STAT 626: Outline of Lectures [redacted]

Seasonal ARIMA (SARIMA) Models (§5.3)

1. Review of ARMA Models [redacted] Introducing SARIMA Models,
   Steps for SARIMA Model Building.

2. **Plot the Data**

3. **Induce Stationarity by Seasonal Differencing or Other Means**

4. **Model Formulation: Use the ACF and PACF to Select** p, q, P, Q

5. **Model Estimation: Find the MLE of the Parameters**

6. **Model Diagnostic: Check the Residuals for Independence**
   $H_0 : \rho(1) = \ldots, \rho(H) = 0.$ **Residuals are uncorrelated (WN)**
   vs.
   H_a : **Residuals are correlated.**

7. **If Not Happy, Or** H_0 **is Rejected , Go to Step 2 and Repeat the PROCESS**

Example 5.11. A Seasonal AR Series

$$(1 - \Phi B^{12})x_t = w_t.$$

Seasonal MA(1):

$$x_t = (1 + \Theta B^{12})w_t.$$

Example: A Seasonal ARMA Series

$$(1 - \Phi B^{12})x_t = (1 + \Theta B^{12})w_t.$$

What are the connections with ARMA(1,1) models?

Is it causal? Invertible?

Its MA(∞) representation?

Its autocovariance function?

Its ACF?

Its predictors? Prediction error variance?

Example 5. 12: A Mixed Seasonal Model

$$x_t = \Phi x_{t-12} + w_t + \theta w_{t-1}.$$

What are the connections with ARMA(1,1) models?

Is it causal? Invertible?

Its MA(∞) representation?

Its autocovariance function?
Its ACF?
Its predictors? Prediction error variance?

Example 5.15: Carbon Dioxide and Global Warming

```
b = as.vector(reg$coef)
g = function(t){ b[1] + b[2]*(t-1955) + b[3]*(t-1955)^2 +
           b[4]*(t-1955)^3 + b[5]*(t-1955)^4 + b[6]*(t-1955)^5 +
           b[7]*(t-1955)^6 + b[8]*(t-1955)^7 + b[9]*(t-1955)^8
}
par(mar=c(2,2.5,.5,0)+.5, mgp=c(1.6,.6,0))
curve(g, 1900, 2024, ylab="Population", xlab="Year", main="U.S.
           Population by Official Census", panel.first=Grid(),
           cex.main=1, font.main=1, col=4)
abline(v=seq(1910,2020,by=20), lty=1, col=gray(.9))
points(time(uspop), uspop, pch=21, bg=rainbow(12), cex=1.25)
mtext(expression(""%*% 10^6), side=2, line=1.5, adj=.95)
axis(1, seq(1910,2020,by=20), labels=TRUE)
```

◊

The final step of model fitting is model choice or model selection. That is, we must decide which model we will retain for forecasting. The most popular techniques, AIC, AICc, and BIC, were described in Section 3.1 in the context of regression models.

Example 5.10. Model Choice for the U.S. GNP Series

To follow up on Example 5.7, recall that two models, an AR(1) and an MA(2), fit the GNP growth rate well. In addition, recall that it was shown that the two models are nearly the same and are not in contradiction. To choose the final model, we compare the AIC, the AICc, and the BIC for both models. These values are a byproduct of the `sarima` runs.

```
sarima(diff(log(gnp)), 1, 0, 0)  # AR(1)
  $AIC: -6.456   $AICc: -6.456   $BIC: -6.425
sarima(diff(log(gnp)), 0, 0, 2)  # MA(2)
  $AIC: -6.459   $AICc: -6.459   $BIC: -6.413
```

The AIC and AICc both prefer the MA(2) fit, whereas the BIC prefers the simpler AR(1) model. The methods often agree, but when they do not, the BIC will select a model of smaller order than the AIC or AICc because its penalty is much larger. Ignoring the philosophical considerations that cause nerds to verbally assault each other, it seems reasonable to retain the AR(1) because pure autoregressive models are easier to work with. ◊

5.3 Seasonal ARIMA Models

In this section, we introduce several modifications made to the ARIMA model to account for seasonal behavior. Often, the dependence on the past tends to occur most strongly at multiples of some underlying seasonal lag s. For example, with monthly economic data, there is a strong yearly component occurring at lags that are multiples of $s = 12$, because of the strong connections of all activity to the calendar year. Data taken quarterly will exhibit the yearly repetitive period at $s = 4$ quarters. Natural phenomena such as temperature also have strong components corresponding to seasons. Hence, the natural variability of many physical, biological, and economic

processes tends to match with seasonal fluctuations. Because of this, it is appropriate to introduce autoregressive and moving average polynomials that identify with the seasonal lags. The resulting *pure seasonal autoregressive moving average model*, say, ARMA$(P,Q)_s$, then takes the form

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t, \tag{5.14}$$

where the operators

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \tag{5.15}$$

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \tag{5.16}$$

are the **seasonal autoregressive operator** and the **seasonal moving average operator** of orders P and Q, respectively, with seasonal period s.

Example 5.11. A Seasonal AR Series

A first-order seasonal autoregressive series that might run over months, denoted SAR$(1)_{12}$, is written as

$$(1 - \Phi B^{12})x_t = w_t$$

or

$$x_t = \Phi x_{t-12} + w_t.$$

This model exhibits the series x_t in terms of past lags at the multiple of the yearly seasonal period $s = 12$ months. It is clear that estimation and forecasting for such a process involves only straightforward modifications of the unit lag case already treated. In particular, the causal condition requires $|\Phi| < 1$.

We simulated 3 years of data from the model with $\Phi = .9$, and exhibit the *theoretical* ACF and PACF of the model in Figure 5.9.

```
set.seed(666)
phi = c(rep(0,11),.9)
sAR = ts(arima.sim(list(order=c(12,0,0), ar=phi), n=37), freq=12) + 50
layout(matrix(c(1,2, 1,3), nc=2), heights=c(1.5,1))
par(mar=c(2.5,2.5,2,1), mgp=c(1.6,.6,0))
plot(sAR, xaxt="n", col=gray(.6), main="seasonal AR(1)", xlab="YEAR",
           type="c", ylim=c(45,54))
abline(v=1:4, lty=2, col=gray(.6))
axis(1,1:4);  box()
abline(h=seq(46,54,by=2), col=gray(.9))
Months = c("J","F","M","A","M","J","J","A","S","O","N","D")
points(sAR, pch=Months, cex=1.35, font=4, col=1:4)
ACF = ARMAacf(ar=phi, ma=0, 100)[-1]
PACF = ARMAacf(ar=phi, ma=0, 100, pacf=TRUE)
LAG = 1:100/12
plot(LAG, ACF, type="h", xlab="LAG", ylim=c(-.1,1), axes=FALSE)
segments(0,0,0,1)
```

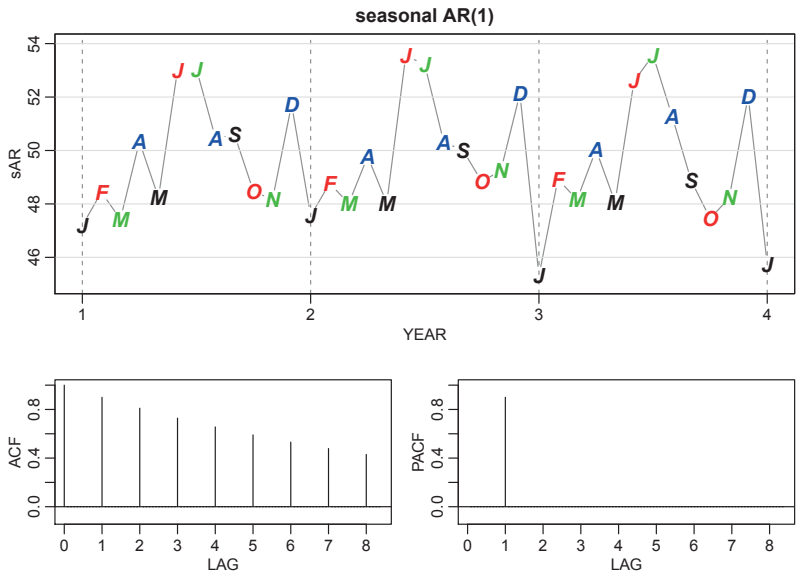

Figure 5.9 *Data generated from an SAR$(1)_{12}$ model, and the true ACF and PACF of the model $(x_t - 50) = .9(x_{t-12} - 50) + w_t$. LAG is in terms of seasons.*

```
axis(1, seq(0,8,by=1));  axis(2);  box();  abline(h=0)
plot(LAG, PACF, type="h", xlab="LAG", ylim=c(-.1,1), axes=FALSE)
axis(1, seq(0,8,by=1));  axis(2);  box();  abline(h=0)
```

◊

For the first-order seasonal ($s = 12$) MA model, $x_t = w_t + \Theta w_{t-12}$, it is easy to verify that

$$\begin{aligned} \gamma(0) &= (1+\Theta^2)\sigma^2 \\ \gamma(\pm 12) &= \Theta\sigma^2 \\ \gamma(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

Thus, the only nonzero correlation, aside from lag zero, is

$$\rho(\pm 12) = \Theta/(1+\Theta^2).$$

For the first-order seasonal ($s = 12$) AR model, using the techniques of the nonseasonal AR(1), we have

$$\begin{aligned} \gamma(0) &= \sigma^2/(1-\Phi^2) \\ \gamma(\pm 12k) &= \sigma^2\Phi^k/(1-\Phi^2) \quad k = 1, 2, \ldots \\ \gamma(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

In this case, the only non-zero correlations are

$$\rho(\pm 12k) = \Phi^k, \quad k = 0, 1, 2, \ldots .$$

Table 5.1 *Behavior of the ACF and PACF for Pure SARMA Models*

	$\text{AR}(P)_s$	$\text{MA}(Q)_s$	$\text{ARMA}(P,Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots$, are zero.

These results can be verified using the general result that

$$\gamma(h) = \Phi\gamma(h-12) \quad \text{for } h \geq 1.$$

For example, when $h = 1$, $\gamma(1) = \Phi\gamma(11)$, but when $h = 11$, we have $\gamma(11) = \Phi\gamma(1)$, which implies that $\gamma(1) = \gamma(11) = 0$. In addition to these results, the PACF have the analogous extensions from nonseasonal to seasonal models. These results are demonstrated in Figure 5.9.

As an initial diagnostic criterion, we can use the properties for the pure seasonal autoregressive and moving average series listed in Table 5.1. These properties may be considered as generalizations of the properties for nonseasonal models that were presented in Table 4.1.

In general, we can combine the seasonal and nonseasonal operators into a *multiplicative seasonal autoregressive moving average model*, denoted by $\text{ARMA}(p,q) \times (P,Q)_s$, and write

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t \tag{5.17}$$

as the overall model. Although the diagnostic properties in Table 5.1 are not strictly true for the overall mixed model, the behavior of the ACF and PACF tends to show rough patterns of the indicated form. In fact, for mixed models, we tend to see a mixture of the facts listed in Table 4.1 and Table 5.1.

Example 5.12. A Mixed Seasonal Model

Consider an $\text{ARMA}(p = 0, q = 1) \times (P = 1, Q = 0)_{s=12}$ model

$$x_t = \Phi x_{t-12} + w_t + \theta w_{t-1},$$

where $|\Phi| < 1$ and $|\theta| < 1$. Then, because x_{t-12}, w_t, and w_{t-1} are uncorrelated, and x_t is stationary, $\gamma(0) = \Phi^2\gamma(0) + \sigma_w^2 + \theta^2\sigma_w^2$, or

$$\gamma(0) = \frac{1+\theta^2}{1-\Phi^2}\,\sigma_w^2.$$

Multiplying the model by x_{t-h}, $h > 0$, and taking expectations, we have $\gamma(1) = \Phi\gamma(11) + \theta\sigma_w^2$, and $\gamma(h) = \Phi\gamma(h-12)$, for $h \geq 2$. Thus, the model ACF is

$$\rho(12h) = \Phi^h \quad h = 1, 2, \ldots.$$

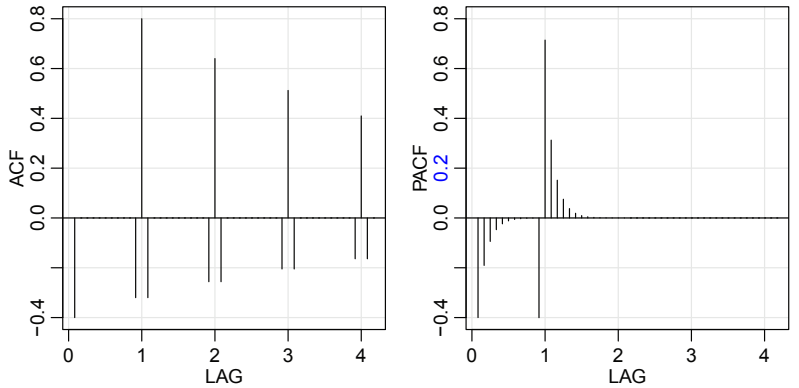

Figure 5.10 *ACF and PACF of the mixed seasonal ARMA model* $x_t = .8x_{t-12} + w_t - .5w_{t-1}$.

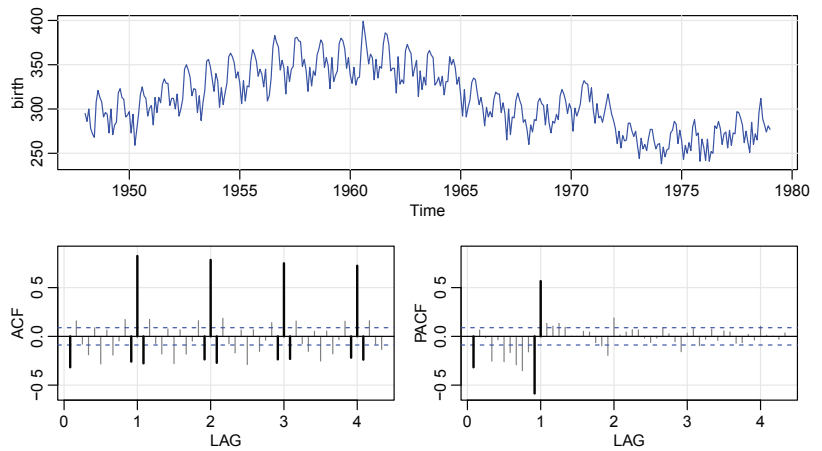

Figure 5.11 *Monthly live births in thousands for the United States during the "baby boom," 1948–1979. Sample ACF and PACF of the data with certain lags highlighted.*

$$\begin{aligned} \rho(12h-1) &= \rho(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h \quad h = 0, 1, 2, \ldots, \\ \rho(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

The ACF and PACF for this model with $\Phi = .8$ and $\theta = -.5$ are shown in Figure 5.10. These types of correlation relationships, although idealized here, are typically seen with seasonal data.

To compare these results to actual data, consider the seasonal series `birth`, which are the monthly live births in thousands for the United States surrounding the "baby boom." The data are plotted in Figure 5.11. Also shown in the figure are the sample ACF and PACF of the growth rate in births. We have highlighted certain values so that it may be compared to the idealized case in Figure 5.10.

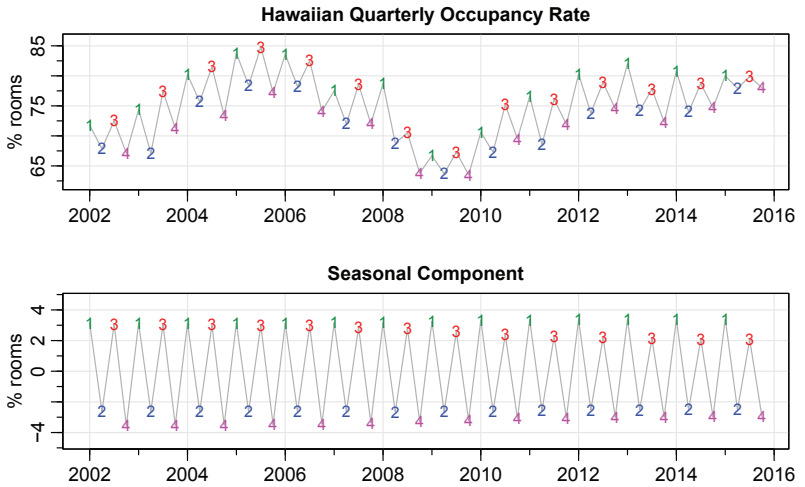

Figure 5.12 *Seasonal persistence: The quarterly occupancy rate of Hawaiian hotels and the extracted seasonal component, say $S_t \approx S_{t-4}$, where t is in quarters.*

```
##-- Figure 5.10 --##
phi  = c(rep(0,11),.8)
ACF  = ARMAacf(ar=phi, ma=-.5, 50)[-1]
PACF = ARMAacf(ar=phi, ma=-.5, 50, pacf=TRUE)
LAG  = 1:50/12
par(mfrow=c(1,2))
plot(LAG, ACF,  type="h", ylim=c(-.4,.8), panel.first=Grid())
abline(h=0)
plot(LAG, PACF, type="h", ylim=c(-.4,.8), panel.first=Grid())
abline(h=0)
##--  birth series --##
tsplot(birth)           # monthly number of births in US
acf2( diff(birth) )     # P/ACF of the differenced birth rate
```

◊

Seasonal persistence occurs when the process is nearly constant in the season. For example, consider the quarterly occupancy rate of Hawaiian hotels shown in Figure 5.12. The seasonal component from structural model fit is shown below the data; recall Example 3.20. Note that the occupancy rate for the first and third quarters is always up 2% to 4%, while the occupancy rate for the second and fourth quarters is always down 2% to 4%. In this case, we might think of the seasonal component, say S_t, as satisfying $S_t \approx S_{t-4}$, or

$$S_t = S_{t-4} + v_t\,,$$

where v_t is white noise.

```
x = window(hor, start=2002)
```

```
par(mfrow = c(2,1))
tsplot(x, main="Hawaiian Quarterly Occupancy Rate", ylab=" % rooms",
           ylim=c(62,86), col=gray(.7))
text(x, labels=1:4, col=c(3,4,2,6), cex=.8)
Qx = stl(x,15)$time.series[,1]
tsplot(Qx, main="Seasonal Component", ylab=" % rooms",
           ylim=c(-4.7,4.7), col=gray(.7))
text(Qx, labels=1:4, col=c(3,4,2,6), cex=.8)
```

The tendency of data to follow this type of behavior will be exhibited in a sample ACF that is large and decays very slowly at lags $h = sk$, for $k = 1, 2, \ldots$. In the occupancy rate example, suppose x_t is the rate with the trend component removed, then a reasonable model might be

$$x_t = S_t + w_t\,,$$

where w_t is white noise. If we subtract the effect of successive years from each other, we find that, with $s = 4$,

$$\begin{aligned}(1 - B^s)x_t = x_t - x_{t-4} &= S_t + w_t - (S_{t-4} + w_{t-4}) \\ &= (S_t - S_{t-4}) + w_t - w_{t-4} = v_t + w_t - w_{t-4}\,,\end{aligned}$$

is stationary and its ACF will have a peak only at lag $s = 4$.

In general, seasonal differencing is indicated when the ACF decays slowly at multiples of some season s. Then, a *seasonal difference of order D* is defined as

$$\nabla_s^D x_t = (1 - B^s)^D x_t, \tag{5.18}$$

where $D = 1, 2, \ldots$, takes positive integer values. Typically, $D = 1$ is sufficient to obtain seasonal stationarity. Incorporating these ideas into a general model leads to the following definition.

Definition 5.13. *The multiplicative* **seasonal autoregressive integrated moving average** *model, or* **SARIMA** *model is given by*

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \alpha + \Theta_Q(B^s)\theta(B)w_t, \tag{5.19}$$

where w_t is the usual Gaussian white noise process. The general model is denoted as $\mathbf{ARIMA}(p, d, q) \times (P, D, Q)_s$. *The ordinary autoregressive and moving average components are represented by $\phi(B)$ and $\theta(B)$ of orders p and q, respectively, and the seasonal autoregressive and moving average components by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ of orders P and Q and ordinary and seasonal difference components by $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$.*

Example 5.14. An SARIMA Model

Consider the following model, which often provides a reasonable representation for seasonal, nonstationary, economic time series. We exhibit the equations for the model, denoted by $\text{ARIMA}(0, 1, 1) \times (0, 1, 1)_{12}$ in the notation given above, where

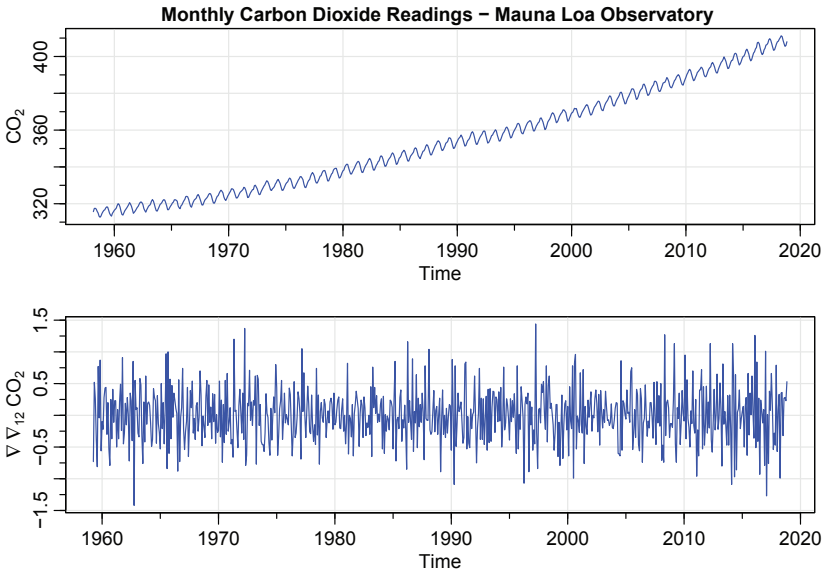

Figure 5.13 *Monthly CO_2 levels (ppm) taken at the Mauna Loa, Hawaii observatory (top) and the data differenced to remove trend and seasonal persistence (bottom).*

the seasonal fluctuations occur every 12 months. Then, with $\alpha = 0$, the model (5.19) becomes

$$\nabla_{12}\nabla x_t = \Theta(B^{12})\theta(B)w_t$$

or

$$(1 - B^{12})(1 - B)x_t = (1 + \Theta B^{12})(1 + \theta B)w_t. \tag{5.20}$$

Expanding both sides of (5.20) leads to the representation

$$(1 - B - B^{12} + B^{13})x_t = (1 + \theta B + \Theta B^{12} + \Theta\theta B^{13})w_t,$$

or in difference equation form

$$x_t = x_{t-1} + x_{t-12} - x_{t-13} + w_t + \theta w_{t-1} + \Theta w_{t-12} + \Theta\theta w_{t-13}.$$

Note that the multiplicative nature of the model implies that the coefficient of w_{t-13} is the product of the coefficients of w_{t-1} and w_{t-12} rather than a free parameter. The multiplicative model assumption seems to work well with many seasonal time series data sets while reducing the number of parameters that must be estimated. ◊

Selecting the appropriate model for a given set of data is a simple step-by-step process. First, consider obvious differencing transformations to remove trend (d) and to remove seasonal persistence (D) if they are present. Then look at the ACF and the PACF of the possibly differenced data. Consider the seasonal components (P and Q) by looking at the seasonal lags only and keeping Table 5.1 in mind. Then look at the first few lags and consider values for within seasonal components (p and q) keeping Table 4.1 in mind.

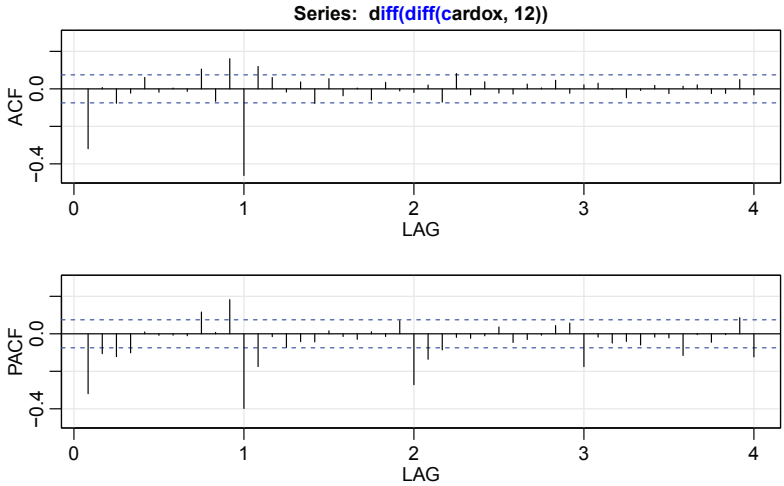

Figure 5.14 *Sample ACF and PACF of the differenced CO_2 data.*

Example 5.15. Carbon Dioxide and Global Warming

Concentration of CO_2 in the atmosphere, which is the primary cause of global warming, has now reached an unprecedented level. In March 2015, the average of all of the global measuring sites showed a concentration above 400 parts per million (ppm). This follows the individual observatory high points of 400 ppm in 2012 at the Barrow observatory in Alaska, and the 2013 high of 400 ppm at the Mauna Loa observatory in Hawaii. Mauna Loa has been running consistently above 400 ppm since late 2015. Scientists advising the United Nations recommend the world should act to keep the CO_2 levels below 400-450 ppm in order to prevent even more irreversible and disastrous climate change effects.

The data shown in Figure 5.13 are the CO_2 readings, say x_t, from March 1958 to November 2018 at the Mauna Loa observatory, which is the oldest continuous monitoring station of carbon dioxide. The trend and seasonal persistence are evident in the plot, so we also exhibit the trend and seasonally differenced data, $\nabla\nabla_{12}x_t$, in the figure. The data are in `cardox`.[1]

```
par(mfrow=c(2,1))
tsplot(cardox, col=4, ylab=expression(CO[2]))
title("Monthly Carbon Dioxide Readings - Mauna Loa Observatory",
            cex.main=1)
tsplot(diff(diff(cardox,12)), col=4,
            ylab=expression(nabla~nabla[12]~CO[2]))
```

The sample ACF and PACF of the differenced data are shown in Figure 5.14.

```
acf2(diff(diff(cardox,12)))
```

[1]The R datasets package already has data sets with names `co2`, which are the same data but only until 1997, and `CO2`, which is unrelated to this example.

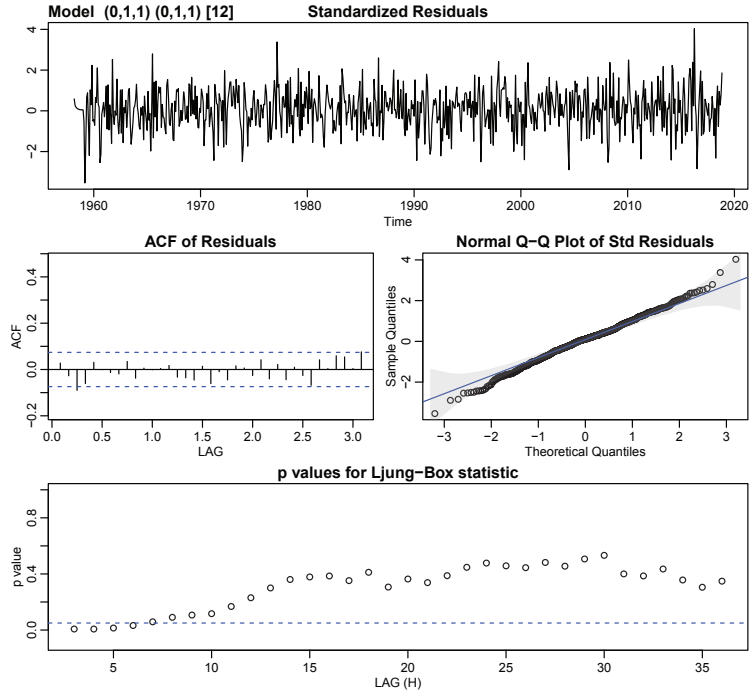

Figure 5.15 *Residual analysis for the ARIMA*$(0,1,1)\times(0,1,1)_{12}$ *fit to the* CO_2 *data set.*

Seasonal: It appears that at the seasons, the ACF is cutting off a lag $1s$ ($s = 12$), whereas the PACF is tailing off at lags $1s, 2s, 3s, 4s$. These results imply an SMA(1), $P = 0$, $Q = 1$, in the seasonal component.

Non-Seasonal: Inspecting the sample ACF and PACF at the first few lags, it appears as though the ACF cuts off at lag 1, whereas the PACF is tailing off. This suggests an MA(1) within the seasons, $p = 0$ and $q = 1$.

Thus, we first try an ARIMA$(0,1,1)\times(0,1,1)_{12}$ on the CO_2 data:

```
sarima(cardox, p=0,d=1,q=1, P=0,D=1,Q=1,S=12)
      Estimate      SE   t.value  p.value
 ma1   -0.3875  0.0390   -9.9277        0
 sma1  -0.8641  0.0192  -45.1205        0
 --
 sigma^2 estimated as 0.09634
 $AIC: 0.5174486  $AICc: 0.5174712  $BIC: 0.5300457
```

The residual analysis is exhibited in Figure 5.15 and the results look decent, however, there may still be a small amount of autocorrelation remaining in the residuals.

The next step is to add a parameter to the within-seasons component. In this case, adding another MA parameter ($q = 2$) gives non-significant results. However, adding an AR parameter does yield significant results.

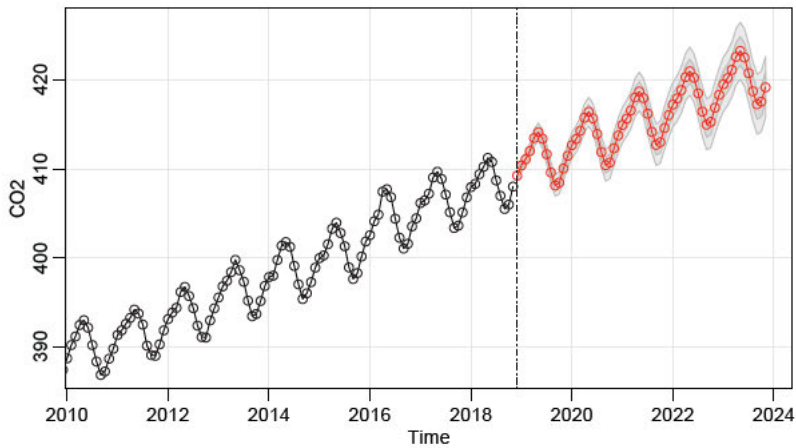

Figure 5.16 *Five-year-ahead forecasts using the ARIMA*$(1,1,1) \times (0,1,1)_{12}$ *model on the Mauna Loa carbon dioxide readings.*

```
sarima(cardox, 1,1,1, 0,1,1,12)
      Estimate      SE   t.value  p.value
 ar1    0.1941  0.0953    2.0374   0.042
 ma1   -0.5578  0.0813   -6.8634   0.000
 sma1  -0.8648  0.0189  -45.7161   0.000
 --
 sigma^2 estimated as 0.09585
 $AIC: 0.5152905  $AICc: 0.5153359  $BIC: 0.5341862
```

The residual analysis (not shown) indicates an improvement to the fit. We do note that while the AIC and AICc prefer the second model, the BIC prefers the first model. In addition, there is a substantial difference in the MA(1) parameter estimate and its standard error. In the final analysis, the predictions from the two models will be close, so we will use the second model for forecasting.

The forecasts out five years are shown in Figure 5.16.

```
sarima.for(cardox, 60, 1,1,1, 0,1,1,12)
abline(v=2018.9, lty=6)
##-- for comparison, try the first model --##
sarima.for(cardox, 60, 0,1,1, 0,1,1,12)  # not shown
```

It is clear that without intervention, atmospheric CO_2 concentrations will continue to grow to dangerous levels. Unfortunately, the carbon dioxide that we have released will remain in the atmosphere for thousands of years. Only after many millennia will it return to rocks, for example, through the formation of calcium carbonate. Once released, carbon dioxide is in our environment essentially forever. It does not go away, unless we, ourselves, remove it. ◊

5.4 Regression with Autocorrelated Errors *

In Section 3.1, we covered classical regression with uncorrelated errors w_t. In this section, we discuss the modifications that might be considered when the errors are correlated. That is, consider the regression model

$$y_t = \beta_1 z_{t1} + \cdots + \beta_r z_{tr} + x_t = \sum_{j=1}^{r} \beta_j z_{tj} + x_t \tag{5.21}$$

where x_t is a process with some covariance function $\gamma_x(s,t)$. In ordinary least squares, the assumption is that x_t is white Gaussian noise, in which case $\gamma_x(s,t) = 0$ for $s \neq t$ and $\gamma_x(t,t) = \sigma^2$, independent of t. If this is not the case, then weighted least squares should be used.

In the time series case, it is often possible to assume a stationary covariance structure for the error process x_t that corresponds to a linear process and try to find an ARMA representation for x_t. For example, if we have a pure AR(p) error, then

$$\phi(B)x_t = w_t,$$

and $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the linear transformation that, when applied to the error process, produces the white noise w_t. Multiplying the regression equation through by the transformation $\phi(B)$ yields,

$$\underbrace{\phi(B)y_t}_{y_t^*} = \beta_1 \underbrace{\phi(B)z_{t1}}_{z_{t1}^*} + \cdots + \beta_r \underbrace{\phi(B)z_{tr}}_{z_{tr}^*} + \underbrace{\phi(B)x_t}_{w_t},$$

and we are back to the linear regression model where the observations have been transformed so that $y_t^* = \phi(B)y_t$ is the dependent variable, $z_{tj}^* = \phi(B)z_{tj}$ for $j = 1, \ldots, r$, are the independent variables, but the βs are the same as in the original model. For example, suppose we have the regression model

$$y_t = \alpha + \beta z_t + x_t$$

where $x_t = \phi x_{t-1} + w_t$ is AR(1). Then, transform the data as $y_t^* = y_t - \phi y_{t-1}$ and $z_t^* = z_t - \phi z_{t-1}$ so that the new model is

$$\underbrace{y_t - \phi y_{t-1}}_{y_t^*} = \underbrace{(1-\phi)\alpha}_{\alpha^*} + \underbrace{\beta(z_t - \phi z_{t-1})}_{\beta z_t^*} + \underbrace{(x_t - \phi x_{t-1})}_{w_t}$$

In the AR case, we may set up the least squares problem as minimizing the error sum of squares

$$S(\phi, \beta) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} \Big[\phi(B)y_t - \sum_{j=1}^{r} \beta_j \phi(B) z_{tj}\Big]^2$$

with respect to all the parameters, $\phi = \{\phi_1, \ldots, \phi_p\}$ and $\beta = \{\beta_1, \ldots, \beta_r\}$. Of course, this is done using numerical methods.

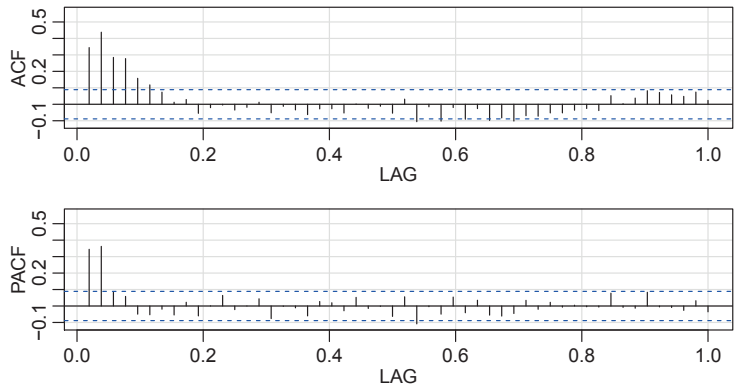

Figure 5.17 *Sample ACF and PACF of the mortality residuals indicating an AR(2) process.*

If the error process is ARMA(p,q), i.e., $\phi(B)x_t = \theta(B)w_t$, then in the above discussion, we transform by $\pi(B)x_t = w_t$ (the π-weights are functions of the ϕs and θs, see Section D.2). In this case the error sum of squares also depends on $\theta = \{\theta_1, \ldots, \theta_q\}$:

$$S(\phi, \theta, \beta) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} \Big[\pi(B)y_t - \sum_{j=1}^{r} \beta_j \pi(B) z_{tj}\Big]^2$$

At this point, the main problem is that we do not typically know the behavior of the noise x_t prior to the analysis. An easy way to tackle this problem was first presented in Cochrane and Orcutt (1949), and with the advent of cheap computing can be modernized.

(i) First, run an ordinary regression of y_t on $z_{t1}, \ldots, z_{tr}$ (acting as if the errors are uncorrelated). Retain the residuals, $\hat{x}_t = y_t - \sum_{j=1}^{r} \hat{\beta}_j z_{tj}$.

(ii) Identify an ARMA model for the residuals $\hat{x}_t$. There may be competing models.

(iii) Run weighted least squares (or MLE) on the regression model(s) with autocorrelated errors using the model(s) specified in step (ii).

(iv) Inspect the residuals $\hat{w}_t$ for whiteness, and adjust the model if necessary.

Example 5.16. Mortality, Temperature, and Pollution

We consider the analyses presented in Example 3.5 relating mean adjusted temperature T_t, and particulate pollution levels P_t to cardiovascular mortality M_t. We consider the regression model

$$M_t = \beta_0 + \beta_1 t + \beta_2 T_t + \beta_3 T_t^2 + \beta_4 P_t + x_t, \tag{5.22}$$

where, for now, we assume that x_t is white noise. The sample ACF and PACF of the residuals from the ordinary least squares fit of (5.22) are shown in Figure 5.17, and

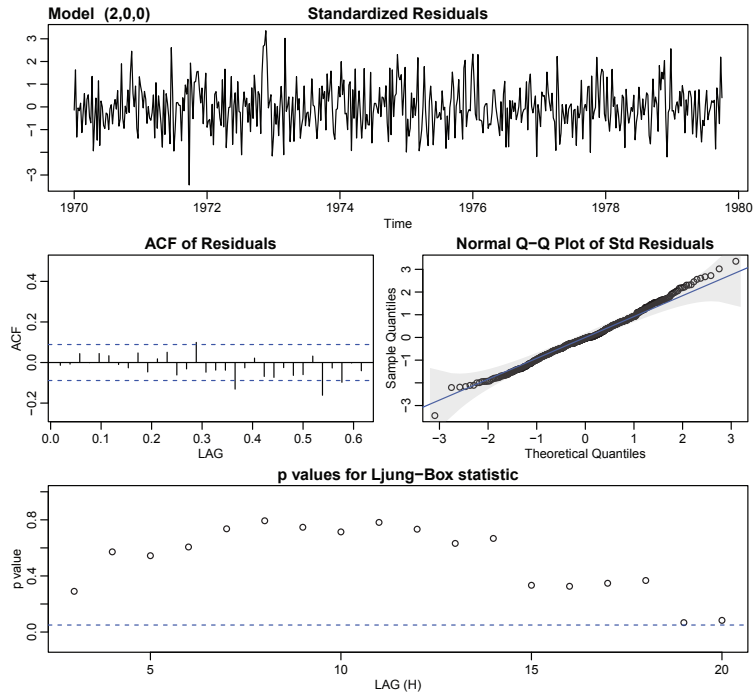

Figure 5.18 *Diagnostics for the regression of mortality on temperature and particulate pollution with autocorrelated errors, Example 5.16.*

the results suggest an AR(2) model for the residuals. The next step is to fit the model (5.22) where x_t is AR(2), $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$ and w_t is white noise. The model can be fit using `sarima` as follows.

```
trend = time(cmort); temp = tempr - mean(tempr); temp2 = temp^2
fit = lm(cmort~trend + temp + temp2 + part, na.action=NULL)
acf2(resid(fit), 52)   # implies AR2
sarima(cmort, 2,0,0, xreg=cbind(trend, temp, temp2, part) )
              Estimate       SE t.value p.value
   ar1          0.3848   0.0436  8.8329  0.0000
   ar2          0.4326   0.0400 10.8062  0.0000
   intercept 3075.1482 834.7157  3.6841  0.0003
   trend       -1.5165   0.4226 -3.5882  0.0004
   temp        -0.0190   0.0495 -0.3837  0.7014
   temp2        0.0154   0.0020  7.6117  0.0000
   part         0.1545   0.0272  5.6803  0.0000
   sigma^2 estimated as 26.01
```

The residual analysis output from `sarima` shown in Figure 5.18 shows no obvious departure of the residuals from whiteness. Also, note that `temp`, T_t, is not significant, but has been centered, $T_t = {}^\circ\mathrm{F}_t - {}^\circ\overline{\mathrm{F}}$ where ${}^\circ\mathrm{F}_t$ is the actual temperature measured in

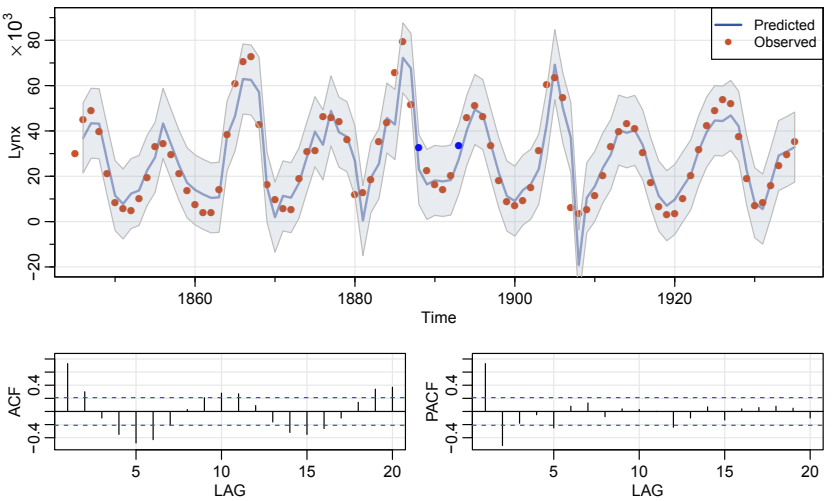

Figure 5.19 *Top: Observed lynx population size (points) and one-year-ahead prediction (line) with ±2 root MSPE (ribbon). Bottom: ACF and PACF of the residuals from* (5.23).

degrees Fahrenheit. Thus `temp2` is $T_t^2 = (^\circ\mathrm{F}_t - ^\circ\overline{\mathrm{F}})^2$, so a linear term for temperature is in the model twice and $^\circ\overline{F}$ was chosen arbitrarily. As is generally true, it's better to leave lower-order terms in the regression to allow more flexibility in the model. ♢

Example 5.17. Lagged Regression: Lynx–Hare Populations

In Example 1.5, we discussed the predator–prey relationship between the lynx and the snowshoe hare populations. Recall that the lynx population rises and falls with that of the hare, even though other food sources may be abundant. In this example, we consider the snowshoe hare population as a leading indicator of the lynx population,

$$L_t = \beta_0 + \beta_1 H_{t-1} + x_t\,, \tag{5.23}$$

where L_t is the lynx population and H_t is the hare population in year t. We anticipate that x_t will be autocorrelated error.

After first fitting OLS, we plotted the sample P/ACF of the residuals, which are shown in the lower part of Figure 5.19. These indicate an AR(2) for the residual process, which was then fit using `sarima`. The residual analysis (not shown) looks good, so we have our final model. The final model was then used to obtain the one-year-ahead predictions of the lynx population, $\hat{L}_t^{t-1}$, which are displayed at the top of Figure 5.19 along with the observations. We note that the model does a good job in predicting the lynx population size one year in advance. The R code for this example, along with some output follows:

```
library(zoo)
lag2.plot(Hare, Lynx, 5)      # lead-lag relationship
pp = as.zoo(ts.intersect(Lynx, HareL1 = lag(Hare,-1)))
```

```
summary(reg <- lm(pp$Lynx~ pp$HareL1))  # results not displayed
acf2(resid(reg))                        # in Figure 5.19
( reg2 = sarima(pp$Lynx, 2,0,0, xreg=pp$HareL1 ))
            Estimate     SE t.value p.value
  ar1         1.3258 0.0732 18.1184  0.0000
  ar2        -0.7143 0.0731 -9.7689  0.0000
  intercept  25.1319 2.5469  9.8676  0.0000
  xreg        0.0692 0.0318  2.1727  0.0326
  sigma^2 estimated as 59.57
prd = Lynx - resid(reg2$fit)   # prediction (resid = obs - pred)
prde = sqrt(reg2$fit$sigma2)   # prediction error
tsplot(prd, lwd=2, col=rgb(0,0,.9,.5), ylim=c(-20,90), ylab="Lynx")
points(Lynx, pch=16, col=rgb(.8,.3,0))
     x = time(Lynx)[-1]
    xx = c(x, rev(x))
    yy = c(prd - 2*prde, rev(prd + 2*prde))
polygon(xx, yy, border=8, col=rgb(.4, .5, .6, .15))
mtext(expression(""%*% 10^3), side=2, line=1.5, adj=.975)
legend("topright", legend=c("Predicted", "Observed"), lty=c(1,NA),
           lwd=2, pch=c(NA,16), col=c(4,rgb(.8,.3,0)), cex=.9)
```

◊

Problems

5.1. For the logarithm of the glacial varve data, say, x_t, presented in Example 4.27, use the first 100 observations and calculate the EWMA, x^n_{n+1}, discussed in Example 5.5, for $n = 1, \ldots, 100$, using $\lambda = .25, .50$, and $.75$, and plot the EWMAs and the data superimposed on each other. Comment on the results.

5.2. In Example 5.6, we fit an ARIMA model to the quarterly GNP series. Repeat the analysis for the US GDP series in `gdp`. Discuss all aspects of the fit as specified in the points at the beginning of Section 5.2 from plotting the data to diagnostics and model choice.

5.3. Crude oil prices in dollars per barrel are in `oil`. Fit an ARIMA(p, d, q) model to the growth rate performing all necessary diagnostics. Comment.

5.4. Fit an ARIMA(p, d, q) model to `gtemp_land`, the land-based global temperature data, performing all of the necessary diagnostics; include a model choice analysis. After deciding on an appropriate model, forecast (with limits) the next 10 years. Comment.

5.5. Repeat Problem 5.4 using the ocean based data in `gtemp_ocean`.

5.6. One of the series collected along with particulates, temperature, and mortality described in Example 3.5 is the sulfur dioxide series, `so2`. Fit an ARIMA(p, d, q) model to the data, performing all of the necessary diagnostics. After deciding on an appropriate model, forecast the data into the future four time periods ahead (about

one month) and calculate 95% prediction intervals for each of the four forecasts. Comment.

5.7. Fit a seasonal ARIMA model to the R data set `AirPassengers`, which are the monthly totals of international airline passengers taken from Box and Jenkins (1970).

5.8. Plot the theoretical ACF of the seasonal ARIMA$(0,1) \times (1,0)_{12}$ model with $\Phi = .8$ and $\theta = .5$ out to lag 50.

5.9. Fit a seasonal ARIMA model of your choice to the chicken price data in `chicken`. Use the estimated model to forecast the next 12 months.

5.10. Fit a seasonal ARIMA model of your choice to the unemployment data, `UnempRate`. Use the estimated model to forecast the next 12 months.

5.11. Fit a seasonal ARIMA model of your choice to the U.S. Live Birth Series, `birth`. Use the estimated model to forecast the next 12 months.

5.12. Fit an appropriate seasonal ARIMA model to the log-transformed Johnson & Johnson earnings series (`jj`) of Example 1.1. Use the estimated model to forecast the next 4 quarters.

5.13.* Let S_t represent the monthly sales data in `sales` ($n = 150$), and let L_t be the leading indicator in `lead`.

(a) Fit an ARIMA model to S_t, the monthly sales data. Discuss your model fitting in a step-by-step fashion, presenting your (A) initial examination of the data, (B) transformations and differencing orders, if necessary, (C) initial identification of the dependence orders, (D) parameter estimation, (E) residual diagnostics and model choice.

(b) Use the CCF and lag plots between ∇S_t and ∇L_t to argue that a regression of ∇S_t on ∇L_{t-3} is reasonable. [*Note:* In `lag2.plot()`, the first named series is the one that gets lagged.]

(c) Fit the regression model $\nabla S_t = \beta_0 + \beta_1 \nabla L_{t-3} + x_t$, where x_t is an ARMA process (explain how you decided on your model for x_t). Discuss your results.

5.14.* One of the remarkable technological developments in the computer industry has been the ability to store information densely on a hard drive. In addition, the cost of storage has steadily declined causing problems of *too much data* as opposed to *big data*. The data set for this assignment is `cpg`, which consists of the median annual retail price per GB of hard drives, say c_t, taken from a sample of manufacturers from 1980 to 2008.

(a) Plot c_t and describe what you see.

(b) Argue that the curve c_t versus t behaves like $c_t \approx \alpha e^{\beta t}$ by fitting a linear regression of $\log c_t$ on t and then plotting the fitted line to compare it to the logged data. Comment.

(c) Inspect the residuals of the linear regression fit and comment.

(d) Fit the regression again, but now using the fact that the errors are autocorrelated. Comment.

5.15.* Redo Problem 3.2 without assuming the error term is white noise.

5.16.* In Example 3.14 we fit the model

$$R_t = \beta_0 + \beta_1 S_{t-6} + \beta_2 D_{t-6} + \beta_3 D_{t-6} S_{t-6} + w_t,$$

where R_t is Recruitment, S_t is SOI, and D_t is a dummy variable that is 0 if $S_t < 0$ and 1 otherwise. However, residual analysis indicated that the residuals are not white noise.

(a) Plot the ACF and PACF of the residuals and discuss why an AR(2) model might be appropriate.

(b) Fit the dummy variable regression model assuming that the noise is correlated noise and compare your results to the results of Example 3.14 (compare the estimated parameters and the corresponding standard errors).

(c) Now fit a seasonal model for the noise in the previous part.

5.17. In this problem we show how to verify that IMA(1,1) model given in (5.7) leads to EWMA forecasting shown in (5.8). Most of the details are given here, the exercise is to verify (5.24) and (5.25) below.

Write $y_t = x_t - x_{t-1}$ so that $y_t = w_t - \lambda w_{t-1}$. Because $|\lambda| < 1$, there is an invertible representation,

$$w_t = \sum_{j=0}^{\infty} \lambda^j y_{t-j}\,.$$

Replace y_t by $x_t - x_{t-1}$ and simplify to get

$$x_t = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{t-j} + w_t\,, \tag{5.24}$$

supposing that we have an infinite history available. Using (5.24),

$$x_n^{n-1} = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n-j}$$

because $w_n^{n-1} = 0$. Consequently,

$$x_{n+1}^n = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n+1-j} = (1-\lambda)x_n + \lambda x_n^{n-1}\,. \tag{5.25}$$

The mean-square prediction error can be approximated using (5.3) by noting that $\psi(z) = (1-\lambda z)/(1-z) = 1 + (1-\lambda)\sum_{j=1}^{\infty} z^j$ for $|z| < 1$. Thus, for large n, (5.3) leads to (5.9).