-1}] + V[\psi_2 Z_{t-2}] + \cdots \\ &= V[Z_t] + \psi_1^2 V[Z_{t-1}] + \psi_2^2 V[Z_{t-2}] + \cdots \\ &= \left(1 + \psi_1^2 + \psi_2^2 + \cdots\right) \sigma_Z^2 \end{aligned}$$

because the white noise terms are mutually independent random variables with common finite population variance σ_Z^2 (see Definition 7.1). Not all values of $\psi_1, \psi_2, \ldots$ will result in a finite population variance of X_t. Setting $\psi_1 = \psi_2 = \cdots = 1$, for example, results in an infinite population variance of X_t. In order to get a finite population variance, the ψ values must decrease in magnitude rapidly enough so that

$$\psi_1^2 + \psi_2^2 + \cdots < \infty.$$

One way to achieve this condition is to have finite values for the first q coefficients $\psi_1, \psi_2, \ldots, \psi_q$ then zeros thereafter. Any general linear model of the first formulation with coefficients that "cut off" in this fashion will satisfy the constraint. Another way of considering this constraint is to write this model using the backshift operator. Using Definition 8.1, the first formulation of the general linear model is

$$X_t = \psi(B) Z_t = \left(1 + \psi_1 B + \psi_2 B^2 + \cdots\right) Z_t.$$

The polynomial in the backshift operator

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$$

will be considered for B values that can assume complex values. So B can have the form $B = a + bi$. The constraint on the coefficients $\psi_1, \psi_2, \ldots$ is equivalent to $\psi(B)$ converging for all B values falling on or inside the unit circle. In other words, $|B| \le 1$.

The population autocovariance function for the general linear model stated in the form

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots$$

with coefficients $\psi_1, \psi_2, \ldots$ satisfying the constraint can be calculated by using the definition of the population covariance:

$$\begin{aligned} \gamma(k) &= \text{Cov}(X_t, X_{t+k}) \\ &= \text{Cov}(Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots,\ Z_{t+k} + \psi_1 Z_{t+k-1} + \psi_2 Z_{t+k-2} + \cdots) \\ &= \text{Cov}(Z_t, \psi_k Z_{t+k-k}) + \text{Cov}\left(\psi_1 Z_{t-1},\ \psi_{k+1} Z_{t+k-(k+1)}\right) + \cdots \\ &= \psi_k \sigma_Z^2 + \psi_1 \psi_{k+1} \sigma_Z^2 + \psi_2 \psi_{k+2} \sigma_Z^2 + \cdots \\ &= (\psi_k + \psi_1 \psi_{k+1} + \psi_2 \psi_{k+2} + \cdots) \sigma_Z^2 \end{aligned}$$

for $k = 1, 2, \ldots$ because of the mutual independence of the terms in the white noise time series. As expected from the previous derivation, the autocovariance at lag 0 is the population variance of X_t:

$$\gamma(0) = V[X_t] = \left(1 + \psi_1^2 + \psi_2^2 + \cdots\right) \sigma_Z^2,$$

where ψ_0, the coefficient of Z_t, equals 1. The associated autocorrelation function is

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\sigma_Z^2 (\psi_k + \psi_1 \psi_{k+1} + \psi_2 \psi_{k+2} + \cdots)}{\sigma_Z^2 \left(1 + \psi_1^2 + \psi_2^2 + \cdots\right)} = \frac{\psi_k + \psi_1 \psi_{k+1} + \psi_2 \psi_{k+2} + \cdots}{1 + \psi_1^2 + \psi_2^2 + \cdots}$$

for $k = 1, 2, \ldots$. Notice that $\rho(0) = 1$ as expected.

The derivation so far has been general, so when specific values of the coefficients $\psi_1, \psi_2, \ldots$ are specified, we now have formulas to determine the population autocovariance function and the population autocorrelation function. Computing these two functions will be illustrated in the next example.

Example 8.3 Consider a time series model $\{X_t\}$ described by

$$X_t = Z_t - \frac{3}{2}Z_{t-1} + \frac{3}{4}Z_{t-2},$$

where $\{Z_t\} \sim WN\left(0, \sigma_Z^2\right)$. Determine whether this time series is stationary and calculate the population autocovariance function and autocorrelation function.

This time series model is a special case of the first formulation of the general linear model from Definition 8.1 which expresses X_t as a linear combination of the white noise terms with coefficients $\psi_1 = -3/2$, $\psi_2 = 3/4$ and $\psi_j = 0$ for $j = 3, 4, \ldots$. The time series is stationary because

$$\psi_1^2 + \psi_2^2 + \cdots = \left(-\frac{3}{2}\right)^2 + \left(\frac{3}{4}\right)^2 = \frac{45}{16} < \infty.$$

The population autocovariance function is

$$\begin{aligned}
\gamma(k) &= (\psi_k + \psi_1\psi_{k+1} + \psi_2\psi_{k+2} + \cdots)\,\sigma_Z^2 \\
&= \begin{cases} \left(1 + (-3/2)^2 + (3/4)^2\right)\sigma_Z^2 & k = 0 \\ \left(-3/2 + (-3/2)(3/4)\right)\sigma_Z^2 & k = 1 \\ (3/4)\,\sigma_Z^2 & k = 2 \\ 0 & k = 3, 4, \ldots \end{cases} \\
&= \begin{cases} 61\sigma_Z^2/16 & k = 0 \\ -21\sigma_Z^2/8 & k = 1 \\ 3\sigma_Z^2/4 & k = 2 \\ 0 & k = 3, 4, \ldots, \end{cases}
\end{aligned}$$

where $\psi_0 = 1$ is the coefficient of Z_t. The associated population autocorrelation function $\rho(k) = \gamma(k)/\gamma(0)$ is

$$\rho(k) = \begin{cases} 1 & k = 0 \\ -42/61 & k = 1 \\ 12/61 & k = 2 \\ 0 & k = 3, 4, \ldots, \end{cases}$$

which is graphed in Figure 8.1. The population autocorrelation function “cuts off” after spikes at lags 1 and 2.

The constraint that has been placed on the values of $\psi_1, \psi_2, \ldots$ can be formalized in this definition of the *causal* representation of the general linear model.

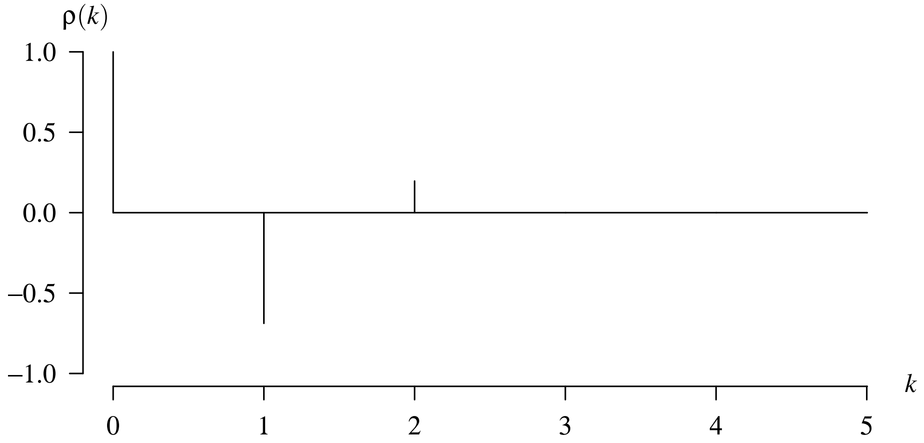

Figure 8.1: Population autocorrelation function for $X_t = Z_t - \frac{3}{2}Z_{t-1} + \frac{3}{4}Z_{t-2}$.

Definition 8.2 A time series $\{X_t\}$ is *causal* if it can be written as

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots,$$

where $\psi_1, \psi_2, \ldots$ are real-valued coefficients that satisfy

$$\psi_1^2 + \psi_2^2 + \cdots < \infty.$$

A time series model that can be written in the causal form is stationary.

The next example illustrates how to convert a general linear model into the causal form in order to establish stationarity.

Example 8.4 Consider the special case of the general linear model

$$\left(1 - \frac{2}{5}B\right) X_t = Z_t.$$

Convert this time series model to the causal representation.

The causal form from Definition 8.2 is

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots.$$

So for the specific case given here,

$$\left(1 - \frac{2}{5}B\right) (Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots) = Z_t.$$

Expanding the left-hand side of this equation gives

$$Z_t + \left(\psi_1 - \frac{2}{5}\right) Z_{t-1} + \left(\psi_2 - \frac{2}{5}\psi_1\right) Z_{t-2} + \left(\psi_3 - \frac{2}{5}\psi_2\right) Z_{t-3} + \cdots = Z_t.$$

Equating the coefficients on the left-hand side and right-hand side of this equation as illustrated in Table 8.1 allows us to solve for $\psi_1, \psi_2, \ldots$. So the causal form of the time series model is

$$X_t = Z_t + \left(\frac{2}{5}\right) Z_{t-1} + \left(\frac{2}{5}\right)^2 Z_{t-2} + \left(\frac{2}{5}\right)^3 Z_{t-3} + \cdots,$$

which has coefficients $\psi_j = (2/5)^j$, for $j = 1, 2, \ldots$. Notice that

$$\psi_1^2 + \psi_2^2 + \psi_3^2 + \cdots = \left(\frac{2}{5}\right)^2 + \left(\frac{2}{5}\right)^4 + \left(\frac{2}{5}\right)^6 + \cdots = \frac{4/25}{1 - 4/25} = \frac{4}{21} < \infty,$$

so the time series is causal because Definition 8.2 is satisfied. Since the time series is causal, this implies that it is also stationary.

term	equation	solution
Z_{t-1}	$\psi_1 - \frac{2}{5} = 0$	$\psi_1 = \frac{2}{5}$
Z_{t-2}	$\psi_2 - \frac{2}{5}\psi_1 = 0$	$\psi_2 = \left(\frac{2}{5}\right)^2$
Z_{t-3}	$\psi_3 - \frac{2}{5}\psi_2 = 0$	$\psi_3 = \left(\frac{2}{5}\right)^3$
$\vdots$	$\vdots$	$\vdots$

Table 8.1: Matching coefficients.

When the second formulation of the general linear model that uses the coefficients $\pi_1, \pi_2, \ldots$ is used, there is an analogous property known as *invertibility* which is defined next. In this case the coefficients $\pi_1, \pi_2, \ldots$ need to decrease in magnitude rapidly enough so that

$$\pi_1^2 + \pi_2^2 + \cdots < \infty.$$

Loosely speaking, a time series model is invertible if there is a one-to-one correspondence between the coefficients $\pi_1, \pi_2, \ldots$ and the associated population autocorrelation function.

Definition 8.3 A time series $\{X_t\}$ is *invertible* if it can be written as

$$X_t = Z_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots,$$

where $\pi_1, \pi_2, \ldots$ are real-valued coefficients that satisfy

$$\pi_1^2 + \pi_2^2 + \cdots < \infty.$$

An invertible time series model has a one-to-one correspondence between the coefficients and the autocorrelation function.

So causality and invertibility are dual properties. Causality indicates that a time series model can be written in the first formulation of the general linear model from Definition 8.1 with coefficients that result in a stationarity model. Invertibility indicates that a time series model can be written in the second formulation of the general linear model from Definition 8.1 with coefficients that ensure a one-to-one correspondence between the coefficients and the population autocorrelation function.

There are three unsettling aspects to the general linear model. First, it only considers *linear* relationships between the X's and the Z's. Situations might arise in which a quadratic term, for example, might be appropriate. Second, the general linear model has an infinite number of parameters: the coefficients $\psi_1, \psi_2, \ldots$ for the first formulation and the coefficients $\pi_1, \pi_2, \ldots$ for the second formulation. ARMA (autoregressive moving average) models, which are special cases of general linear models that are introduced in the next section, limit the number of parameters in the model. The third shortcoming concerns the population mean. Taking the expected value of both sides of the first formulation of the general linear model

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots,$$

for example, gives $E[X_t] = 0$. But the vast majority of real-world time series are not centered around zero. These problems associated with an infinite number of parameters and nonzero mean value will be overcome by the ARMA models introduced in the next section.

8.1.2 An Introduction to ARMA Models

The autoregressive moving average time series model, universally known as the ARMA model, provides two twists on the general linear model. First, the ARMA model limits the number of terms, and therefore limits the number of parameters. Second, the ARMA model includes both types of terms in the two formulations of the general linear model given in Definition 8.1.

There are several reasons for the popularity of the ARMA time series model. First, the population autocorrelation function $\rho(k)$ for an ARMA model can take on a wide variety of shapes, which makes it an appropriate time series model in a wide variety of applications. Second, the ARMA model is parsimonious in the sense that it typically requires only a small number of parameters to achieve an adequate representation of the probability model governing a time series. The notion of parsimony appears in all branches of statistics in which there is interest in finding an approximate probability model using the smallest number of parameters. Third, the ARMA model has been around for several decades, which means that dozens of software packages have been developed over the years for model identification, parameter estimation, forecasting, etc. Although the emphasis here will be on the R language, there are many other software packages that support time series modeling.

The general linear model from Definition 8.1 used the parameters $\psi_1, \psi_2, \ldots$ for the first formulation and $\pi_1, \pi_2, \ldots$ for the second formulation. Of course both of these formulations have the additional parameter σ_Z^2, which is the population variance of the white noise. Tradition dictates that in the conversion from the first formulation of the general linear model to the ARMA model, the Greek letter ψ used for coefficients in the general linear model is replaced by θ, and there are q of these coefficients: $\theta_1, \theta_2, \ldots, \theta_q$. Likewise, in the conversion from the second formulation of the general linear model to the ARMA model, the Greek letter π used for the coefficients in the general model is replaced by ϕ, and there are p of these coefficients: $\phi_1, \phi_2, \ldots, \phi_p$.

So two key parameters in specifying an ARMA model are p and q, which are both nonnegative integers. The parameter p is the number of coefficient parameters in the autoregressive portion of the model; the parameter q is the number of coefficient parameters in the moving average portion of the model. The format for specifying the orders p and q of an ARMA model with p autoregressive terms and q moving average terms is ARMA(p, q).

Definition 8.4 The ARMA(p, q) time series model is

$$X_t = \overbrace{\phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} +}^{\text{autoregressive portion}} \underbrace{Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}}_{\text{moving average portion}},$$

where $\{X_t\}$ is the time series of interest, $\{Z_t\}$ is a time series of white noise, $\phi_1, \phi_2, \ldots, \phi_p$ are real-valued parameters associated with the AR portion of the model, and $\theta_1, \theta_2, \ldots, \theta_q$ are real-valued parameters associated with the MA portion of the model.

The autoregressive portion of this time series model is aptly named because the current value of the time series X_t is regressed on the p previous values of itself. White noise is injected into the model through $\{Z_t\}$ because it is the widest class of the three noise processes from Definition 7.1 which gives the probabilistic properties that are derived in this chapter.

If an ARMA model only involves, for example, the autoregressive portion of the model with two terms (that is, no moving average terms because $\theta_1 = \theta_2 = \cdots = \theta_q = 0$), then this ARMA(2, 0) model is specified as an AR(2) model. Likewise, if an ARMA model only involves, for example, the moving average portion of the model with four terms (that is, no autoregressive terms because $\phi_1 = \phi_2 = \cdots = \phi_p = 0$), then this ARMA(0, 4) model is specified as an MA(4) model. An ARMA(0, 0) model is just a time series of white noise, which was analyzed in Examples 7.9 and 7.15.

The ARMA(p, q) time series model from Definition 8.4 can also be written in terms of the backshift operator B. Taking the original form of the ARMA(p, q) model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q},$$

and separating the autoregressive terms on the left-hand side of the equation and the moving average terms on the right-hand side of the equation results in

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}.$$

This can be written in terms of the backshift operator as

$$X_t - \phi_1 B X_t - \phi_2 B^2 X_t - \cdots - \phi_p B^p X_t = Z_t + \theta_1 B Z_t + \theta_2 B^2 Z_t + \cdots + \theta_q B^q Z_t$$

or

$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) X_t = \left(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q\right) Z_t$$

or more compactly as

$$\phi(B) X_t = \theta(B) Z_t,$$

where the polynomials in B are

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

and

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q,$$

and these are often referred to as the *characteristic polynomials*. This algebra constitutes a proof of the alternative representation of the ARMA(p, q) time series model using polynomials.

Theorem 8.2 The ARMA(p, q) time series model can be written using the backshift operator B as

$$\phi(B)X_t = \theta(B)Z_t,$$

where the characteristic polynomials in B are

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

and

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

Being able to convert between the purely algebraic formulation of an ARMA(p, q) model and the backshift operator formulation is an important skill in time series analysis. The next three examples illustrate how to perform these conversions.

Example 8.5 For the ARMA time series model

$$X_t = 5X_{t-1} - 2X_{t-2} + Z_t - 4Z_{t-1} + 2Z_{t-2} - Z_{t-3},$$

(a) identify the time series model, and

(b) write the time series model in terms of the backshift operator B.

(a) Since there are two terms in the autoregressive portion of the time series model with coefficients

$$\phi_1 = 5 \qquad \text{and} \qquad \phi_2 = -2$$

and three terms in the moving average portion of the time series model with coefficients

$$\theta_1 = -4, \qquad \theta_2 = 2, \qquad \text{and} \qquad \theta_3 = -1,$$

this is an ARMA(2, 3) model.

(b) The time series model

$$X_t = 5X_{t-1} - 2X_{t-2} + Z_t - 4Z_{t-1} + 2Z_{t-2} - Z_{t-3}$$

can be separated into autoregressive and moving average portions as

$$X_t - 5X_{t-1} + 2X_{t-2} = Z_t - 4Z_{t-1} + 2Z_{t-2} - Z_{t-3}.$$

This can be written in terms of B as

$$X_t - 5BX_t + 2B^2X_t = Z_t - 4BZ_t + 2B^2Z_t - B^3Z_t$$

or

$$\left(1 - 5B + 2B^2\right)X_t = \left(1 - 4B + 2B^2 - B^3\right)Z_t.$$

So the polynomials in B that define the coefficients for the ARMA(2, 3) time series model written in the form $\phi(B)X_t = \theta(B)Z_t$ are

$$\phi(B) = 1 - 5B + 2B^2 \qquad \text{and} \qquad \theta(B) = 1 - 4B + 2B^2 - B^3.$$

The previous example converted an ARMA time series model from a purely algebraic formulation to a formulation that uses the backshift operator. The next example goes in the other direction.

Example 8.6 For the ARMA time series model

$$\phi(B)X_t = \theta(B)Z_t,$$

where $\phi(B) = 1 - 0.3B$ and $\theta(B) = 1$,

(a) identify the time series model, and

(b) write the time series model in purely algebraic form.

(a) Since $\phi(B)$ is a first degree polynomial, $p = 1$. Since $\theta(B)$ is a zero degree polynomial, $q = 0$. So this is an ARMA(1, 0) model, which is more commonly referred to as an AR(1) model.

(b) The time series model is

$$(1 - 0.3B)X_t = 1 \cdot Z_t$$

or

$$X_t - 0.3BX_t = Z_t$$

or

$$X_t - 0.3X_{t-1} = Z_t.$$

Isolating X_t on the left-hand side of the equation, the purely algebraic formulation of this AR(1) model with $\phi_1 = 0.3$ is

$$X_t = 0.3X_{t-1} + Z_t.$$

The third and final example of converting between the purely algebraic formulation and backshift formulation of the ARMA(p, q) model would certainly be classified as a trick question. The example emphasizes the importance of looking for common factors between the $\phi(B)$ and $\theta(B)$ polynomials.

Example 8.7 For the ARMA time series model

$$X_t = -3X_{t-1} + X_{t-2} + 3X_{t-3} + Z_t - 3Z_{t-1} - 4Z_{t-2},$$

(a) identify the time series model, and

(b) write the time series model using the backshift operator.

(a) Since there are three terms in the autoregressive portion of the model and two terms in the moving average portion of the model, one might be temped to conclude that this is an ARMA(3, 2) model with autoregressive coefficients

$$\phi_1 = -3, \qquad \phi_2 = 1, \qquad \text{and} \qquad \phi_3 = 3,$$

and moving average coefficients

$$\theta_1 = -3 \qquad \text{and} \qquad \theta_2 = -4.$$

But that conclusion is wrong. It is actually an ARMA(2, 1) model because $\phi(B)$ and $\theta(B)$ have a common factor, as will be seen in part (b). Writing the time series model using the backshift operator B makes it easier to recognize this common factor.

(b) The time series model

$$X_t = -3X_{t-1} + X_{t-2} + 3X_{t-3} + Z_t - 3Z_{t-1} - 4Z_{t-2}$$

can be separated into autoregressive and moving average portions as

$$X_t + 3X_{t-1} - X_{t-2} - 3X_{t-3} = Z_t - 3Z_{t-1} - 4Z_{t-2}$$

or

$$X_t + 3BX_t - B^2X_t - 3B^3X_t = Z_t - 3BZ_t - 4B^2Z_t$$

or

$$\left(1 + 3B - B^2 - 3B^3\right) X_t = \left(1 - 3B - 4B^2\right) Z_t$$

or

$$\phi(B)X_t = \theta(B)Z_t,$$

where

$$\phi(B) = 1 + 3B - B^2 - 3B^3 \qquad \text{and} \qquad \theta(B) = 1 - 3B - 4B^2.$$

The model still looks like an ARMA(3, 2) model. But factoring $\phi(B)$ and $\theta(B)$ reveals that both polynomials contain a common factor:

$$\phi(B) = 1 + 3B - B^2 - 3B^3 = (1 + B)\left(1 + 2B - 3B^2\right)$$

and

$$\theta(B) = 1 - 3B - 4B^2 = (1 + B)(1 - 4B).$$

The common factor $(1 + B)$ in the two polynomials cancels, which means that the ARMA model can be reduced to

$$\phi(B)X_t = \theta(B)Z_t,$$

where

$$\phi(B) = 1 + 2B - 3B^2 \qquad \text{and} \qquad \theta(B) = 1 - 4B,$$

which is an ARMA(2, 1) model. Written in purely algebraic form, this ARMA(2, 1) model is

$$X_t + 2X_{t-1} - 3X_{t-2} = Z_t - 4Z_{t-1},$$

or

$$X_t = -2X_{t-1} + 3X_{t-2} + Z_t - 4Z_{t-1},$$

so the autoregressive coefficients are $\phi_1 = -2$ and $\phi_2 = 3$, and the moving average coefficient is $\theta_1 = -4$.

Based on this example involving a common factor in the $\phi(B)$ and $\theta(B)$ polynomials, we will henceforth assume that the modeler has removed any redundant factors in an ARMA(p, q) time series model. So any ARMA(p, q) model you see going forward will in this sense be presented in lowest terms with no common factors between $\phi(B)$ and $\theta(B)$.

Since an AR(p) model has a finite number of coefficients $\phi_1, \phi_2, \ldots, \phi_p$ in the autoregressive portion of the model, they always satisfy

$$\phi_1^2 + \phi_2^2 + \cdots + \phi_p^2 < \infty,$$

so AR(p) models are always invertible per Definition 8.3. Likewise, since an MA(q) model has a finite number of coefficients $\theta_1, \theta_2, \ldots, \theta_q$ in the moving average portion of the model, they always satisfy

$$\theta_1^2 + \theta_2^2 + \cdots + \theta_q^2 < \infty,$$

so MA(q) models are always stationary per Definition 8.2. In an advanced class in time series, you will prove that an AR(p) model is stationary when all of the p complex roots of the polynomial $\phi(B) = 0$ lie outside of the unit circle in the complex plane. Likewise, an MA(q) model is invertible when all of the q complex roots of the polynomial $\theta(B) = 0$ lie outside of the unit circle in the complex plane. An ARMA(p, q) model is stationary when all of the p complex roots of $\phi(B) = 0$ lie outside of the unit circle in the complex plane. An ARMA(p, q) model is invertible when all of the q complex roots of $\theta(B) = 0$ lie outside of the unit circle in the complex plane. These results are summarized below.

Theorem 8.3 The AR(p) model $\phi(B)X_t = Z_t$ is

- always invertible, and
- stationary when the p roots of $\phi(B) = 0$ lie outside the unit circle in the complex plane.

The MA(q) model $X_t = \theta(B)Z_t$ is

- always stationary, and
- invertible when the q roots of $\theta(B) = 0$ lie outside the unit circle in the complex plane.

The ARMA(p, q) model $\phi(B)X_t = \theta(B)Z_t$ is

- stationary when the p roots of $\phi(B) = 0$ lie outside the unit circle in the complex plane, and
- invertible when the q roots of $\theta(B) = 0$ lie outside the unit circle in the complex plane.

We now revisit the first numeric example of a time series model that we encountered earlier in this chapter to check and see if it is both stationary and invertible.

Example 8.8 Consider the time series model for $\{X_t\}$ that first appeared in Example 8.3 described by

$$X_t = Z_t - \frac{3}{2}Z_{t-1} + \frac{3}{4}Z_{t-2},$$

where $\{Z_t\} \sim WN\left(0, \sigma_Z^2\right)$. Identify this time series model and determine whether it is stationary and invertible.

Since the current and two previous white noise values included in this time series model, this is an MA(2) model. By Theorem 8.3, all MA(2) models are stationary. To see whether this model is invertible, we want to calculate the roots of $\theta(B) = 0$ and see if they lie outside of the unit circle in the complex plane. The purely algebraic form of the time series model

$$X_t = Z_t - \frac{3}{2}Z_{t-1} + \frac{3}{4}Z_{t-2},$$

can be written in terms of the backshift operator as

$$X_t = Z_t - \frac{3}{2}BZ_t + \frac{3}{4}B^2Z_t$$

or

$$X_t = \left(1 - \frac{3}{2}B + \frac{3}{4}B^2\right)Z_t,$$

so $\theta(B) = 1 - \frac{3}{2}B + \frac{3}{4}B^2$. To find the values of B that solve $\theta(B) = 0$ requires solving

$$\frac{3}{4}B^2 - \frac{3}{2}B + 1 = 0,$$

which is equivalent to the quadratic equation

$$3B^2 - 6B + 4 = 0.$$

Using the quadratic formula, the roots of this quadratic equation are

$$B = \frac{6 \pm \sqrt{36 - 48}}{6}$$

or

$$B = 1 \pm \frac{\sqrt{3}}{3}\,i.$$

Since $\theta(B)$ is a second-order polynomial, the complex roots are necessarily complex conjugates. We now need to determine whether these two roots lie outside of the unit circle in the complex plane. There are two ways to proceed. The first is to simply plot these two roots in the complex plane and see if they fall outside of the unit circle. Figure 8.2 shows that the two roots do indeed fall outside of the unit circle. The second way to determine whether the roots fall outside the unit circle is to take the sum of squares of the real and imaginary parts of the roots and see if they exceed 1. In this case,

$$(1)^2 + \left(\frac{\sqrt{3}}{3}\right)^2 = 1 + \frac{1}{3} = \frac{4}{3} > 1.$$

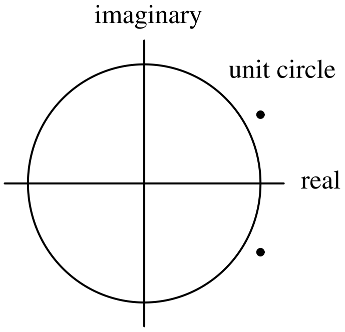

Figure 8.2: Unit circle in the complex plane and the roots of $\theta(B) = 0$.

Both techniques draw the same conclusion: the two roots of $\theta(B) = 0$ fall outside of the unit circle in the complex plane, which means that the time series model is invertible. In conclusion, this MA(2) time series model is both stationary and invertible.

We will get some further practice with these calculations involving the polynomials $\phi(B)$ and $\theta(B)$ when we investigate special cases of the ARMA(p, q) model in more detail in the sections that follow.

Shifted ARMA models

We now address a major shortcoming of the ARMA(p, q) model that–fortunately–is easily overcome. For a stationary ARMA(p, q) model as it has been defined in Definition 8.4, the expected value of X_t is $E\left[X_t\right] = 0$. But most real-world stationary time series are not centered about 0; they are typically centered about some nonzero constant value. The reason that we have waited this long to bring up the topic of a time series centered around a value other than zero is that when we shift the time series, there will be no change in the population autocovariance and autocorrelation functions because population covariance and correlation are unaffected by shifting the time series. The mathematics involved with determining these important functions is much cleaner if you assume that the time series model is centered about zero. There are two ways to tweak the ARMA(p, q) model to allow for it to be centered about some constant value. These two alterations are presented next.

The first way to introduce a nonzero central value for an ARMA(p, q) time series model is to subtract μ from all of the values in the time series. In other words, transform the usual ARMA(p, q) time series model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$$

to the shifted ARMA(p, q) time series model

$$X_t - \mu = \phi_1\left(X_{t-1} - \mu\right) + \phi_2\left(X_{t-2} - \mu\right) + \cdots + \phi_p\left(X_{t-p} - \mu\right) + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}.$$

This can be written compactly in terms of the backshift operator B as

$$\phi(B)\left(X_t - \mu\right) = \theta(B) Z_t,$$

where $\phi(B)$ is the usual polynomial of degree p in B associated with the autoregressive portion of the model:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p,$$

and $\theta(B)$ is the usual polynomial of degree q in B associated with the moving average portion of the model:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q.$$

In this particular formulation of a shifted ARMA(p, q) model, the population mean of the process is $E\left[X_t\right] = \mu$ when the model is stationary. This can be established by taking the expected value of both sides of the shifted ARMA(p, q) time series model.

A second way to formulate a shifted ARMA(p, q) time series model with a nonzero population mean is to simply add a constant, denoted by $\tilde{\mu}$, to the right-hand side of the model:

$$X_t = \tilde{\mu} + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$$

This can be written in terms of the backshift operator as

$$\phi(B) X_t = \tilde{\mu} + \theta(B) Z_t.$$

The reason that a tilde has been placed above μ in this formulation is that $\tilde{\mu}$ is *not* the population mean of the time series model. The two ways of formulating a shifted ARMA(p, q) time series model in these two fashions are summarized as follows.

Definition 8.5 A shifted ARMA(p, q) time series model with a nonzero population mean μ can be written in purely algebraic form as

$$X_t - \mu = \phi_1 (X_{t-1} - \mu) + \phi_2 (X_{t-2} - \mu) + \cdots + \phi_p (X_{t-p} - \mu) + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q},$$

or equivalently using the backshift operator B as

$$\phi(B)(X_t - \mu) = \theta(B) Z_t.$$

A second way to formulate a shifted ARMA(p, q) time series model with a nonzero population mean can be written in purely algebraic form as

$$X_t = \tilde{\mu} + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q},$$

or equivalently using the backshift operator B as

$$\phi(B) X_t = \tilde{\mu} + \theta(B) Z_t,$$

where $\phi(B)$ and $\theta(B)$ are the usual polynomials in the backshift operator B given in Theorem 8.2.

The example that follows illustrates how to convert a shifted time series model from one of these forms to the other.

Example 8.9 The shifted ARMA(1, 1) model defined by

$$X_t = 8 + 0.6 X_{t-1} + Z_t - 0.1 Z_{t-1}$$

is written in the second form from Definition 8.5 with $\tilde{\mu} = 8$. Convert it to the first form.

Moving all autoregressive terms and the constant term to the left-hand side of the equation results in

$$X_t - 0.6 X_{t-1} - 8 = Z_t - 0.1 Z_{t-1}.$$

Using the backshift operator, this can be written as

$$(1 - 0.6B) X_t - 8 = (1 - 0.1B) Z_t.$$

We would like to fold the constant 8 into position on the left-hand side of the equation to match the first formulation from Definition 8.5. We multiply and divide 8 by $(1 - 0.6B)$, keeping in mind that the backshift operator applied to a constant is just the constant:

$$(1 - 0.6B) X_t - 8 \cdot \frac{1 - 0.6B}{1 - 0.6B} = (1 - 0.1B) Z_t$$

or

$$(1 - 0.6B) X_t - (1 - 0.6B) \cdot \frac{8}{0.4} = (1 - 0.1B) Z_t$$

or

$$(1 - 0.6B)(X_t - 20) = (1 - 0.1B) Z_t.$$

So this shifted ARMA(1, 1) time series model is now written in the first formulation from Definition 8.5, which is $\phi(B)(X_t - \mu) = \theta(B)$. The expected value of X_t is $\mu = E[X_t] = 20$. One way to check that we have done all of the algebra correctly is to use $\mu = 20$ as an argument in the first formulation of the model from Definition 8.5 and perform the algebra to see whether it is equivalent to the second formulation.

The previous example can be generalized from the shifted ARMA(1, 1) model to the shifted ARMA(p, q) model. The following theorem gives the relationship between μ and $\tilde{\mu}$ for the two formulations of the shifted ARMA(p, q) models in Definition 8.5.

Theorem 8.4 The parameters $\mu = E[X_t]$ and $\tilde{\mu}$ for the two shifted ARMA(p, q) models from Definition 8.5 are related by

$$\mu = \frac{\tilde{\mu}}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}$$

when the coefficients $\phi_1, \phi_2, \ldots, \phi_p$ correspond to a stationary model.

Proof The second shifted ARMA(p, q) model from Definition 8.5 is

$$X_t = \tilde{\mu} + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}.$$

Taking the expected value of both sides of this equation yields

$$E[X_t] = \tilde{\mu} + \phi_1 E[X_{t-1}] + \phi_2 E[X_{t-2}] + \cdots + \phi_p E[X_{t-p}] + 0$$

because all of the white noise terms have expected value zero. Since the time series is assumed to be stationary, $E[X_t] = E[X_{t-1}] = E[X_{t-2}] = \cdots = E[X_{t-p}]$, and this equation becomes

$$E[X_t] = \tilde{\mu} + \phi_1 E[X_t] + \phi_2 E[X_t] + \cdots + \phi_p E[X_t].$$

Solving for $\mu = E[X_t]$ gives

$$\mu = \frac{\tilde{\mu}}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}. \qquad \square$$

In the previous example, the value of $\mu = E[X_t]$ could have been calculated by appealing to Theorem 8.4 with $\tilde{\mu} = 8$ and $\phi_1 = 0.6$, which gives

$$\mu = E[X_t] = \frac{8}{1 - 0.6} = 20.$$

This provides an illustration of how Theorem 8.4 provides a mechanism for converting between the two forms of the shifted ARMA(p, q) models given in Definition 8.5.

This section has provided an introduction to linear models. The first subsection surveyed the two formulations of the general linear model and introduced the causality and invertibility properties. The second subsection introduced a special case of the general linear model known as the ARMA (autoregressive moving average) model. These time series models are inherently probabilistic in nature. The next section introduces some of the associated statistical topics in time series analysis: parameter estimation, forecasting, model assessment, and model selection.

8.2 Statistical Methods

The previous section introduced two linear probability models for time series: the general linear model and the ARMA model. These models contain parameters which can be used to tune the model to a particular application. This chapter introduces the statistical methods that are used to estimate these parameters and assess whether the model with its fitted parameters provides an adequate representation of the probabilistic mechanism governing the time series. As you read the rest of this book, you should be continually asking yourself whether the new material is associated with a probability model or presents a statistical method. The statistical methods are presented here in a somewhat generic manner; the specific implementations on a time series of observations occurs subsequently. The first subsection in this section introduces three methods for estimating the parameters in an ARMA model: the method of moments, least squares, and maximum likelihood. This is followed by a subsection that considers the important topic of forecasting future observations in a time series. Subsections on model assessment and model selection complete the section.

8.2.1 Parameter Estimation

The emphasis now shifts from a time series model, which is developed using probability theory, to statistical questions associated with a realization of a time series. The observed values of this realization are denoted by $X_1, X_2, \ldots, X_n$ when considered abstractly; when specific values are considered, they are denoted by $x_1, x_2, \ldots, x_n$.

Before considering parameter estimation, we consider the topic of *model identification*. Since p and q are nonnegative integers, there are an infinite number of ARMA(p, q) models from which to choose. Which model is appropriate for a particular application? Most statistical software packages that perform the analysis of a time series have functions that estimate parameters and forecast future values of the time series. So those two aspects of time series analysis are largely automated. The part of the process that requires some insight from the modeler is the specification of an appropriate time series model for a particular application. By what criteria do we decide whether an MA(1), AR(2), or ARMA(2, 1) is a tentative or a final time series model? The two steps associated with model identification for an ARMA(p, q) model are given next.

1. **Inspect the time series plot**. The process of identifying a time series model always begins with a careful inspection of a plot of the time series. Take a few minutes to look for cyclic variation, trends, step changes, outliers, and nonconstant variance in the plot of the time series. Visually assess the time series for any serial correlation. The human eye can spot subtleties that an algorithm might miss. Only you can perform this step. We assume for now that no trends, step changes, outliers, cyclic variation, or nonconstant variance in the time series have been identified, so a stationary model for the time series is sought. Modeling cyclic variation, trends, and nonconstant variance will be taken up subsequently.

2. **Inspect the plots of r_k and r_k^***. Inspecting plots of the sample autocorrelation function and the sample partial autocorrelation function is an attempt to conduct a visual pattern match between the sample autocorrelation patterns with a known inventory of population autocorrelation patterns for the various ARMA(p, q) models. The minimum length of a time series in order to make meaningful visual comparisons between the sample and population autocorrelation functions is about $n = 60$ or $n = 70$ observations. As will be seen in subsequent chapters, the shape of the sample autocorrelation function and the sample partial autocorrelation function can provide clues as to an appropriate time series model. In some cases, the values of p and q in the ARMA(p, q) model become immediately apparent upon viewing these

three plots. In other cases, the situation is murky, and there might be two or three potential ARMA(p, q) models that seem to be plausible. Since we have assumed that the time series is stationary in the previous paragraph, there is no need to transform or difference the data based on these plots in the current setting. The p and q values for the ARMA time series model identified from this step will be known as the *tentative model*. Once a tentative model has been identified, the next step is to estimate the parameters, which accounts for the remainder of this section.

We would like to estimate the parameters of a stationary and invertible tentative ARMA(p, q) model. It is assumed that the number of autoregressive terms p and the number of moving average terms q have been established for a tentative ARMA(p, q) time series model based on an inspection of the sample autocorrelation and sample partial autocorrelation functions. There are a total of $p+q+1$ unknown parameters in a standard ARMA(p, q) model from Definition 8.4: the autoregressive coefficients $\phi_1, \phi_2, \ldots, \phi_p$, the moving average coefficients $\theta_1, \theta_2, \ldots, \theta_q$, and the population variance of the white noise σ_Z^2. The shifted ARMA(p, q) model from Definition 8.5 has the additional parameter μ.

Consistent with conventional notation in statistics, hats on unknown parameters denote their point estimators. The point estimator of the unknown parameter ϕ_1, for example, is $\hat{\phi}_1$. The point estimators developed here are random variables that take on one particular value for an observed time series $x_1, x_2, \ldots, x_n$. Point estimators are typically paired with a $100(1-\alpha)$% confidence interval that gives a sense of the precision of the point estimator. A confidence interval for the unknown parameter ϕ_1, for example, is typically expressed in the form $L < \phi_1 < U$, where L is the random lower bound of the confidence interval and U is the random upper bound of the confidence interval.

In most practical problems involving a time series model, a shifted ARMA(p, q) model is used because very few time series are centered around zero. Since the ARMA(p, q) time series model is generally assumed to be stationary and invertible, it is common practice in time series analysis to estimate the population mean parameter μ with the sample mean $\bar{X}$. This is justified by the fact that $E[X_t] = \mu$ for a stationary and invertible shifted ARMA(p, q) model. This is consistent with the method of moments approach. Once μ has been estimated, the new time series which is shifted by $\hat{\mu} = \bar{X}$ is

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}.$$

This time series can be fitted to a standard ARMA(p, q) model from Definition 8.4. This new time series has a sample mean value of zero because

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}\bar{x} = \bar{x} - \frac{1}{n}\cdot n \cdot \bar{x} = 0.$$

So for now we dispatch with the parameter μ and assume that it will typically be estimated by $\bar{x}$ for a stationary and invertible ARMA(p, q) model by centering the time series as described above. Both the original time series and the centered time series will be denoted by as $\{X_t\}$ or $\{x_t\}$ in order to avoid introducing a new letter (Y_t or y_t) into the notation. The parameter estimation techniques that follow will be applied to a standard ARMA(p, q) model centered around zero, which assumes that μ has been estimated in the shifted model. This will make the notation somewhat more compact. The population variance of $\bar{X}$ for mutually independent and identically distributed observations $X_1, X_2, \ldots, X_n$ is the well-known formula

$$V\left[\bar{X}\right] = \frac{\sigma_X^2}{n}.$$

But for a stationary ARMA(p, q) time series model with population autocovariance function $\gamma(k)$ and population autocorrelation function $\rho(k)$, the population variance of the sample mean is

$$\begin{aligned}
V\left[\bar{X}\right] &= V\left[\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)\right] \\
&= \frac{1}{n^2} V\left[X_1 + X_2 + \cdots + X_n\right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}\left(X_i, X_j\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n} V\left[X_i\right] + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{Cov}\left(X_i, X_j\right)\right] \\
&= \frac{1}{n^2}\left[n\gamma(0) + 2\sum_{k=1}^{n-1}(n-k)\gamma(k)\right] \\
&= \frac{\sigma_X^2}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho(k)\right].
\end{aligned}$$

Notice that this formula collapses to $V\left[\bar{X}\right] = \sigma_X^2/n$ when $\rho(1) = \rho(2) = \cdots = \rho(n-1) = 0$ as expected. This formula should be kept in mind whenever statistical inferences, such as confidence intervals or hypothesis tests, are made concerning the population mean from a realization of a time series. The sample mean is a meaningful summary statistic for a time series only when appropriate transformations have been applied to the time series in order to reduce it to a stationary time series.

Three techniques for the estimation of parameters in a time series model will be introduced here: the method of moments, least squares, and maximum likelihood estimation. There are three reasons why just one parameter estimation technique is not adequate. First, an AR(3) model, for example, might be well fitted with one estimation technique, but an MA(2) model, on the other hand, might be more compatible with another estimation technique. Second, it is often the case that one technique will provide initial estimates for a numerical method associated with a second technique. Third, some of the estimation techniques provide estimators which have degraded statistical properties near the boundaries of the stationarity or invertibility regions. The three techniques will be discussed generally below, and then will be illustrated with examples subsequently using real time series data.

Method of Moments

The essence of the method of moments technique is to equate low-order population and sample moments and solve for all unknown parameters. This method was developed by English mathematician and biostatistician Karl Pearson. This approach often seems arresting to those encountering it for the first time because population moments are constants and sample moments are random variables. Equating constants and random variables is simply a device that is used to get a perfect match between low-order population and sample moments.

In a non-time-series context with data values $X_1, X_2, \ldots, X_n$ and m unknown population parameters, the m equations

$$E\left[X_t\right] = \frac{1}{n}\sum_{t=1}^{n} X_t$$

$$E\left[X_t^2\right] = \frac{1}{n}\sum_{t=1}^{n} X_t^2$$

$$\vdots$$

$$E\left[X_t^m\right] = \frac{1}{n}\sum_{t=1}^{n} X_t^m$$

can be solved to arrive at the *m* method of moments estimators of the unknown parameters. In some settings this can be done analytically, but in other settings numerical methods are required.

Returning to a time-series context, the stationarity assumption (see Definition 7.6) places requirements on only the first two population moments $E[X_t]$ and $E\left[X_t^2\right]$. Stationarity places no requirements on the third and higher order moments. But stationarity does imply that the autocorrelation between two observations depends only on the lag, and this can be exploited to generate the necessary number of equations to employ the method of moments technique. Consider a stationary and invertible ARMA(p, q) model, for example, that has four unknown parameters. Solving the set of four equations in the four unknown parameters

$$\begin{aligned}
E\left[X_t\right] &= \frac{1}{n}\sum_{t=1}^{n} X_t \\
E\left[X_t^2\right] &= \frac{1}{n}\sum_{t=1}^{n} X_t^2 \\
\rho(1) &= r_1 \\
\rho(2) &= r_2
\end{aligned}$$

yields the method of moments estimators for the four unknown parameters. The usual approach to fitting a time series model to a realization of a time series by the method of moments technique is to use the first two of these equations, and then equate population and sample autocorrelations at enough low-order lags in order to account for all unknown parameters. In this way the population and the sample autocorrelations will match at lower-order lags.

Least Squares Estimation

The least squares estimation technique is used nearly universally in regression analysis. This method developed by German mathematician Carl Friedrich Gauss. The essence of the least squares technique is to find the values of the unknown parameters that minimize the sum of squares of the error terms in a model. In the time series setting, we want to find the values of the parameters that minimize

$$S = \sum_{t=1}^{n} Z_t^2.$$

The use of least squares for ARMA(p, q) models requires two steps. First, solve the target model for Z_t, and then substitute that expression into the equation above. At this point, S is written in terms of the unknown parameters. Second, take the partial derivatives of S with respect to all unknown parameters and solve for the unknown parameters. The set of equations to solve is often referred to as the *orthonormal equations*. The solution to these equations yields the least squares estimates of the unknown parameters. In some cases these equations can be solved analytically; in other cases numerical methods are required.

Maximum Likelihood Estimation

Maximum likelihood estimation is the most prevalent technique for estimating unknown parameters from a data set in the field of statistics, particularly outside of regression. The method was popularized by English statistician Sir Ronald Fisher. The essence of the maximum likelihood estimation technique, whether applied in time series analysis or otherwise, is to select the parameters in a hypothesized model that are the most likely ones to have resulted in the observed data values. The *maximum likelihood estimators* of the unknown parameters are found by maximizing the *likelihood function*, which is the joint probability density function of the data values evaluated at their observed values. The likelihood function is a function of the unknown parameters in the model with the data values fixed at their observed values. We begin by using maximum likelihood estimation on an ARMA(0, 0) model in order to establish some of the issues associated with the use of the maximum likelihood estimation technique to estimate the parameters in a time series model.

Example 8.10 Let $x_1, x_2, \ldots, x_n$ be a realization of observations from an ARMA(0, 0) time series model that is simply white noise:

$$X_t = Z_t,$$

where $Z_t \sim WN\left(0, \sigma_Z^2\right)$. Find the maximum likelihood estimator of σ_Z^2, determine whether the maximum likelihood estimator is unbiased and consistent, and derive an exact two-sided $100(1-\alpha)\%$ confidence interval for σ_Z^2.

The ARMA(0, 0) time series model has just a single unknown parameter σ_Z^2, the population variance of the white noise, that needs to be estimated. The likelihood function is the joint probability density function of the observations:

$$L\left(\sigma_Z^2\right) = f(x_1, x_2, \ldots, x_n).$$

The $x_1, x_2, \ldots, x_n$ arguments on L and the σ_Z^2 argument on f are suppressed for brevity. We are lucky with the ARMA(0, 0) model because we can exploit the fact that the observations in the time series are mutually independent, which means that the joint probability density function of the observed values $x_1, x_2, \ldots, x_n$ is the product of the marginal probability density functions:

$$L\left(\sigma_Z^2\right) = f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n),$$

where $f(x)$ is the probability density function of a single observation in the time series, which is just white noise. We won't be so lucky for general ARMA(p, q) models. The assumption of white noise is vague in the sense that we do not know the functional form of $f(x)$. We only know that it is a probability distribution with population mean 0 and population variance σ_Z^2. In order to apply the maximum likelihood estimation technique, we must make an additional assumption about the distribution of $X_1, X_2, \ldots, X_n$. So at this point we make the additional assumption that the white noise terms are in fact Gaussian white noise terms:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma_Z^2}}\, e^{-x_i^2/\left(2\sigma_Z^2\right)} \qquad -\infty < x_i < \infty,$$

for $i = 1, 2, \ldots, n$, which is the probability density function of a $N\left(0, \sigma_Z^2\right)$ random variable. The assumption of normally-distributed error terms in order to use the maximum likelihood estimation technique is nearly universal in time series analysis. The associated likelihood function is

$$L\left(\sigma_Z^2\right) = \prod_{i=1}^{n} f(x_i) = \left(2\pi\sigma_Z^2\right)^{-n/2} e^{-\sum_{i=1}^{n} x_i^2/\left(2\sigma_Z^2\right)}.$$

The maximum likelihood estimator of σ_Z^2 is the value of σ_Z^2 that maximizes the likelihood function:

$$\hat{\sigma}_Z^2 = \underset{\Omega}{\operatorname{argmax}}\, L\left(\sigma_Z^2\right),$$

where Ω is the parameter space $\Omega = \left\{\sigma_Z^2 \,|\, \sigma_Z^2 > 0\right\}$. It is often the case that the mathematics associated with maximizing the natural logarithm of the likelihood function is easier than the mathematics of maximizing the likelihood function. Both functions are maximized at the same value because the natural logarithm is a monotonic transformation. The log likelihood function is

$$\ln L\left(\sigma_Z^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma_Z^2\right) - \frac{1}{2\sigma_Z^2}\sum_{i=1}^{n} x_i^2.$$

The derivative of the log likelihood function with respect to the unknown parameter σ_Z^2 is

$$\frac{\partial \ln L\left(\sigma_Z^2\right)}{\partial \sigma_Z^2} = -\frac{n}{2\sigma_Z^2} + \frac{1}{2\sigma_Z^4}\sum_{i=1}^{n} x_i^2.$$

Equating this derivative to zero and solving for σ_Z^2 gives the maximum likelihood estimator

$$\hat{\sigma}_Z^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

The maximum likelihood estimator is an unbiased estimator of σ_Z^2 because

$$E\left[\hat{\sigma}_Z^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} V\left[X_i\right] = \frac{1}{n}\cdot n\cdot\sigma_Z^2 = \sigma_Z^2$$

based on the shortcut formula for the population variance and the fact that $E[X_i] = 0$. This means that although the maximum likelihood estimator might miss the true parameter value σ_Z^2 on the low side or on the high side, it is pointing at the correct target because its expected value (long-run average) is the true parameter value.

By standardizing the X_i values, we find that a function of the maximum likelihood estimator has the chi-square distribution because it can be written as the sum of squares of mutually independent standard normal random variables:

$$\frac{n\hat{\sigma}_Z^2}{\sigma_Z^2} = \sum_{i=1}^{n}\left(\frac{X_i - 0}{\sigma_Z}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i}{\sigma_Z}\right)^2 \sim \chi^2(n).$$

The population variance of the maximum likelihood estimator is

$$V\left[\hat{\sigma}_Z^2\right] = \frac{\sigma_Z^4}{n^2}\cdot V\left[\frac{n\hat{\sigma}_Z^2}{\sigma_Z^2}\right] = \frac{\sigma_Z^4}{n^2}\cdot 2n = \frac{2\sigma_Z^4}{n}$$

because the population variance of a chi-square random variable with n degrees of freedom is $2n$. The maximum likelihood estimator is a consistent estimator of σ_Z^2 because it is unbiased and $\lim_{n\to\infty} V\left[\hat{\sigma}_Z^2\right] = 0$. The maximum likelihood estimator $\hat{\sigma}_Z^2$ will approach the true parameter value σ_Z^2 in the limit as n increases. In other words, for any positive constant ε,

$$\lim_{n\to\infty} P\left(\left|\hat{\sigma}_Z^2 - \sigma_Z^2\right| < \varepsilon\right) = 1.$$

The unbiased and consistent point estimator $\hat{\sigma}_Z^2$ does not convey any sense of the precision of the point estimator, however. That information is best conveyed in this setting by a confidence interval. An appropriate pivotal quantity is

$$\frac{n\hat{\sigma}_Z^2}{\sigma_Z^2} \sim \chi^2(n),$$

which implies that

$$\chi^2_{n,\,1-\alpha/2} < \frac{n\hat{\sigma}_Z^2}{\sigma_Z^2} < \chi^2_{n,\,\alpha/2}$$

with probability $1-\alpha$. The second subscript on the quantile of the chi-square distribution is a right-hand tail probability. Performing the algebra required to isolate σ_Z^2 in the center of the inequality results in the exact two-sided $100(1-\alpha)\%$ confidence interval

$$\frac{n\hat{\sigma}_Z^2}{\chi^2_{n,\,\alpha/2}} < \sigma_Z^2 < \frac{n\hat{\sigma}_Z^2}{\chi^2_{n,\,1-\alpha/2}}.$$

Common values for α are 0.1, 0.05, and 0.01, which are known as 90%, 95%, and 99% confidence intervals, respectively. The proper interpretation of a confidence interval like this one is critical. An *incorrect* interpretation of this exact confidence interval for, say, $\alpha = 0.05$, is:

> "The probability that this confidence interval contains σ_Z^2 is 0.95"

because once the data has been collected and the interval is calculated, it either contains the unknown parameter σ_Z^2 or it does not. A probability statement like this one does not make sense because there are no random variables after the data values are collected. The correct interpretation of this exact confidence interval for σ_Z^2 with nominal coverage 0.95 is as follows.

> "The confidence interval I have calculated might contain σ_Z^2 or it might not. However, if (a) all of the assumptions that I have made concerning the ARMA(0, 0) time series model with Gaussian white noise are correct, (b) many realizations of the time series of size n are collected, and (c) the same procedure was used for calculating a confidence interval for each of the realizations, then 0.95 is the expected fraction of these confidence intervals that will contain the true parameter σ_Z^2."

Obviously, one would not want to repeat this tedious explanation every time a confidence interval is calculated. So statisticians shorten this by simply saying:

> "I am 95% confident that my confidence interval contains the unknown parameter σ_Z^2."

The brevity and avoidance of the use of "probability" in this statement aids the proper interpretation of the confidence interval.

Finally, we consider an application area in which the ARMA(0, 0) might be appropriate. The ARMA(0, 0) model has industrial applications in quality control. When formulating a model for a continuous measurement associated with a product (such as a ball bearing diameter or the pre-cooked weight of a quarter-pound hamburger) that

> is produced repeatedly over time, management prefers a stationary time series model with mutually independent consecutive observations. In this particular setting, a shifted ARMA(0, 0) is appropriate and justified. This model is used in practice to help detect when the continuous measurement trends away from the mean value in a shifted ARMA(0, 0) time series model in what is known in quality control as a *control chart*.

Applying the maximum likelihood estimation technique to the ARMA(0, 0) time series model was ideal in that the point estimator for σ_Z^2 could be expressed in closed form and an exact two-sided confidence interval for σ_Z^2 could be derived to give an indication of the precision of the point estimator. There are three key take-aways from the ARMA(0, 0) example involving maximum likelihood estimation.

- We needed to narrow the assumption of white noise error terms to Gaussian white noise error terms in order to implement the maximum likelihood estimation technique.
- We were fortunate that the likelihood function could be factored into the product of the marginal probability density functions because of the mutual independence of the observations. This will not be the case with the ARMA(p, q) model with $p > 0$ and/or $q > 0$.
- We were fortunate in the sense that we could establish an exact two-sided $100(1-\alpha)\%$ confidence interval for σ_Z^2 based on a pivotal quantity. For ARMA(p, q) models with $p > 0$ and/or $q > 0$ we will generally have only approximate confidence intervals which are based on asymptotic results.

We now address the third take-away concerning confidence intervals for parameters in ARMA models that go beyond the ARMA(0, 0) model illustrated in the previous example. The mathematics associated with deriving the exact distribution of some pivotal quantity becomes too difficult once autocorrelation is injected into a model, so we use asymptotic results concerning the parameter estimates in order to arrive at approximate confidence intervals. To frame the conversation concerning these asymptotic results, some notation must be established. Let

$$\beta = (\beta_1, \beta_2, \ldots, \beta_r)'$$

be a vector that denotes the r unknown parameters in a time series model. In the case of a shifted ARMA(p, q) model, for example, the elements of β are the $p+q+2$ unknown parameters ϕ_1, ϕ_2, ..., ϕ_p, θ_1, θ_2, ..., θ_q, μ, and σ_Z^2. Let $x_1, x_2, \ldots, x_n$ denote a realization of the time series observations. The likelihood function is

$$L(\beta) = f(x_1, x_2, \ldots, x_n)$$

and the associated log likelihood function is

$$\ln L(\beta) = \ln f(x_1, x_2, \ldots, x_n).$$

The jth element of the score vector is

$$\frac{\partial \ln L(\beta)}{\partial \beta_j}$$

for $j = 1, 2, \ldots, r$. Equating the elements of the score vector to zero and solving for the unknown parameters yields the maximum likelihood estimators $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_r$. The (j, k) element of the Fisher information matrix $I(\beta)$ is

$$E\left[-\frac{\partial^2 \ln L(\beta)}{\partial \beta_j \partial \beta_k}\right]$$

for $j = 1, 2, \ldots, r$ and $k = 1, 2, \ldots, r$, when the expected values exist. The Fisher information matrix is estimated by the observed information matrix $O\big(\hat{\beta}\big)$, whose (j, k) element is

$$\left[-\frac{\partial^2 \ln L(\beta)}{\partial\beta_j \partial\beta_k}\right]_{\beta=\hat{\beta}}$$

for $j = 1, 2, \ldots, r$ and $k = 1, 2, \ldots, r$. The inverse of the observed information matrix is the asymptotic variance–covariance matrix of the parameter estimates. If one is willing to ignore the off-diagonal elements of this matrix, the square roots of the diagonal elements are estimates of the standard errors of the point estimators. The asymptotic normality of maximum likelihood estimators allows one to construct approximate confidence intervals for the unknown parameters.

We were able to obtain an exact two-sided confidence interval for σ_Z^2 for the ARMA(0, 0) model in the previous example; the next example goes through the appropriate steps for the model had we not been so lucky. We return to the analysis of the standard ARMA(0, 0) time series model because it is the only ARMA(p, q) model with a single unknown parameter and associated tractable mathematics.

Example 8.11 Find an asymptotically exact two-sided $100(1-\alpha)\%$ confidence interval for σ_Z^2 for an ARMA(0, 0) model based on the asymptotic normality of the maximum likelihood estimator $\hat{\sigma}_Z^2$. Estimate the actual coverage of this confidence interval for $n = 100$, $\sigma_Z^2 = 1$, and $\alpha = 0.05$. What is the impact of n on the actual coverage?

Although we know that there is an exact confidence interval for σ_Z^2 from the previous example, we pretend that we are unaware of such an interval and try to find an asymptotically exact interval based on the inverse of the observed information matrix. This is done to illustrate the mechanics of constructing the asymptotically exact confidence interval. From Example 8.10, the maximum likelihood estimator of σ_Z^2 is

$$\hat{\sigma}_Z^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 .$$

Once again treating σ_Z^2 as a unit, the second partial derivative of the log likelihood function with respect to σ_Z^2 is

$$\frac{\partial^2 \ln L\left(\sigma_Z^2\right)}{\partial\left(\sigma_Z^2\right)^2} = \frac{n}{2\sigma_Z^4} - \frac{1}{\sigma_Z^6}\sum_{i=1}^{n} x_i^2 .$$

The single entry in the 1×1 Fisher information matrix is the expected value of the negative of this partial derivative:

$$I\left(\sigma_Z^2\right) = E\left[-\frac{\partial^2 \ln L\left(\sigma_Z^2\right)}{\partial\left(\sigma_Z^2\right)^2}\right] = -\frac{n}{2\sigma_Z^4} + \frac{1}{\sigma_Z^6}\sum_{i=1}^{n} V\left[X_i\right] = \frac{n}{2\sigma_Z^4} .$$

Since σ_Z^2 is an unknown parameter, the Fisher information matrix cannot be determined from the observations from a time series. The 1×1 observed information matrix provides an estimate of the Fisher information matrix from the data values:

$$O\left(\hat{\sigma}_Z^2\right) = \left[-\frac{\partial^2 \ln L\left(\sigma_Z^2\right)}{\partial\left(\sigma_Z^2\right)^2}\right]_{\sigma_Z^2=\hat{\sigma}_Z^2} = -\frac{n}{2\hat{\sigma}_Z^4} + \frac{1}{\hat{\sigma}_Z^6}\sum_{i=1}^{n} x_i^2 = \frac{n^3}{2\left(\sum_{i=1}^{n} x_i^2\right)^2} .$$

The inverse of this 1 × 1 matrix is just the reciprocal of the single entry:

$$O^{-1}\left(\hat{\sigma}_Z^2\right) = \frac{2\left(\sum_{i=1}^{n} x_i^2\right)^2}{n^3}.$$

For large values of n, this quantity converges to the variance of $\hat{\sigma}_Z^2$. So since

$$\hat{\sigma}_Z^2 \xrightarrow{D} N\left(\sigma_Z^2, \frac{2\left(\sum_{i=1}^{n} x_i^2\right)^2}{n^3}\right),$$

an asymptotically exact $100(1-\alpha)\%$ confidence interval for σ_Z^2 is

$$\hat{\sigma}_Z^2 - z_{\alpha/2}\sqrt{\frac{2\left(\sum_{i=1}^{n} x_i^2\right)^2}{n^3}} < \sigma_Z^2 < \hat{\sigma}_Z^2 + z_{\alpha/2}\sqrt{\frac{2\left(\sum_{i=1}^{n} x_i^2\right)^2}{n^3}}.$$

We know that the actual coverage of this two-sided confidence interval converges to the exact coverage as $n \to \infty$. But how does the confidence interval perform for finite values of n? This can only be assessed by a Monte Carlo simulation experiment.

The Monte Carlo simulation given by the R code below simulates four million time series of length $n = 100$ generated from an ARMA(0, 0) model with Gaussian white noise having variability $\sigma_Z^2 = 1$ and estimates the actual coverage of the approximate 95% confidence interval by printing the fraction of the simulated confidence intervals that contain the arbitrarily-assigned true parameter value $\sigma_Z^2 = 1$.

```
nrep  = 4000000
count = 0
n     = 100
alpha = 0.05
crit  = qnorm(1 - alpha / 2)
for (i in 1:nrep) {
  x   = rnorm(n)
  ssq = sum(x ^ 2)
  mle = ssq / n
  std = sqrt(2 * ssq ^ 2 / n ^ 3)
  lo  = mle - crit * std
  hi  = mle + crit * std
  if (lo < 1 && hi > 1) count = count + 1
}
print(count / nrep)
```

After a call to `set.seed(3)` to establish the random number stream, five runs of this simulation yield the following estimated confidence interval coverages:

0.9402 0.9399 0.9400 0.9401 0.9401.

Although the stated (or nominal) coverage for this confidence interval is 0.95, the Monte Carlo simulation reveals that the actual coverage is 0.940.

The final question concerns the impact of n on the actual coverage. The Monte Carlo simulation experiment given above is executed for several other values of n. The actual coverage values are shown in Figure 8.3. These values confirm what we suspect about an asymptotic confidence interval: the actual coverage asymptotically approaches the stated coverage (indicated by the dashed horizontal line in Figure 8.3). This behavior is typical of asymptotic confidence intervals.

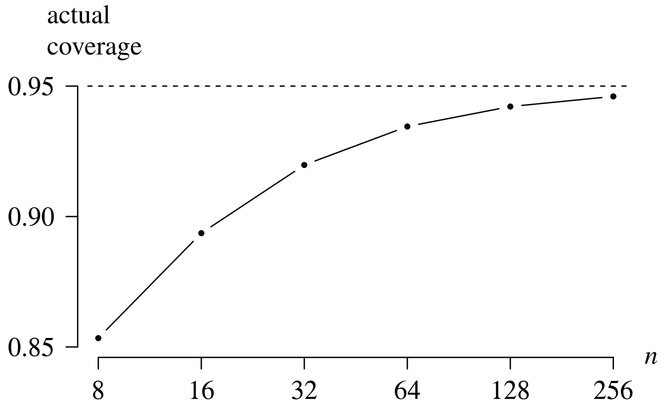

Figure 8.3: Asymptotic 95% confidence interval actual coverage for $n = 8, 16, 32, \ldots, 256$.

This ends the discussion of the important topic of parameter estimation. The time series model that emerges from this step is known as a *fitted tentative model*. Three techniques for parameter estimation have been introduced: the method of moments, least squares, and maximum likelihood estimation. In time series analysis, exact confidence intervals for the unknown parameters are typically mathematically intractable, so we must settle for asymptotically exact confidence intervals.

The next section introduces another important statistical topic that arises frequently in time series analysis: the prediction of future values in a time series based on a realization of n observations of a time series, which is typically known as *forecasting*.

8.2.2 Forecasting

The purpose of forecasting is to predict one or more future values of a time series based on observed values of a time series $x_1, x_2, \ldots, x_n$. Forecasting future values of a time series often plays a critical role in policy decisions. The closing price of the Dow Jones Industrial Average tomorrow, the number of oysters in the Chesapeake Bay next year, the high temperature in Tuscaloosa on Saturday, and a company's profit next quarter are examples of applications of forecasting.

The term "forecasting" is synonymous with "prediction" and the two terms will be used interchangeably. Forecasting is a slightly more popular term in the time series literature. Both terms can be interpreted as "telling before."

Forecasting involves extrapolation of the time series model outside of the time frame associated with the observed values $x_1, x_2, \ldots, x_n$, typically into the future. The notion of backcasting, which is predicting values in the past, will not be considered here. Care must be taken to ensure that the fitted probability model still applies in the time range in which the extrapolation occurs. If future observations are governed by the same probability model as previous observations, then a forecasted value is meaningful. Furthermore, if an ARMA(p, q) model is used, it is subject to errors in identification (for example, the wrong values of p and q or perhaps an ARMA model is used when a non-ARMA model is appropriate) and estimation (for example, due to random sampling variability or choosing an inferior parameter estimation procedure).

There are several choices for forecasting notation. We assume that the values of a time series $\{X_t\}$ are given by the observed values $x_1, x_2, \ldots, x_n$. We would like to predict the value of the time series h (for "horizon") time units into the future, given that we know the values of $x_1, x_2, \ldots, x_n$ and our forecast is being made at time n. The notation that we will use for this future value of the time series will be the random variable X_{n+h}. Its associated predicted value will be denoted by $\hat{X}_{n+h}$. This predicted value is defined as the conditional expected value of the future value given the values of the n observed values:

$$\hat{X}_{n+h} = E\left[X_{n+h} \,|\, X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right].$$

We will use the alternative notation $\hat{X}_n(h)$ for the forecast whenever there might be some ambiguity associated with the origin of the forecast. The default assumption for forecasting in this book is that we are making a forecast based on n observed values, and the forecast is being made at time origin n for h time units into the future. The forecasted value at time $n+h$ can be thought of as the average of all future possibilities given the history up to time n. But why use the conditional expectation? Might a quantile of the probability distribution of X_{n+h}, for example, the population median, provide a better forecast? The rationale behind using the conditional expectation is that it minimizes the mean square error of the predicted value, which is defined as

$$E\left[\left(X_{n+h} - \hat{X}_{n+h}\right)^2\right],$$

among all linear functions of the observed values $x_1, x_2, \ldots, x_n$. For this reason, the forecasted value given by the conditional expectation is often known as the *best linear predictor* of X_{n+h} in the sense of minimizing the mean square error of the predicted value.

Figure 8.4 illustrates the case of a (tiny) time series of just $n = 4$ observations: x_1, x_2, x_3, x_4. (Recall that $n = 60$ or $n = 70$ is the minimum value of n in practice. This example with a tiny value of n is for illustrative purposes only.) The observed values of the time series are indicated by points which are connected by lines. Each of the three forecasted values, $\hat{X}_5, \hat{X}_6, \hat{X}_7$, is indicated by a ○.

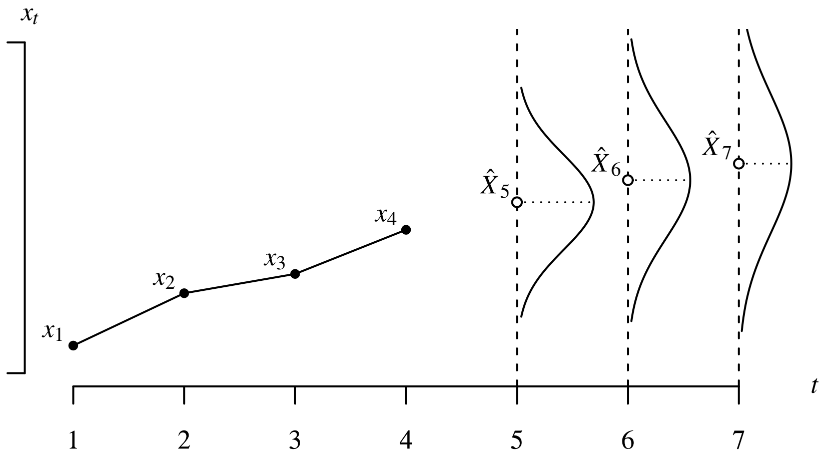

Figure 8.4: Forecasting three future values from $n = 4$ observations.

The three forecasts, associated with $h = 1$, $h = 2$, and $h = 3$, are made at time $t = n = 4$. In addition, there are three probability density functions, each rotated clockwise 90°, which indicate the probability distributions of the random future observations X_5, X_6, X_7. There are three key observations associated with this figure.

- The time series values x_1, x_2, x_3, x_4 increase over time, and the associated forecasted values $\hat{X}_5, \hat{X}_6, \hat{X}_7$ continue this trend.
- The population variance of the probability distributions of X_5, X_6, X_7 increases as the forecasting time horizon increases. This is consistent with weather prediction, for example, in that the weather prediction three days from now is less precise than the weather prediction tomorrow.
- The random sampling variability that is apparent in the four observed values x_1, x_2, x_3, x_4 is not apparent in the forecasted values $\hat{X}_5, \hat{X}_6, \hat{X}_7$. Observed time series values typically exhibit random sampling variability; forecasted values tend to be smooth.

Our goal in this subsection is to discuss forecasting generally and to introduce techniques for determining point estimates and interval estimates for future values in a time series. The example that follows assumes that a valid ARMA model has been specified and the parameters in a time series model are known, rather than estimated from a realization of the time series. For a long realization (large n) or significant amounts of previous history associated with a particular time series, this assumption might not pose any problem. In order to derive a prediction interval for X_{n+h}, the white noise terms are assumed to be Gaussian white noise for mathematical tractability. The reason for this assumption will be apparent in the following example.

Example 8.12 Consider the shifted stationary AR(1) time series model

$$X_t - \mu = \phi\,(X_{t-1} - \mu) + Z_t,$$

where $\{Z_t\}$ is Gaussian white noise and $-1 < \phi < 1$, μ, and $\sigma_Z^2 > 0$ are fixed, known parameters. Let $x_1, x_2, \ldots, x_n$ be one realization of the time series.

(a) Find a point estimate and an exact two-sided $100(1-\alpha)\%$ prediction interval for X_{n+1}.

(b) Find a point estimate and an exact two-sided $100(1-\alpha)\%$ prediction interval for X_{n+2}.

Notice that ϕ is a constant here and should not be confused with the polynomial $\phi(B)$. This is an unusual case because the three parameters ϕ, μ, and σ_Z^2 are known. In addition, it is assumed that the AR(1) model is a perfect stochastic model to govern the time series. Neither of these assumptions are typically satisfied perfectly in practice.

(a) Writing the AR(1) time series model with X_{n+1} on the left-hand side:

$$X_{n+1} - \mu = \phi\,(x_n - \mu) + Z_{n+1}$$

or

$$X_{n+1} = \mu + \phi\,(x_n - \mu) + Z_{n+1}.$$

Notice that X_{n+1} and Z_{n+1} are random future values which are set in uppercase, but x_n has already been observed, so it is set in lowercase. Taking the conditional expected value of both sides of this equation yields the one-step-ahead forecast

$$E\left[X_{n+1}\,|\,X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] = \mu + \phi\,(x_n - \mu)$$

because the expected value of a constant is a constant and the future Gaussian white noise term has conditional expected value 0. Taking the conditional population variance of both sides of the equation yields

$$V\left[X_{n+1}\,|\,X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] = \sigma_Z^2$$

because μ, ϕ, and x_n are all constants and the population variance is unaffected by a shift. So the point estimate of X_{n+1} is

$$\hat{X}_{n+1} = \mu + \phi\,(x_n - \mu)\,.$$

Since X_{n+1} is a constant, $\mu + \phi\,(x_n - \mu)$, plus a normal random variable, Z_{n+1}, it too is normally distributed with conditional mean $\hat{X}_{n+1}$ and conditional population variance σ_Z^2. So an exact two-sided $100(1-\alpha)\%$ prediction interval for X_{n+1} is

$$\hat{X}_{n+1} - z_{\alpha/2}\sigma_Z < X_{n+1} < \hat{X}_{n+1} + z_{\alpha/2}\sigma_Z,$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

(b) Writing the AR(1) time series model with X_{n+2} on the left-hand side:

$$X_{n+2} - \mu = \phi\,(X_{n+1} - \mu) + Z_{n+2}$$

or

$$X_{n+2} = \mu + \phi\,(X_{n+1} - \mu) + Z_{n+2}.$$

All of the X and Z variables are random future values, so they are set in uppercase. Taking the conditional expected value of both sides of this equation yields the two-step-ahead forecast

$$\begin{aligned}
&E\left[X_{n+2}\,|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] \\
&\qquad = \mu + \phi\big(E\left[X_{n+1}\,|\,X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] - \mu\big) \\
&\qquad = \mu + \phi\big(\phi(x_n - \mu)\big) \\
&\qquad = \mu + \phi^2(x_n - \mu)
\end{aligned}$$

because the conditional expected value of Z_{n+2} is zero. Taking the conditional population variance of both sides of the equation yields

$$\begin{aligned}
&V\left[X_{n+2}\,|\,X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] \\
&\quad = \phi^2 V\left[X_{n+1}\,|\,X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] + V\left[Z_{n+2}\,|\,X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right] \\
&\quad = \left(\phi^2 + 1\right)\sigma_Z^2.
\end{aligned}$$

So the point estimate of X_{n+2} is

$$\hat{X}_{n+2} = \mu + \phi^2\,(x_n - \mu)\,.$$

Since X_{n+2} is written as a constant, μ, plus the linear combination of two normally distributed random variables, $\phi\,(X_{n+1} - \mu)$ and Z_{n+2}, which is itself normally distributed, an exact two-sided $100(1-\alpha)\%$ prediction interval for X_{n+2} is

$$\hat{X}_{n+2} - z_{\alpha/2}\sqrt{\phi^2 + 1}\,\sigma_Z < X_{n+2} < \hat{X}_{n+2} + z_{\alpha/2}\sqrt{\phi^2 + 1}\,\sigma_Z.$$

Notice that for $\phi \neq 0$, the prediction interval for X_{n+2} is wider than the prediction interval for X_{n+1} for the same time series values and the same α value. This is consistent with intuition because we are less certain as we forecast further out into the future. This is the typical case in practice. On the other hand, the two prediction intervals have identical width when $\phi = 0$ because the AR(1) time series model reduces to Gaussian white noise in this case, and each future observation will have the same precision because of the mutual independence of the X_t values in this case.

This case was ideal in the sense that all three of the parameters, ϕ, μ, and σ_Z^2, are fixed and known. When these parameters are replaced by their point estimates, $\hat{\phi}$, $\hat{\mu}$, and $\hat{\sigma}_Z^2$, the prediction intervals become approximate rather than exact.

The previous example has illustrated the process for determining forecasted values and associated prediction intervals for an AR(1) time series model with known parameters. Consider generalizing this process for the h-step-ahead forecast. In order to obtain a point estimate for the forecast, take the conditional expected value of both sides of the model with X_{n+h} isolated on the left-hand side, which effectively results in: (a) present and past values of X_t are replaced by their observed values; (b) future values of Z_t are replaced by their conditional expected values, which are zero; and (c) future values of X_t are replaced by their conditional expected values. After simplification, this results in the forecast value $\hat{X}_{n+h}$.

As is typically the case in statistics, a point estimate is usually accompanied by an interval estimate which gives an indication of the precision of the point estimate. In a time series setting, a *prediction interval* for X_{n+h} has the generic form

$$\hat{X}_{n+h} \pm z_{\alpha/2} \sqrt{V\left[X_{n+h} \,|\, X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right]}.$$

This formula assumes that the random future value at time $n + h$, denoted by X_{n+h}, is normally distributed. This is usually achieved by assuming that the white noise terms consist of Gaussian white noise. Unlike confidence intervals, prediction intervals typically do not have widths that shrink to zero as the sample size n increases.

This ends the important topic of forecasting. Many more examples of forecasting will appear in subsequent sections in this chapter when special cases of ARMA(p, q) models are introduced. We now turn to another important statistical topic, which is model assessment.

8.2.3 Model Assessment

It is often the case that we have little or no information concerning the underlying physical mechanism governing a time series, so we must resort to an entirely data-driven approach to developing a time series model that adequately approximates the underlying probability mechanism. The usual approach to building a times series model consists of iterating through the following steps until a suitable model is formulated. The model building process is—by design—both iterative and interactive, making R an ideal platform for carrying out the process.

1. Identify a tentative time series model.

2. Estimate the unknown parameters of the tentative time series model.

3. Assess the adequacy of the fitted time series model.

The third step is considered in this section. As an instance of this approach, let's say we decide (based on inspecting plots of the time series, the sample autocorrelation function, and the sample partial autocorrelation function) that a shifted AR(2) time series model is a strong candidate for modeling a particular time series. After the parameters μ, ϕ_1, ϕ_2, and σ_Z^2 are estimated, we hope that the fitted model adequately models the underlying probability mechanism for the time series. If this is the case, then the *signal* associated with the time series has been captured, and all that should remain is *noise*. So how do we test whether or not the fitted model provides an adequate representation of the time series? One common approach taken in time series modeling is to assess whether the random shocks $\{Z_t\}$ are mutually independent and identically distributed random variables with population mean zero and common population variance σ_Z^2. But these Z_t values are not observed by the modeler, so instead we inspect the *residuals*, which are estimates of the Z_t values. In time series analysis, this important step is known as *diagnostic checking* or *residual analysis*. (This step is analogous to the similar step in regression analysis.) This process is the rough equivalent of *goodness-of-fit testing* from classical statistical theory. A residual value is defined as

$$[\text{residual}] = [\text{observed value}] - [\text{predicted value}].$$

The predicted value is the one-step-ahead forecast from the time $t-1$. Using the notation from the forecasting section, the residual at time t can be written as

$$\hat{Z}_t = X_t - \hat{X}_t.$$

This is one instance in which a more precise notation for a forecasted value would be helpful; this is more clearly written as

$$\hat{Z}_t = X_t - \hat{X}_{t-1}(1).$$

The hat is added to Z_t in order to indicate that the parameters in the fitted model have been estimated from the observed time series. Only in a simulated time series with known parameters do we observe Z_t. The residuals are ordered in time, so they can be viewed as a time series in their own right. If the hypothesized and fitted model are adequate, then the time series plot of the residuals will approximate a time series of white noise. The question here is how closely the residuals resemble white noise terms.

The behavior of the residuals is an indicator of whether the time series has been adequately modeled. If the model has been specified correctly and the parameter estimates are near their associated population values, then the residuals should appear to be white noise values, with common population mean zero and common population standard deviation. If this is not the case, then the search for an adequate time series model should continue.

A plot of the residuals over time is a crucial initial step in assessing whether they resemble white noise terms. Carefully examine the plot for any signs of trend, seasonality, or serial correlation. An example of a plot of Gaussian white noise was given in Figure 7.3. This step is just as important in residual analysis as was the inspection of the plot of the original time series. In addition, a plot of the sample autocorrelation function and the sample partial autocorrelation function of the residuals can be helpful in assessing whether the residuals closely approximate white noise. But rather than just a subjective visual inspection, we also want to confirm our intuition with a formal statistical test. The next four paragraphs briefly survey four statistical tests to assess the following null and alternative hypotheses:

$$H_0 : \text{the residuals are mutually independent and identically distributed random variables}$$

versus

$$H_1 : \text{the residuals are not mutually independent and identically distributed random variables.}$$

If there is no apparent visual trend, seasonality, or serial correlation in the residuals, then any one of the four hypothesis tests that follow can be conducted to confirm that the residuals do not exhibit any of these characteristics.

Count the number of significant spikes in the sample autocorrelation function. This test begins with a plot of the sample autocorrelation function of the residuals. If the residuals are well approximated by white noise terms, then the time series model can be judged to be adequate. The sample autocorrelation function values for white noise terms are approximately mutually independent and identically distributed $N(0,\, 1/n)$ random variables. So if the residuals closely approximate white noise, then any sample autocorrelation function value will fall between $-1.96/\sqrt{n}$ and $1.96/\sqrt{n}$ with approximate probability 0.95. We would like to conduct a hypothesis test in which the null hypothesis is that the sample autocorrelation function values of the residuals are independent $N(0,\, 1/n)$ random variables. A large number of sample autocorrelation values falling outside of the limits (which serves as the test statistic here) will result in rejecting the null hypothesis. So if each sample autocorrelation function value can be thought of as a toss of a biased coin in the case of the residuals being approximately white noise, then for, say, the first $m = 40$ such values, we expect $40 \cdot 0.05 = 2$ to fall outside of the limits $\pm 1.96/\sqrt{n}$. (Of course, the lag 0 sample autocorrelation $r_0 = 1$ is not included in the count.) In order to achieve an approximate level of significance $\alpha = 0.05$, if four or fewer of the 40 sample autocorrelation function values associated with the residuals fall outside of $\pm 1.96/\sqrt{n}$, we fail to reject H_0. The time series model is deemed to be adequate. But if five or more of the 40 sample autocorrelation function values associated with the residuals fall outside of $\pm 1.96/\sqrt{n}$, this is evidence *against* the hypothesized model and we reject H_0. The time series model is deemed to be inadequate. The p-value associated with four or fewer of the 40 sample autocorrelation function values associated with the residuals falling outside of the limits $\pm 1.96/\sqrt{40}$ can be calculated with the R statement

```
1 - pbinom(4, 40, 0.05)
```

This statement returns

```
[1] 0.04802826
```

So the exact level of significance for this test is $\alpha = 0.048$, which is quite close to the desired level of significance of 0.05. Rather than using trial and error with the `pbinom` function to determine the number of lags to use as the critical value, the `qbinom` function can be used to determine the cutoff.

```
qbinom(0.95, 40, 0.05)
```

This statement returns

```
[1] 4
```

A similar analysis can be applied to lag counts other than the $m = 40$ sample autocorrelation function values illustrated above. This analysis assumes that the sample autocorrelation function values of the residuals are independent and identically distributed normal random variables. One weakness of this approach is that it simply counts the number of sample autocorrelation function values falling outside the 95% confidence interval limits and ignores (a) how far outside of the limits the values fall or (b) how close to the limits they fall when they lie within the limits. This weakness prompts us to seek a statistical test that captures all of the sample autocorrelation function values associated with the residuals and includes their magnitudes.

Box–Pierce test. Let r_k be the lag k sample autocorrelation function value associated with the *residuals* of the fitted time series. As before, we only consider the first m such sample autocorrelation function values $r_1, r_2, \ldots, r_m$. It is approximately true that for mutually independent and identically distributed residuals,

$$r_k \sim N(0, 1/n).$$

By the transformation technique, this implies that

$$\sqrt{n}\, r_k \sim N(0, 1).$$

Squaring this random variable gives

$$n r_k^2 \sim \chi^2(1).$$

Assuming that the sample autocorrelation function values are uncorrelated, the sum of the first m of these random variables is

$$n \sum_{k=1}^{m} r_k^2 \sim \chi^2(m).$$

In the case in which r unknown model parameters have been estimated, the degrees of freedom are reduced by r:

$$n \sum_{k=1}^{m} r_k^2 \sim \chi^2(m-r).$$

This is the test statistic for the Box–Pierce test for serial correlation. Large values of this test statistic lead to rejecting H_0 and indicate a poor fit. The null hypothesis is rejected at level of significance α when this test statistic is greater than $\chi^2_{m-r,\alpha}$, where the first subscript is the number of degrees of freedom and the second subscript is the right-hand tail probability associated with this quantile of the chi-square distribution. There have been several approximations that occurred in formulating this statistical test. First, the r_k values are only approximately normally distributed. Second, the r_k values have variances which are less than $1/n$ for small lag values k. To compound this approximation, these smaller initial variances are dependent on the model under consideration. Third, the r_k values exhibit some serial correlation even when the residuals are mutually independent and identically distributed. These three weaknesses prompted a modification of the Box–Pierce test which provides a test statistic whose distribution more closely approximates the $\chi^2(m-r)$ distribution.

Ljung–Box test. The Box–Pierce test statistic was modified by Ljung and Box as

$$n(n+2) \sum_{k=1}^{m} \frac{r_k^2}{n-k},$$

which is approximately $\chi^2(m-r)$, where r is the number of parameters estimated in the model. Comparing the Box–Pierce and Ljung–Box test statistics, since

$$\frac{n+2}{n-k} > 1$$

for $k = 1, 2, \ldots, m$, the Ljung–Box test statistic always exceeds the Box–Pierce test statistic. The Box–Pierce test is more likely to accept a time series model with a poor fit than the Ljung–Box test for the same set of residuals. The Ljung–Box test should be used over the Box–Pierce because the probability distribution of its test statistic is closer to a $\chi^2(m-r)$ random variable under H_0.

Turning point test. As opposed to focusing on the sample autocorrelation function associated with the residuals, the turning point test considers the number of turning points in the time series of residuals. A turning point in a time series is defined to be a value associated with a local minimum or a local maximum. A local minimum occurs when $\hat{Z}_{t-1} > \hat{Z}_t$ and $\hat{Z}_t < \hat{Z}_{t+1}$. A local maximum occurs when $\hat{Z}_{t-1} < \hat{Z}_t$ and $\hat{Z}_t > \hat{Z}_{t+1}$. The random number of turning points in a time series of length n comprised of strictly continuous observations is denoted by T. The strictly continuous assumption is in place to avoid ties in adjacent values. A turning point cannot occur at the first or last value of the time series. Keep in mind that there might be fewer residuals than original observations. The n that is used here is the number of residuals. As given in an exercise at the end of this chapter, if the residuals are mutually independent and identically distributed continuous random variables, then

$$E[T] = \frac{2(n-2)}{3} \qquad \text{and} \qquad V[T] = \frac{16n-29}{90}.$$

Furthermore, even though T is a discrete random variable, it is well approximated by the normal distribution with population mean $E[T]$ and population variance $V[T]$ for a time series of mutually independent and identically distributed observations and large n. Thus, an appropriate test statistic for testing H_0 is

$$\frac{T-2(n-2)/3}{\sqrt{(16n-29)/90}},$$

which is approximately standard normal for large values of n. The null hypothesis is rejected in favor of the alternative hypothesis whenever the test statistic is less than $-z_{\alpha/2}$ (which indicates fewer turning points than expected, which is an indicator of *positive* serial correlation among the residuals) or the test statistic is greater than $z_{\alpha/2}$ (which indicates more turning points than expected, which is an indicator of *negative* serial correlation among the residuals).

This completes the brief introduction to four statistical tests concerning the mutual independence of the residuals. There are several other such tests, some of which are introduced in the exercises at the end of the chapter, but these four are representative of how such tests work. Three questions are given below concerning issues associated with the analysis of the residuals.

1. What if two time series models are deemed adequate by these statistical tests?

   Instances frequently arise in which two or more candidate time series models fail to be rejected by the statistical tests on residuals that were just surveyed. In these cases, the modeler has four guiding principles. First, there might be physical considerations that might favor one model over another. An engineer, for example, might provide some engineering design insight concerning why one time series model would be favored over another. Second, the model with the best value of one of the *model-selection statistics* outlined in the next section, might be the appropriate choice. Third, if the modeler is torn between two time series models, selecting the model with the fewer parameters follows the parsimony principle. We would like a time series model that adequately captures the probabilistic aspects of the time series with the minimum number of parameters. Fourth, the purpose of the model, for example, description, explanation, prediction, or simulation, might drive the final choice of the model.

2. If a time series model is deemed inadequate, can the analysis of the residuals guide the modeler toward a more suitable model?

   In some cases, the analysis of the residuals can indeed guide the modeler toward a more suitable time series model. Here is one instance. Let's say that a shifted AR(1) model is being considered as a potential time series model:

   $$X_t - \mu = \phi\,(X_{t-1} - \mu) + Z_t.$$

   Isolating the white noise term, this model can be written as

   $$X_t - \mu - \phi\,(X_{t-1} - \mu) = Z_t.$$

The parameters μ, ϕ, and σ_Z^2 are estimated from the observed time series, and the associated residuals are calculated and plotted. Rather than appearing as white noise, let's say that the residuals appear to look like observations from an MA(1) time series model

$$Z_t = W_t + \theta W_{t-1},$$

where $\{W_t\}$ is a time series of white noise. Combining the two previous equations, this would lead us in the direction of considering the model

$$X_t - \mu - \phi\,(X_{t-1} - \mu) = W_t + \theta W_{t-1},$$

which can be recognized as a shifted ARMA(1, 1) time series model. Thus, the ARMA(1, 1) composite model has been constructed from the two simpler models. We would then revisit parameter estimation procedures for the parameters μ, ϕ, θ, and σ_Z^2, and perform model adequacy tests on the associated residual values on the fitted ARMA(1, 1) model.

3. If a time series model is deemed adequate, should the noise terms be modeled as white noise or Gaussian white noise?

   The four statistical tests for autocorrelation do not assess the *normality* of the residuals. Drawing a histogram of the residuals is an important first step in terms of determining whether the residuals are normally distributed. If the histogram appears to be bell-shaped, then the Gaussian white noise aspect of the model is justified. Some time series analysts prefer to view a histogram of the standardized residuals, and the vast majority of these values should lie between -3 and 3. A QQ (quantile–quantile) plot is also useful for visually assessing normality, which can be graphed with the R function `qqnorm`. A QQ plot which is linear is an indication of normality. The behavior at the extremes of a QQ plot is typically more variable than at the center, so some analysts prefer to focus on the behavior between, say, the first and third quartiles. Assessing the normality of a histogram or the linearity of a QQ plot is subjective. Objective statistical tests for the normality of the residuals include the Shapiro–Wilk, Anderson–Darling, Cramer–von Mises, and Kolmogorov–Smirnov tests.

Analyzing the residuals is not the only way to assess the adequacy of a time series model. Another technique is known as *overfitting*. ARMA models with a single additional term are fitted to the original time series. This approach is analogous to *forward selection* in the stepwise approach to multiple regression. We will refer to the time series model under consideration as the *tentative* model and the overfitted models as *enhanced* models. For example, if an MA(1) model is being contemplated as a tentative time series model, then

- adding an additional moving average term yields the enhanced MA(2) model, and
- adding an autoregressive term yields the enhanced ARMA(1, 1) model.

The parameters for these two enhanced models should be fit to the original time series in the usual fashion. If both of the following two criteria are met, then the tentative time series model should be accepted as the final model.

- The parameter estimates in the enhanced models are close to the parameter estimates in the tentative model.
- The additional parameter in the enhanced models does not differ significantly from zero.

So in the example given above, the parameters in the tentative MA(1) model, θ_1 and σ_Z^2, should be estimated from the original time series. Then the parameters in the enhanced MA(2) model, θ_1, θ_2, and σ_Z^2, should be estimated from the original time series. If a confidence interval for θ_2 contains zero (or you fail to reject the null hypothesis $H_0 : \theta_2 = 0$ versus the alternative hypothesis $H_1 : \theta_2 \neq 0$), and the other parameter estimates do not vary significantly between the two models, then the modeler concludes that the extra parameter in the MA(2) model is not necessary. The same type of thinking applies to the enhanced ARMA(1, 1) model. So in addition to a careful examination of the residuals, it is also helpful to overfit the model in the autoregressive and moving average directions to assess whether the additional term significantly improves the fit.

The model assessment techniques described in this subsection will be applied to actual time series later in this chapter.

8.2.4 Model Selection

Model-selection statistics are helpful when there are two or more tentative fitted ARMA(p, q) models for a time series which have been deemed adequate by the model assessment techniques outlined in the previous subsection. One naive approach to model selection is to just add additional terms to an ARMA(p, q) model and check the resulting sum of the squared residuals. This approach violates the parsimony principle because it is typically the case that adding parameters to a model results in a lower sum of squared residuals. Just blindly adding terms to minimize the sum of squares is likely to produce time series models with superfluous terms that contain no real explanatory value, which can potentially cause problems in the application of the model.

We seek some statistical measure that strikes a balance between simplicity and capturing the essence of the probabilistic mechanism governing the time series model. Some statistical measure which reflects the benefit of an additional parameter, but extracts a penalty for adding parameters would be helpful to strike this balance.

In the case in which the analyst is presented with multiple plausible tentative fitted models, a model-selection statistic such as Akaike's Information Criterion might prove helpful in determining the best model. This statistic strikes a harmony between a simple model (which might not capture certain probabilistic aspects of the mechanism governing the time series) and a more complex model (which might contain unnecessary terms). This is the notion of a *parsimonious* model which uses as few parameters as possible to achieve adequate explanatory power. Akaike's Information Criterion (AIC), named after Japanese statistician Hirotugu Akaike (1927–2009), extracts a penalty for each additional parameter that is added to the model. The AIC is

$$\text{AIC} = -2\ln\big(L(\cdot)\big) + 2r,$$

where r is the number of unknown parameters that are estimated and L is the likelihood function evaluated at the maximum likelihood estimators for the r unknown parameters. Since $L(\cdot)$ is maximized at the maximum likelihood estimators, the first part of the AIC statistic, namely $-2\ln\big(L(\cdot)\big)$, is minimized at the maximum likelihood estimator values because of the negative sign. The $2r$ term can be thought of as a penalty term for adding additional parameters to the model. Each additional parameter added to the model will probably decrease the first term in the AIC involving the log likelihood function, but will also increase the penalty term because r has been increased. The model with the lowest value of AIC is deemed by this model-selection statistic to be the most appropriate parsimonious time series model.

There are two variants of the AIC that provide improved ability to correctly identify a time series model.

- The AIC estimates the expected value of the Kullback–Leibler divergence of the estimated model from the true model, and there is a slight bias in the AIC which is significant for small values of n. The *corrected* Akaike Information Criterion, usually denoted by AICC, replaces the $2r$ penalty term with $2rn/(n-r-1)$, resulting in

  $$\text{AICC} = -2\ln\big(L(\cdot)\big) + \frac{2rn}{n-r-1}.$$

  Since $n/(n-r-1) > 1$, the AICC always exceeds the AIC for the same time series, meaning that the penalty for adding parameters is increased. The AICC will be more stingy than the AIC when it comes to adding parameters. The AICC model-selection statistic compensates for the AIC's tendency to overfit models.

- Another variant to the AIC is the Bayesian Information Criterion (BIC) which replaces the penalty term $2r$ with $r\ln n$, resulting in

  $$\text{BIC} = -2\ln\big(L(\cdot)\big) + r\ln n.$$

  As shown in Figure 8.5 for a time series of length $n = 50$ and $r = 0, 1, 2, \ldots, 5$ unknown parameters, the BIC places an even higher penalty on additional terms in the time series model than the AIC and the AICC, which will result, on average, with time series models with fewer terms. Since the use of maximum likelihood estimation is required for calculating AIC, AICC, and BIC because all three are a function of the likelihood function L, the white noise terms are assumed to be normally distributed (that is, Gaussian white noise). A visual check of this assumption can be made by looking at a histogram of the residuals or a QQ plot of the residuals.

The time series analyst should consult with people who are familiar with the time series in order to glean whether there might be some aspects of the data set that might suggest one particular model or another. The analyst should also not necessarily assume that one of the models suggested in this

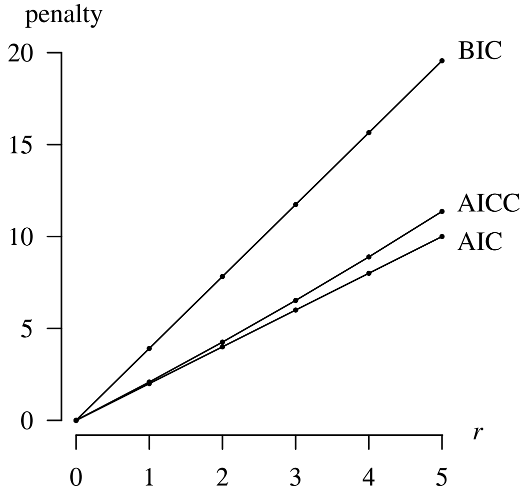

Figure 8.5: Penalty terms for model-selection statistics AIC, AICC, and BIC.

chapter might be appropriate for every setting. There are seldom uniquely correct values for p and q but rather these model-selection statistics are helpful in comparing two fitted tentative models.

In principle, the general linear model and its associated statistical methods are all that is necessary to fit and assess an ARMA(p, q) model. Since each specific ARMA(p, q) model has its own idiosyncrasies, the first few special cases of the autoregressive and moving average models will be examined in the next chapter.

8.3 Exercises

8.1 Show that the general linear model

$$X_t = Z_t + \psi_1 Z_{t-1} + \psi_2 Z_{t-2} + \cdots$$

can be written in the form

$$X_t = Z_t + \pi_1 X_{t-1} + \pi_2 X_{t-2} + \cdots .$$

8.2 For the ARMA time series model

$$X_t = 4X_{t-1} - 3X_{t-2} - 2X_{t-3} + Z_t - 5Z_{t-1} + 6Z_{t-2}.$$

(a) identify the time series model, and

(b) write the time series model in terms of the backshift operator B.

8.3 For the ARMA time series model

$$\phi(B)X_t = \theta(B)Z_t,$$

where $\phi(B) = 1$ and $\theta(B) = 1 - 0.6B + 0.1B^2$,

(a) identify the time series model, and

(b) write the time series model in purely algebraic form.

8.4 For the ARMA time series model

$$X_t = 2X_{t-1} - X_{t-2} + Z_t - Z_{t-2},$$

(a) identify the time series model, and

(b) write the time series model using the backshift operator.

8.5 Consider the special case of the general linear model

$$X_t = \frac{1}{2}X_{t-1} + Z_t - \frac{1}{3}Z_{t-1}.$$

(a) Write this model in its causal representation.

(b) Write this model in its invertible representation.

8.6 Show that $E[X_t] = \mu$ for the stationary shifted ARMA(p, q) model

$$X_t - \mu = \phi_1 (X_{t-1} - \mu) + \phi_2 (X_{t-2} - \mu) + \cdots + \phi_p (X_{t-p} - \mu) + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}.$$

8.7 Find $E[X_t]$ for the shifted ARMA(2, 1) model

$$X_t = 7 + 0.4X_{t-1} - 0.1X_{t-2} + Z_t + 0.3Z_{t-1}.$$

8.8 Let $X_1, X_2, \ldots, X_n$ be observations from an ARMA(0, 0) time series model with Gaussian white noise. The maximum likelihood estimator of the population variance of the Gaussian white noise derived in Example 8.10 is

$$\hat{\sigma}_Z^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

An asymptotically exact confidence interval for σ_Z^2 derived in Example 8.11 is

$$\hat{\sigma}_Z^2 - z_{\alpha/2}\sqrt{\frac{2\left(\sum_{i=1}^{n} X_i^2\right)^2}{n^3}} < \sigma_Z^2 < \hat{\sigma}_Z^2 + z_{\alpha/2}\sqrt{\frac{2\left(\sum_{i=1}^{n} X_i^2\right)^2}{n^3}}.$$

Calculate and plot the actual coverage of a 95% confidence interval for σ_Z^2 as a function of n for $n = 8, 9, \ldots, 256$. Use analytic methods rather than Monte Carlo simulation.

8.9 Let $X_1, X_2, \ldots, X_n$ be observations from an ARMA(0, 0) time series model with Gaussian white noise. Find the probability density function of the maximum likelihood estimator of the population variance of the Gaussian white noise

$$\hat{\sigma}_Z^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

8.10 Let $X_1, X_2, \ldots, X_n$ be observations from an ARMA(0, 0) time series model with Gaussian white noise. As shown in Example 8.10, the maximum likelihood estimator of the population variance of the Gaussian white noise is

$$\hat{\sigma}_Z^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

and a pivotal quantity for developing an exact two-sided $100(1-\alpha)\%$ confidence interval for σ_Z^2 is

$$\frac{n\hat{\sigma}_Z^2}{\sigma_Z^2} \sim \chi^2(n).$$

Find an exact two-sided $100(1-\alpha)\%$ confidence interval for σ_Z^2.

8.11 Let $X_1, X_2, \ldots, X_n$ be observations from an ARMA(0, 0) time series model with Gaussian white noise having finite positive population variance σ_Z^2. The maximum likelihood estimator of the population variance of the Gaussian white noise is

$$\hat{\sigma}_Z^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Conduct a Monte Carlo simulation experiment that provides convincing numerical evidence that

$$\frac{n\hat{\sigma}_Z^2}{\chi^2_{n,\alpha/2}} < \sigma_Z^2 < \frac{n\hat{\sigma}_Z^2}{\chi^2_{n,1-\alpha/2}}$$

is an *exact* $100(1-\alpha)\%$ confidence interval for σ_Z^2 for one particular set of n, α, and σ_Z^2 of your choice.

8.12 Let X_1 and X_2 be jointly distributed random variables. The population mean and variance of X_1 are μ_{X_1} and $\sigma^2_{X_1}$. The population mean and variance of X_2 are μ_{X_2} and $\sigma^2_{X_2}$. The population correlation between X_1 and X_2 is $\rho = \text{Corr}(X_1, X_2)$. The value of X_2 is to be predicted as a linear function of X_1 with $mX_1 + b$. Find the values of m and b which minimize the mean square error of the prediction. In other words, find m and b which minimize

$$E\left[(X_2 - mX_1 - b)^2\right].$$

8.13 Consider an ARMA(0, 0) model with $U(-1, 1)$ white noise terms. Find an exact two-sided 95% prediction interval for X_{n+h}.

8.14 Suppose an ARMA(2, 1) time series model is a strong candidate for modeling a particular time series. A long time series is available for analysis, so n is large. The ARMA(2, 1) model is fitted and residuals are calculated. If the sample autocorrelation function associated with the residuals is calculated for the first 100 lags, how many values need to fall outside of $\pm 1.96/\sqrt{n}$ in order to reject the null hypothesis H_0, which corresponds to a good fit at a significance level that is less than $\alpha = 0.05$?

8.15 Compare the expected p-values for the Box–Pierce and Ljung–Box tests for serial independence of a time series consisting of $n = 100$ mutually independent and identically distributed standard normal random variables. Consider only the first $k = 40$ lag values.

8.16 Let $\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_n$ be residual values associated with a fitted time series model. The Durbin–Watson test statistic defined by

$$D = \sum_{t=2}^{n} \left(\hat{Z}_t - \hat{Z}_{t-1}\right)^2 \Big/ \sum_{t=1}^{n} \hat{Z}_t^2$$

is useful for testing the serial independence of the residuals.

(a) Conduct a Monte Carlo simulation experiment to estimate the expected value of D when $\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_{1000}$ are $n = 1000$ mutually independent and identically distributed standard normal random variables.

(b) Give an explanation for the result that you obtained in part (a).

8.17 The *turning point test* for serial dependence counts the number of turning points (the number of local minima and maxima) T in a time series of length n comprised of strictly continuous observations. A turning point cannot occur at the first or last value of the time series.

(a) Show that $E[T] = 2(n-2)/3$ when the observations in the time series are mutually independent and identically distributed.

(b) Show that $V[T] = (16n - 29)/90$ when the observations in the time series are mutually independent and identically distributed.

(c) Perform a Monte Carlo simulation that supports the values of $E[T]$ and $V[T]$ for a time series of length $n = 101$.

(d) Argue why T is approximately normally distributed with population mean $E[T]$ and population variance $V[T]$ for a time series of mutually independent and identically distributed observations and large n. Suggest an appropriate test statistic for testing the null hypothesis that there is no serial correlation in the time series.

8.18 Let X_1, X_2, X_3, X_4 be a time series of mutually independent and identically distributed continuous random variables. Let T be the number of turning points. Find the probability mass function of T.

8.19 The nonparametric *difference–sign test* for serial dependence counts the number of values in a time series of strictly continuous observations $X_1, X_2, \ldots, X_n$ in which $X_i > X_{i-1}$, for $i = 2, 3, \ldots, n$. Denote this count by T.

(a) Show that $E[T] = (n-1)/2$ when the observations in the time series are mutually independent and identically distributed.

(b) Show that $V[T] = (n+1)/12$ when the observations in the time series are mutually independent and identically distributed.

(c) Perform a Monte Carlo simulation that supports the values of $E[T]$ and $V[T]$ for a time series of length $n = 101$.

(d) Argue why T is approximately normally distributed with population mean $E[T]$ and population variance $V[T]$ for a time series of mutually independent and identically distributed observations and large n. Suggest an appropriate test statistic for testing the null hypothesis that there is no serial correlation in the time series.

8.20 Suppose an AR(1) model is being considered as a tentative time series model based on a realization of the time series. A single autoregressive parameter and a single moving average parameter is added to the tentative model, resulting in an ARMA(2, 1) enhanced model. Describe any problems that might arise by comparing the AR(1) time series model and the ARMA(2, 1) time series model.