For a time series with a seasonal pattern, following are typical values of s:

- 52 for weekly data
- 12 for monthly data
- 7 for daily data

The next section presents a seasonal ARIMA example and describes several techniques and approaches to identify the appropriate model and forecast the future.

8.2.5 Building and Evaluating an ARIMA Model

For a large country, the monthly gasoline production measured in millions of barrels has been obtained for the past 240 months (20 years). A market research firm requires some short-term gasoline production forecasts to assess the petroleum industry's ability to deliver future gasoline supplies and the effect on gasoline prices.

```
library(forecast)

# read in gasoline production time series
# monthly gas production expressed in millions of barrels
gas_prod_input <- as.data.frame( read.csv("c:/data/gas_prod.csv") )

# create a time series object
gas_prod <- ts(gas_prod_input[,2])

#examine the time series
plot(gas_prod, xlab = "Time (months)",
     ylab = "Gasoline production (millions of barrels)")
```

Using R, the dataset is plotted in Figure 8-11.

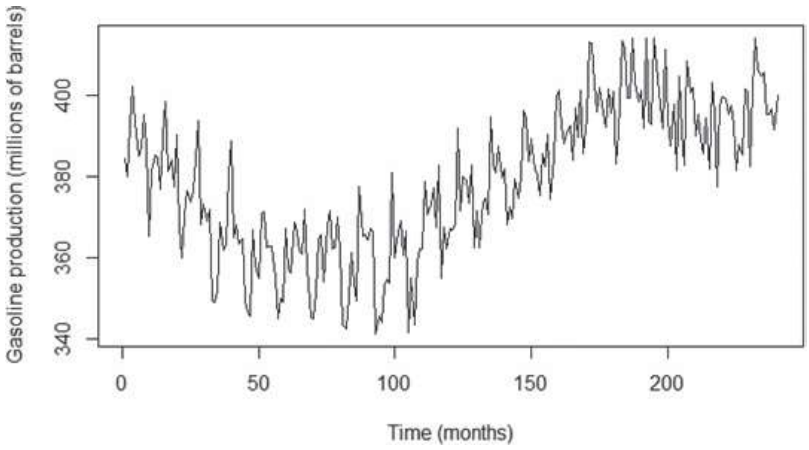

FIGURE 8-11 *Monthly gasoline production*

In R, the `ts()` function creates a time series object from a vector or a matrix. The use of time series objects in R simplifies the analysis by providing several methods that are tailored specifically for handling equally time spaced data series. For example, the `plot()` function does not require an explicitly specified variable for the x-axis.

To apply an ARMA model, the dataset needs to be a stationary time series. Using the `diff()` function, the gasoline production time series is differenced once and plotted in Figure 8-12.

```
plot(diff(gas_prod))
abline(a=0, b=0)
```

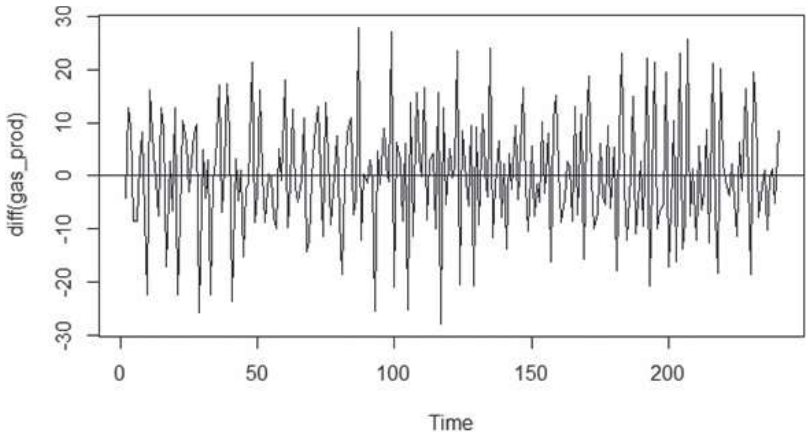

FIGURE 8-12 *Differenced gasoline production time series*

The differenced time series has a constant mean near zero with a fairly constant variance over time. Thus, a stationary time series has been obtained. Using the following R code, the ACF and PACF plots for the differenced series are provided in Figures 8-13 and 8-14, respectively.

```
# examine ACF and PACF of differenced series
acf(diff(gas_prod), xaxp = c(0, 48, 4), lag.max=48, main="")
pacf(diff(gas_prod), xaxp = c(0, 48, 4), lag.max=48, main="")
```

The dashed lines provide upper and lower bounds at a 95% significance level. Any value of the ACF or PACF outside of these bounds indicates that the value is significantly different from zero.

Figure 8-13 shows several significant ACF values. The slowly decaying ACF values at lags 12, 24, 36, and 48 are of particular interest. A similar behavior in the ACF was seen in Figure 8-3, but for lags 1, 2, 3,… Figure 8-13 indicates a seasonal autoregressive pattern every 12 months. Examining the PACF plot in Figure 8-14, the PACF value at lag 12 is quite large, but the PACF values are close to zero at lags 24, 36, and 48. Thus, a seasonal AR(1) model with period = 12 will be considered. It is often useful to address the seasonal portion of the overall ARMA model before addressing the nonseasonal portion of the model.

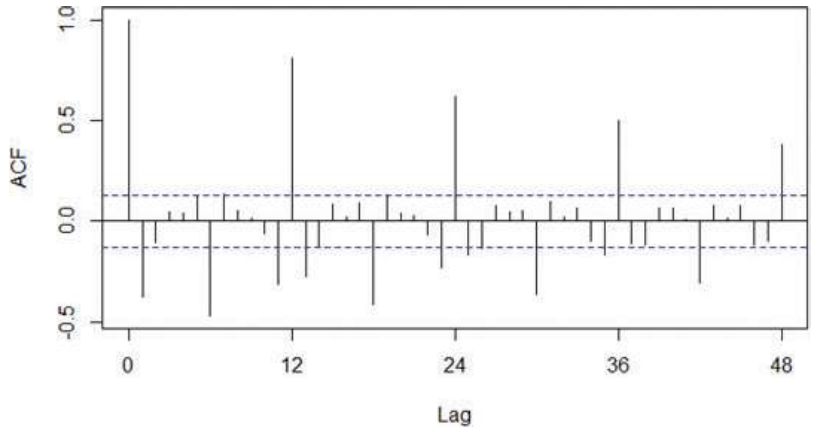

FIGURE 8-13 *ACF of the differenced gasoline time series*

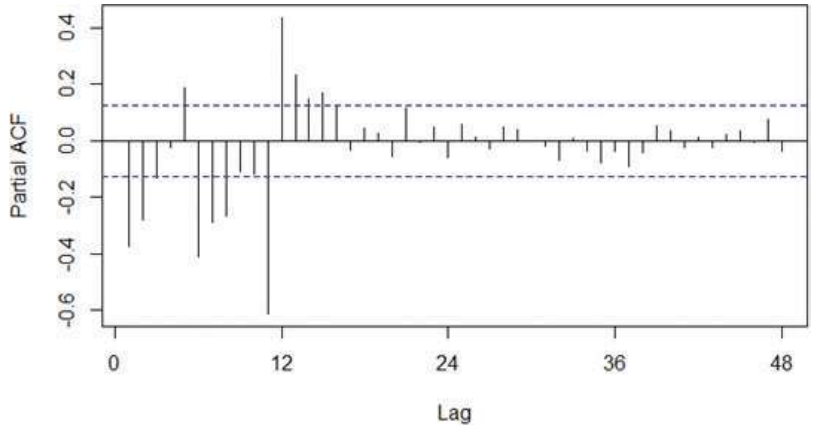

FIGURE 8-14 *PACF of the differenced gasoline time series*

The `arima()` function in R is used to fit a $(0,1,0) \times (1,0,0)_{12}$ model. The analysis is applied to the original time series variable, *gas_prod*. The differencing, d = 1, is specified by the order = c(0,1,0) term.

```
arima_1 <- arima (gas_prod,
                  order=c(0,1,0),
                  seasonal = list(order=c(1,0,0),period=12))
arima_1
```

```
Series: gas_prod
ARIMA(0,1,0)(1,0,0)[12]

Coefficients:
        sar1
      0.8335
```

```
s.e.  0.0324

sigma^2 estimated as 37.29:  log likelihood=-778.69
AIC=1561.38   AICc=1561.43   BIC=1568.33
```

The value of the coefficient for the seasonal AR(1) model is estimated to be 0.8335 with a standard error of 0.0324. Because the estimate is several standard errors away from zero, this coefficient is considered significant. The output from this first pass ARIMA analysis is stored in the variable `arima_1`, which contains several useful quantities including the residuals. The next step is to examine the residuals from fitting the $(0,1,0) \times (1,0,0)_{12}$ ARIMA model. The ACF and PACF plots of the residuals are provided in Figures 8-15 and 8-16, respectively.

```
# examine ACF and PACF of the (0,1,0)x(1,0,0)12 residuals
acf(arima_1$residuals, xaxp = c(0, 48, 4), lag.max=48, main="")
pacf(arima_1$residuals, xaxp = c(0, 48, 4), lag.max=48, main="")
```

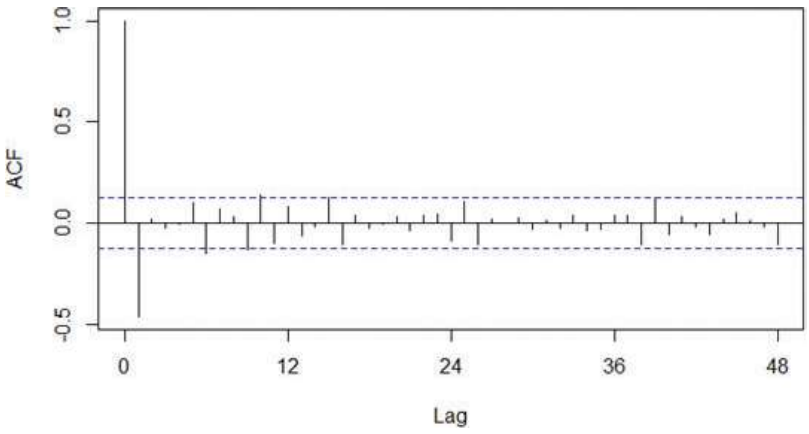

FIGURE 8-15 *ACF of residuals from seasonal AR(1) model*

The ACF plot of the residuals in Figure 8-15 indicates that the autoregressive behavior at lags 12, 24, 26, and 48 has been addressed by the seasonal AR(1) term. The only remaining ACF value of any significance occurs at lag 1. In Figure 8-16, there are several significant PACF values at lags 1, 2, 3, and 4.

Because the PACF plot in Figure 8-16 exhibits a slowly decaying PACF, and the ACF cuts off sharply at lag 1, an MA(1) model should be considered for the nonseasonal portion of the ARMA model on the differenced series. In other words, a $(0,1,1) \times (1,0,0)_{12}$ ARIMA model will be fitted to the original gasoline production time series.

```
arima_2 <- arima (gas_prod,
                  order=c(0,1,1),
                  seasonal = list(order=c(1,0,0),period=12))
arima_2

Series: gas_prod
ARIMA(0,1,1)(1,0,0)[12]

Coefficients:
          ma1     sar1
```

```
      -0.7065  0.8566
s.e.   0.0526  0.0298

sigma^2 estimated as 25.24:  log likelihood=-733.22
AIC=1472.43   AICc=1472.53   BIC=1482.86

acf(arima_2$residuals, xaxp = c(0, 48, 4), lag.max=48, main="")
pacf(arima_2$residuals, xaxp = c(0, 48,4), lag.max=48, main="")
```

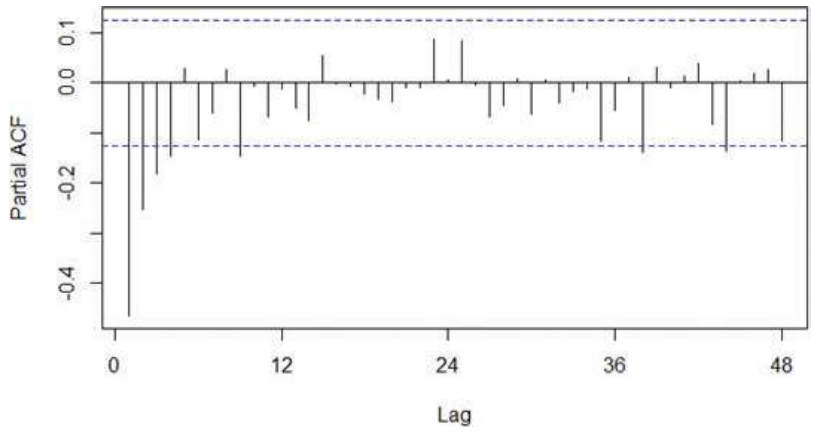

FIGURE 8-16 *PACF of residuals from seasonal AR(1) model*

Based on the standard errors associated with each coefficient estimate, the coefficients are significantly different from zero. In Figures 8-17 and 8-18, the respective ACF and PACF plots for the residuals from the second pass ARIMA model indicate that no further terms need to be considered in the ARIMA model.

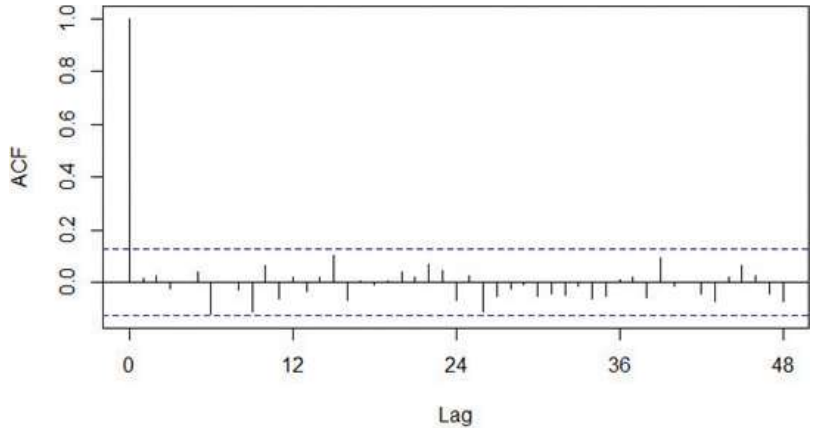

FIGURE 8-17 *ACF for the residuals from the* $(0,1,1) \times (1,0,0)_{12}$ *model*

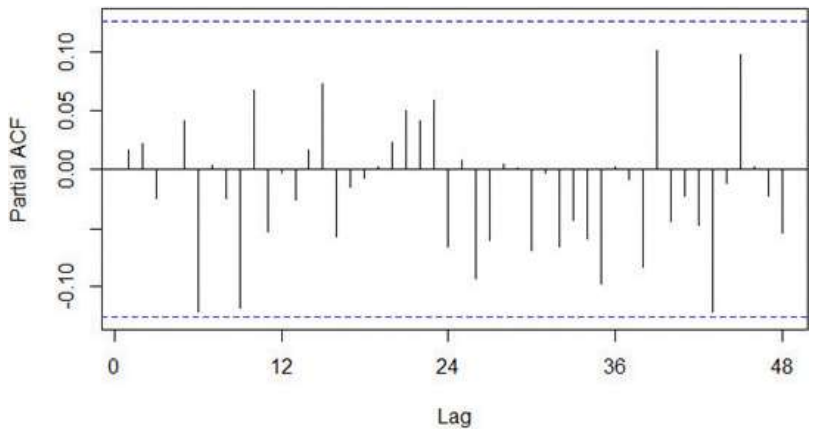

FIGURE 8-18 *PACF for the residuals from the* $(0,1,1)\times(1,0,0)_{12}$ *model*

It should be noted that the ACF and PACF plots each have several points that are close to the bounds at a 95% significance level. However, these points occur at relatively large lags. To avoid overfitting the model, these values are attributed to random chance. So no attempt is made to include these lags in the model. However, it is advisable to compare a reasonably fitting model to slight variations of that model.

Comparing Fitted Time Series Models

The `arima()` function in R uses Maximum Likelihood Estimation (MLE) to estimate the model coefficients. In the R output for an ARIMA model, the log-likelihood ($\log L$) value is provided. The values of the model coefficients are determined such that the value of the log likelihood function is maximized. Based on the $\log L$ value, the R output provides several measures that are useful for comparing the appropriateness of one fitted model against another fitted model. These measures follow:

- AIC (Akaike Information Criterion)
- AICc (Akaike Information Criterion, corrected)
- BIC (Bayesian Information Criterion)

Because these criteria impose a penalty based on the number of parameters included in the models, the preferred model is the fitted model with the smallest AIC, AICc, or BIC value. Table 8-1 provides the information criteria measures for the ARIMA models already fitted as well as a few additional fitted models. The highlighted row corresponds to the fitted ARIMA model obtained previously by examining the ACF and PACF plots.

TABLE 8-1 *Information Criteria to Measure Goodness of Fit*

ARIMA Model $(p,d,q) \times (P,Q,D)_s$	AIC	AICc	BIC
$(0,1,0) \times (1,0,0)_{12}$	1561.38	1561.43	1568.33
$(0,1,1) \times (1,0,0)_{12}$	1472.43	1472.53	1482.86
$(0,1,2) \times (1,0,0)_{12}$	1474.25	1474.42	1488.16
$(1,1,0) \times (1,0,0)_{12}$	1504.29	1504.39	1514.72
$(1,1,1) \times (1,0,0)_{12}$	1474.22	1474.39	1488.12

In this dataset, the $(0,1,1) \times (1,0,0)_{12}$ model does have the lowest AIC, AICc, and BIC values compared to the same criterion measures for the other ARIMA models.

Normality and Constant Variance

The last model validation step is to examine the normality assumption of the residuals in Equation 8-15. Figure 8-19 indicates residuals with a mean near zero and a constant variance over time. The histogram in Figure 8-20 and the Q-Q plot in Figure 8-21 support the assumption that the error terms are normally distributed. Q-Q plots were presented in Chapter 6, "Advanced Analytical Theory and Methods: Regression."

```
plot(arima_2$residuals, ylab = "Residuals")
abline(a=0, b=0)

hist(arima_2$residuals, xlab="Residuals", xlim=c(-20,20))

qqnorm(arima_2$residuals, main="")
qqline(arima_2$residuals)
```

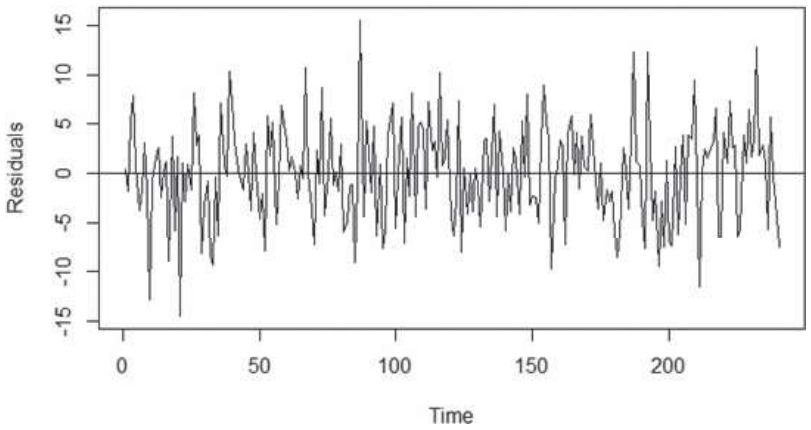

FIGURE 8-19 *Plot of residuals from the fitted $(0,1,1) \times (1,0,0)_{12}$ model*

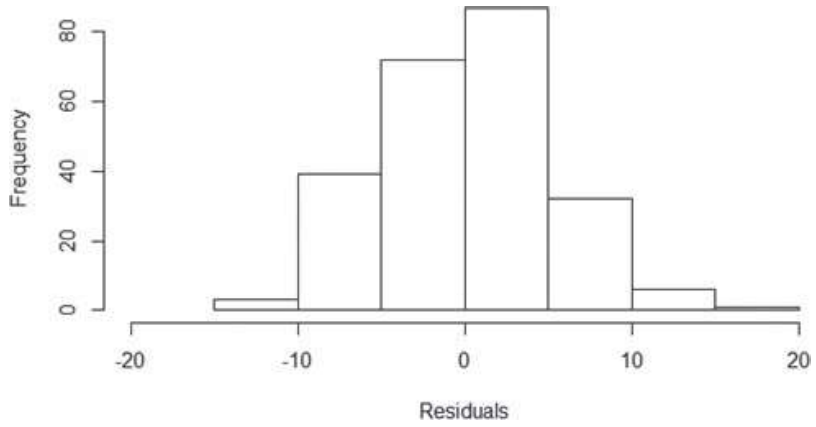

FIGURE 8-20 *Histogram of the residuals from the fitted* $(0,1,1)\times(1,0,0)_{12}$ *model*

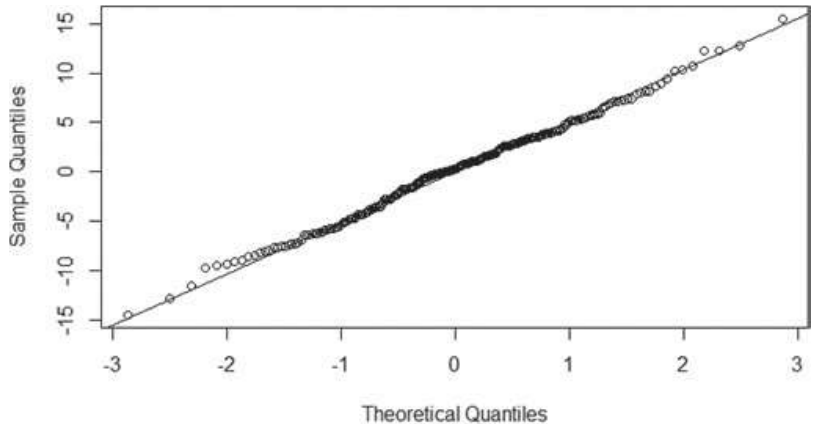

FIGURE 8-21 *Q-Q plot of the residuals from the fitted* $(0,1,1)\times(1,0,0)_{12}$ *model*

If the normality or the constant variance assumptions do not appear to be true, it may be necessary to transform the time series prior to fitting the ARIMA model. A common transformation is to apply a logarithm function.

Forecasting

The next step is to use the fitted $(0,1,1)\times(1,0,0)_{12}$ model to forecast the next 12 months of gasoline production. In R, the forecasts are easily obtained using the `predict()` function and the fitted model already stored in the variable *`arima_2`*. The predicted values along with the associated upper and lower bounds at a 95% confidence level are displayed in R and plotted in Figure 8-22.

```
#predict the next 12 months
arima_2.predict <- predict(arima_2,n.ahead=12)

matrix(c(arima_2.predict$pred-1.96*arima_2.predict$se,
```

```
          arima_2.predict$pred,
          arima_2.predict$pred+1.96*arima_2.predict$se), 12,3,
          dimnames=list( c(241:252) ,c("LB","Pred","UB"))  )

           LB     Pred       UB
241 394.9689 404.8167 414.6645
242 378.6142 388.8773 399.1404
243 394.9943 405.6566 416.3189
244 405.0188 416.0658 427.1128
245 397.9545 409.3733 420.7922
246 396.1202 407.8991 419.6780
247 396.6028 408.7311 420.8594
248 387.5241 399.9920 412.4598
249 387.1523 399.9507 412.7492
250 387.8486 400.9693 414.0900
251 383.1724 396.6076 410.0428
252 390.2075 403.9500 417.6926
plot(gas_prod, xlim=c(145,252),
     xlab = "Time (months)",
     ylab = "Gasoline production (millions of barrels)",
     ylim=c(360,440))
lines(arima_2.predict$pred)
lines(arima_2.predict$pred+1.96*arima_2.predict$se, col=4, lty=2)
lines(arima_2.predict$pred-1.96*arima_2.predict$se, col=4, lty=2)
```

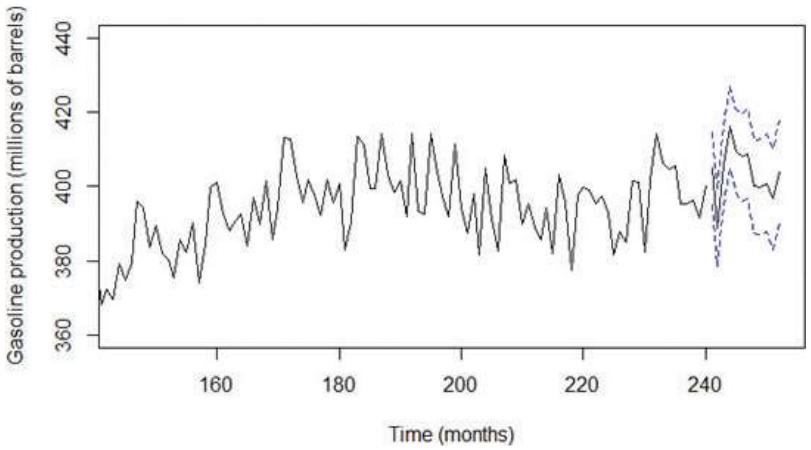

FIGURE 8-22 *Actual and forecasted gasoline production*

8.2.6 Reasons to Choose and Cautions

One advantage of ARIMA modeling is that the analysis can be based simply on historical time series data for the variable of interest. As observed in the chapter about regression (Chapter 6), various input variables