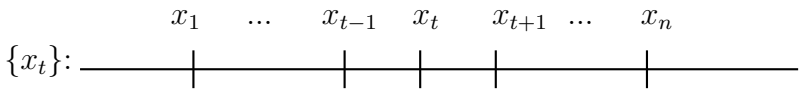
STAT 626: Outline of Lecture 1

Time Series Data: Examples (§1.2)

1. A Quick Review of the Syllabus,

2. **Time Series Data:** $x_1, x_2, \ldots, x_n$.

   A variable measured over time:



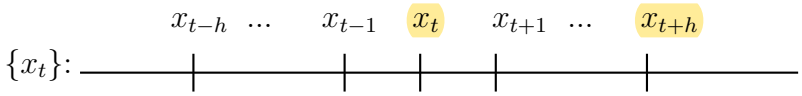
   Time Series Plot: Plot of x_t vs time should reveal some patterns over time like:

   Trend, Cycle, Variability, Dependence,

   Exploratory Data Analysis: Chapters 1-3.

STATIONARY TIME SERIES, Chaps 4-5

Going beyond sample data or independent and identically distributed (i.i.d.) rvs.



Autocovariance Function (ACF) & PACF:

$$\gamma(h) = \text{cov}(x_{t+h}, x_t), \quad h = 0, 1, \ldots.$$

Correlogram: Plot of Correlations vs Lags,

3. Autoregressive Integrated Moving Average (ARIMA) Models.

Goal(s) of Time Series Analysis: Forecasting, Detection,.....

Chapter 1

Time Series Elements

1.1 Introduction

The analysis of data observed at different time points leads to unique problems that are not covered by classical statistics. The dependence introduced by the sampling data over time restricts the applicability of many conventional statistical methods that require random samples. The analysis of such data is commonly referred to as *time series analysis*.

To provide a statistical setting for describing the elements of time series data, the data are represented as a collection of random variables indexed according to the order they are obtained in time. For example, if we collect data on daily high temperatures in your city, we may consider the time series as a sequence of random variables, $x_1, x_2, x_3, \ldots$, where the random variable x_1 denotes the high temperature on day one, the variable x_2 denotes the value for the second day, x_3 denotes the value for the third day, and so on. In general, a collection of random variables, $\{x_t\}$, indexed by t is referred to as a *stochastic process*. In this text, t will typically be discrete and vary over the integers $t = 0, \pm 1, \pm 2, \ldots$ or some subset of the integers, or a similar index like months of a year.

Historically, time series methods were applied to problems in the physical and environmental sciences. This fact accounts for the engineering nomenclature that permeates the language of time series analysis. The first step in an investigation of time series data involves careful scrutiny of the recorded data plotted over time. Before looking more closely at the particular statistical methods, we mention that two separate, but not mutually exclusive, approaches to time series analysis exist, commonly identified as the *time domain approach* (Chapter 4 and 5) and the *frequency domain approach* (Chapter 6 and 7).

1.2 Time Series Data

The following examples illustrate some of the common kinds of time series data as well as some of the statistical questions that might be asked about such data.

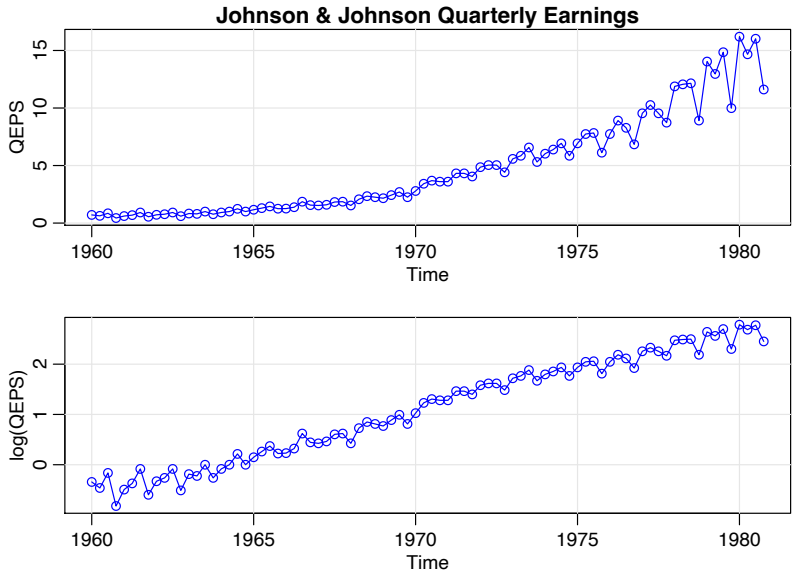


Figure 1.1 *Johnson & Johnson quarterly earnings per share, 1960-I to 1980-IV [top]. The same data logged [bottom].*

Example 1.1. Johnson & Johnson Quarterly Earnings

Figure 1.1 shows quarterly earnings per share (QEPS) for the U.S. company Johnson & Johnson and the data transformed by taking logs. There are 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980. Modeling such series begins by observing the primary patterns in the time history. In this case, note the increasing underlying trend and variability, and a somewhat regular oscillation superimposed on the trend that seems to repeat over quarters. Methods for analyzing data such as these are explored in Chapter 3 (see Problem 3.1) using regression techniques.

If we consider the data as being generated as a small percentage change each year, say r_t (which can be negative), we might write $x_t = (1 + r_t)x_{t-4}$, where x_t is the QEPS for quarter t. If we log the data, then $\log(x_t) = \log(1 + r_t) + \log(x_{t-4})$, implying a linear growth rate; i.e., this quarter's value is the same as last year plus a small amount, $\log(1 + r_t)$. This attribute of the data is displayed by the bottom plot of Figure 1.1.

The R code to plot the data for this example is,[1]

```
library(astsa)    # we leave this line off subsequent examples
par(mfrow=2:1)
tsplot(jj, ylab="QEPS", type="o", col=4, main="Johnson & Johnson
          Quarterly Earnings")
tsplot(log(jj), ylab="log(QEPS)", type="o", col=4)
```

◊

[1]We assume `astsa` version 1.8.6 or later has been installed; see Section A.2

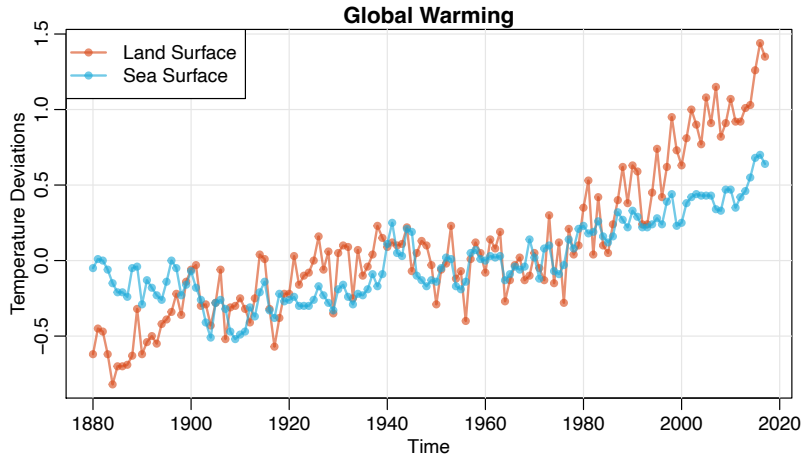


Figure 1.2 *Yearly average global land surface and ocean surface temperature deviations (1880–2017) in °C.*

Example 1.2. Global Warming and Climate Change
Two global temperature records are shown in Figure 1.2. The data are (1) annual temperature anomalies averaged over the Earth's land area, and (2) sea surface temperature anomalies averaged over the part of the ocean that is free of ice at all times (open ocean). The time period is 1880 to 2017 and the values are deviations (°C) from the 1951-1980 average, updated from Hansen et al. (2006). The upward trend in both series during the latter part of the twentieth century has been used as an argument for the climate change hypothesis. Note that the trend is not linear, with periods of leveling off and then sharp upward trends. It should be obvious that fitting a simple linear regression of the either series (x_t) on time (t), say $x_t = \alpha + \beta t + \epsilon_t$, would not yield an accurate description of the trend. Most climate scientists agree the main cause of the current global warming trend is human expansion of the *greenhouse effect*; see https://climate.nasa.gov/causes/. The R code for this example is:

```
culer = c(rgb(.85,.30,.12,.6), rgb(.12,.65,.85,.6))
tsplot(gtemp_land, col = culer[1], lwd=2, type="o", pch=20,
           ylab="Temperature Deviations", main="Global Warming")
lines(gtemp_ocean, col=culer[2], lwd=2, type="o", pch=20)
legend("topleft", col=culer, lty=1, lwd=2, pch=20, legend=c("Land
           Surface", "Sea Surface"), bg="white")
```

◊

Example 1.3. Dow Jones Industrial Average
As an example of financial time series data, Figure 1.3 shows the trading day closings and returns (or percent change) of the Dow Jones Industrial Average (DJIA) from 2006 to 2016. If x_t is the value of the DJIA closing on day t, then the return is

$$r_t = (x_t - x_{t-1})/x_{t-1}\,.$$

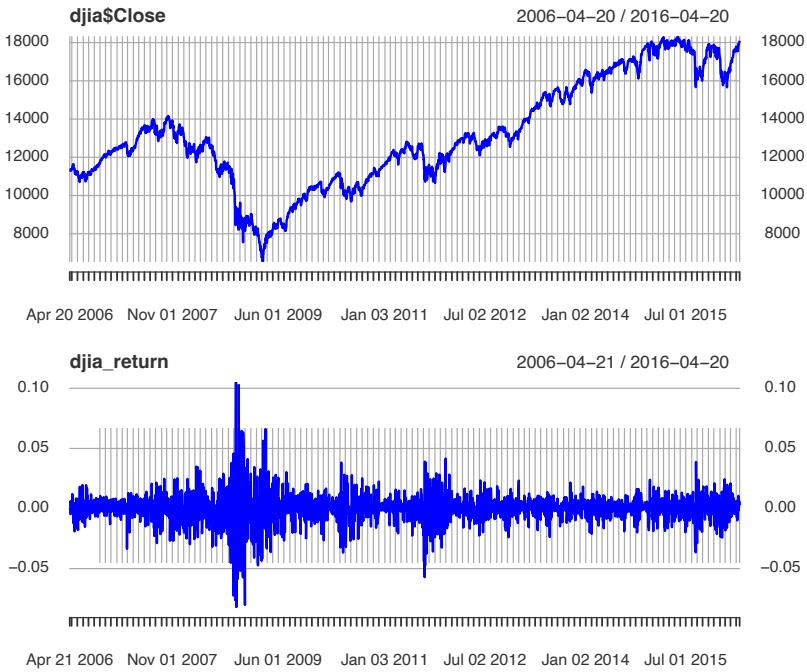


Figure 1.3 *Dow Jones Industrial Average (DJIA) trading days closings [top] and returns [bottom] from April 20, 2006 to April 20, 2016.*

This means that $1 + r_t = x_t / x_{t-1}$ and

$$\log(1 + r_t) = \log(x_t / x_{t-1}) = \log(x_t) - \log(x_{t-1}),$$

just as in Example 1.1. Noting the expansion

$$\log(1 + r) = r - \frac{r^2}{2} + \frac{r^3}{3} - \cdots \quad -1 < r \leq 1,$$

we see that if r is very small, the higher order terms will be negligible. Consequently, because for financial data, $x_t / x_{t-1} \approx 1$, we have

$$\log(1 + r_t) \approx r_t.$$

Note the financial crisis of 2008 in Figure 1.3. The data shown are typical of return data. The mean of the series appears to be stable with an average return of approximately zero, however, the *volatility* (or variability) of data exhibits clustering; that is, highly volatile periods tend to be clustered together. A problem in the analysis of these type of financial data is to forecast the volatility of future returns. Models have been developed to handle these problems; see Chapter 8. The data set is an `xts` data file, so it must be loaded.

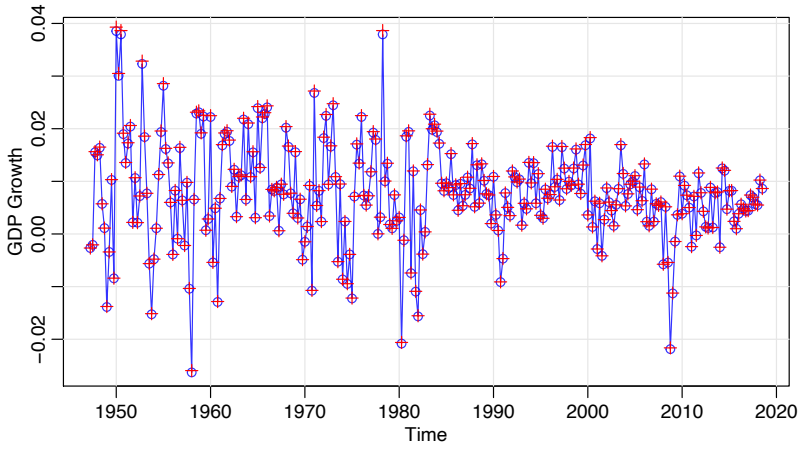


Figure 1.4: *US GDP growth rate calculated using logs (–○–) and actual values (+).*

```
library(xts)
djia_return = diff(log(djia$Close))[-1]
par(mfrow=2:1)
plot(djia$Close,  col=4)
plot(djia_return, col=4)
```

You can see a comparison of r_t and $\log(1+r_t)$ in Figure 1.4, which shows the seasonally adjusted quarterly growth rate, r_t, of US GDP compared to the version obtained by calculating the difference of the logged data.

```
tsplot(diff(log(gdp)), type="o", col=4, ylab="GDP Growth") # diff-log
points(diff(gdp)/lag(gdp,-1), pch=3, col=2)           # actual return
```

It turns out that many time series behave like this, so that logging the data and then taking successive differences is a standard data transformation in time series analysis. ◊

Example 1.4. El Niño – Southern Oscillation (ENSO)
The Southern Oscillation Index (SOI) measures changes in air pressure related to sea surface temperatures in the central Pacific Ocean. The central Pacific warms every three to seven years due to the ENSO effect, which has been blamed for various global extreme weather events. During El Niño, pressure over the eastern and western Pacific reverses, causing the trade winds to diminish and leading to an eastward movement of warm water along the equator. As a result, the surface waters of the central and eastern Pacific warm with far-reaching consequences to weather patterns.

Figure 1.5 shows monthly values of the Southern Oscillation Index (SOI) and associated Recruitment (an index of the number of new fish). Both series are for a period of 453 months ranging over the years 1950–1987. The series show an obvious annual cycle (hot in the summer, cold in the winter), and, though difficult to see, a slower frequency of three to seven years. The study of the kinds of cycles and their

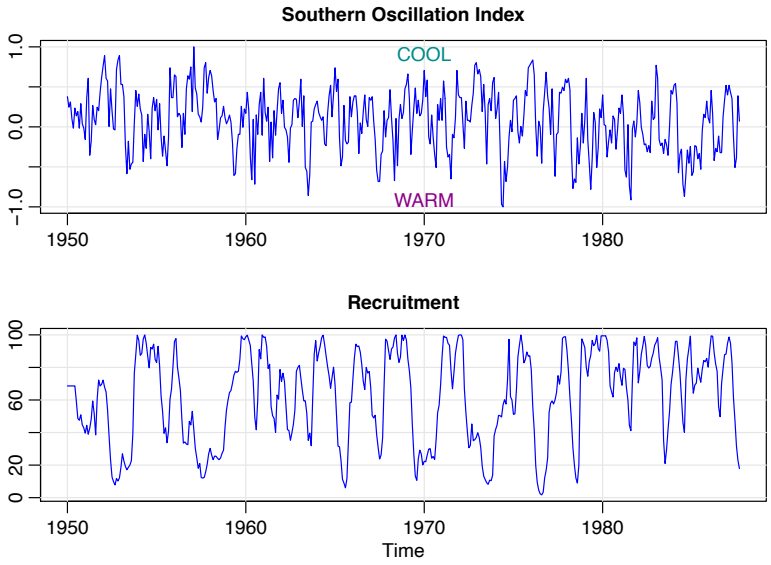


Figure 1.5: *Monthly SOI and Recruitment (estimated new fish), 1950-1987.*

strengths is the subject of Chapter 6 and 7. The two series are also related; it is easy to imagine that fish population size is dependent on the ocean temperature.

The following R code will reproduce Figure 1.5:

```
par(mfrow = c(2,1))
tsplot(soi, ylab="", xlab="", main="Southern Oscillation Index", col=4)
text(1970, .91, "COOL", col="cyan4")
text(1970,-.91, "WARM", col="darkmagenta")
tsplot(rec, ylab="", main="Recruitment", col=4)
```

◊

Example 1.5. Predator-Prey Interactions

While it is clear that predators influence the numbers of their prey, prey affect the number of predators because when prey become scarce, predators may die of starvation or fail to reproduce. Such relationships are often modeled by the Lotka–Volterra equations, which are a pair of simple nonlinear differential equations (e.g., see Edelstein-Keshet, 2005, Ch. 6).

One of the classic studies of predator-prey interactions is the snowshoe hare and lynx pelts purchased by the Hudson's Bay Company of Canada. While this is an indirect measure of predation, the assumption is that there is a direct relationship between the number of pelts collected and the number of hare and lynx in the wild. These predator-prey interactions often lead to cyclical patterns of predator and prey abundance seen in Figure 1.6. Notice that the lynx and hare population sizes are asymmetric in that they tend to increase slowly and decrease quickly ($\nearrow\!\!\downarrow$).

The lynx prey varies from small rodents to deer, with the snowshoe hare being

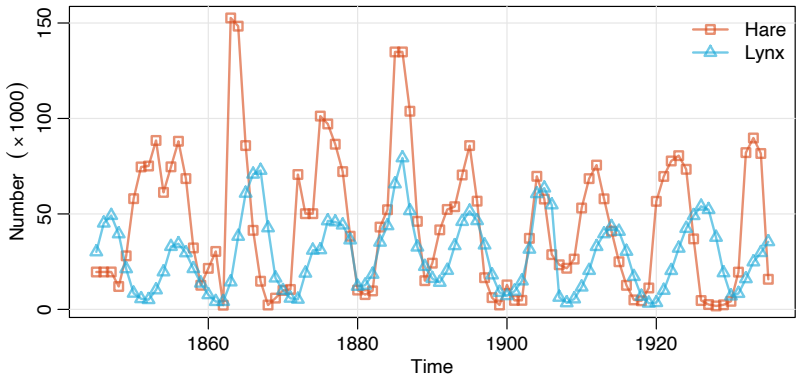


Figure 1.6 *Time series of the predator-prey interactions between the snowshoe hare and lynx pelts purchased by the Hudson's Bay Company of Canada. It is assumed there is a direct relationship between the number of pelts collected and the number of hare and lynx in the wild.*

its overwhelmingly favored prey. In fact, lynx are so closely tied to the snowshoe hare that its population rises and falls with that of the hare, even though other food sources may be abundant. In this case, it seems reasonable to model the size of the lynx population in terms of the snowshoe population. This idea is explored further in Example 5.17.

Figure 1.6 may be reproduced as follows.

```
culer = c(rgb(.85,.30,.12,.6), rgb(.12,.67,.86,.6))
tsplot(Hare, col = culer[1], lwd=2, type="o", pch=0,
           ylab=expression(Number~~~("%*% 1000")))
lines(Lynx, col=culer[2], lwd=2, type="o", pch=2)
legend("topright", col=culer, lty=1, lwd=2, pch=c(0,2),
           legend=c("Hare", "Lynx"), bty="n")
```

◊

Example 1.6. fMRI Imaging

Often, time series are observed under varying experimental conditions or treatment configurations. Such a set of series is shown in Figure 1.7, where data are collected from various locations in the brain via functional magnetic resonance imaging (fMRI).

In fMRI, subjects are put into an MRI scanner and a stimulus is applied for a period of time, and then stopped. This on-off application of a stimulus is repeated and recorded by measuring the blood oxygenation-level dependent (BOLD) signal intensity, which measures areas of activation in the brain. The BOLD contrast results from changing regional blood concentrations of oxy- and deoxy- hemoglobin.

The data displayed in Figure 1.7 are from an experiment that used fMRI to examine the effects of general anesthesia on pain perception by comparing results from anesthetized volunteers while a supramaximal shock stimulus was applied. This stimulus was used to simulate surgical incision without inflicting tissue damage. In

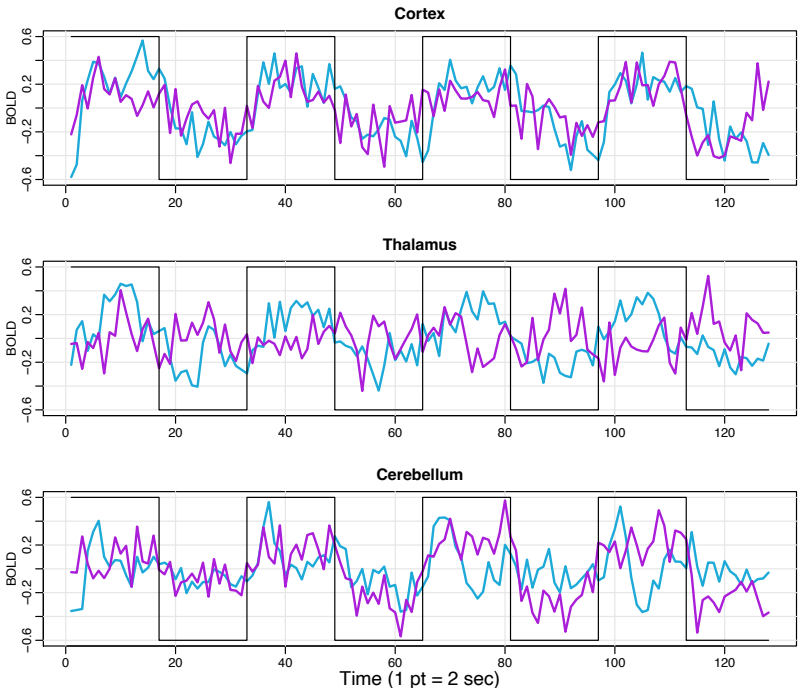


Figure 1.7 *fMRI data from two locations in the cortex, the thalamus, and the cerebellum;* $n = 128$ *points, one observation taken every 2 seconds. The boxed line represents the presence or absence of the stimulus.*

this example, the stimulus was applied for 32 seconds and then stopped for 32 seconds, so that the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$).

Notice that the periodicities appear strongly in the motor cortex series but seem to be missing in the thalamus and perhaps in the cerebellum. In this case, it is of interest to statistically determine if the areas in the thalamus and cerebellum are actually responding to the stimulus. Use the following R commands for the graphic:

```
par(mfrow=c(3,1))
culer = c(rgb(.12,.67,.85,.7), rgb(.67,.12,.85,.7))
u = rep(c(rep(.6,16), rep(-.6,16)), 4)    # stimulus signal
tsplot(fmri1[,4], ylab="BOLD", xlab="", main="Cortex", col=culer[1],
           ylim=c(-.6,.6), lwd=2)
lines(fmri1[,5], col=culer[2], lwd=2)
lines(u, type="s")
tsplot(fmri1[,6], ylab="BOLD", xlab="", main="Thalamus", col=culer[1],
           ylim=c(-.6,.6), lwd=2)
lines(fmri1[,7], col=culer[2], lwd=2)
lines(u, type="s")
```

```
tsplot(fmri1[,8], ylab="BOLD", xlab="", main="Cerebellum",
           col=culer[1], ylim=c(-.6,.6), lwd=2)
lines(fmri1[,9], col=culer[2], lwd=2)
lines(u, type="s")
mtext("Time (1 pt = 2 sec)", side=1, line=1.75)
```

◊

1.3 Time Series Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data, like that encountered in the previous section.

The fundamental visual characteristic distinguishing the different series shown in Example 1.1 – Example 1.6 is their differing degrees of smoothness. A parsimonious explanation for this smoothness is that adjacent points in time are correlated, so the value of the series at time t, say, x_t, depends in some way on the past values $x_{t-1}, x_{t-2}, \ldots$. This idea expresses a fundamental way in which we might think about generating realistic looking time series.

Example 1.7. White Noise
A simple kind of generated series might be a collection of *uncorrelated* random variables, w_t, with mean 0 and finite variance σ_w^2. The time series generated from uncorrelated variables is used as a model for noise in engineering applications where it is called *white noise*; we shall sometimes denote this process as $w_t \sim wn(0, \sigma_w^2)$. The designation white originates from the analogy with white light (details in Chapter 6). A special version of white noise that we use is when the variables are independent and identically distributed normals, written $w_t \sim \text{iid } \mathrm{N}(0, \sigma_w^2)$.

The upper panel of Figure 1.8 shows a collection of 500 independent standard normal random variables ($\sigma_w^2 = 1$), plotted in the order in which they were drawn. The resulting series bears a resemblance to portions of the DJIA returns in Figure 1.3. ◊

If the stochastic behavior of all time series could be explained in terms of the white noise model, classical statistical methods would suffice. Two ways of introducing serial correlation and more smoothness into time series models are given in Example 1.8 and Example 1.9.

Example 1.8. Moving Averages, Smoothing and Filtering
We might replace the white noise series w_t by a moving average that smoothes the series. For example, consider replacing w_t in Example 1.7 by an average of its current value and its immediate two neighbors in the past. That is, let

$$v_t = \tfrac{1}{3}\left(w_{t-1} + w_t + w_{t+1}\right), \tag{1.1}$$

new time series value at time t

which leads to the series shown in the lower panel of Figure 1.8. This series is much smoother than the white noise series and has a smaller variance due to averaging. It should also apparent that averaging removes some of the high frequency (fast

which smoothenes the graph, by doing linear combination of curent, pre,next -->sliding sum. As we assumed white noise is uncorrelated, variance of sum becomes sum of variances and we see that varaicne becomes by 1/3. -- see notes.

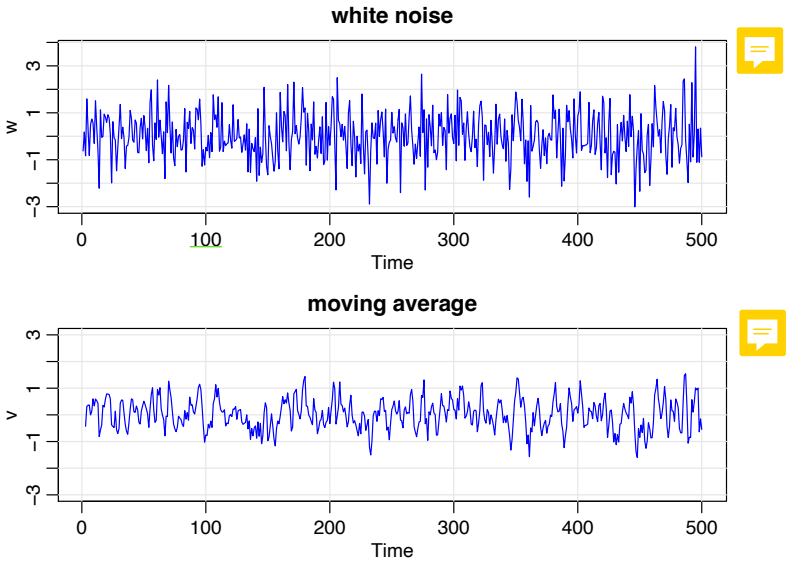


Figure 1.8 *Gaussian white noise series (top) and three-point moving average of the Gaussian white noise series (bottom).*

oscillations) behavior of the noise. We begin to notice a similarity to some of the non-cyclic fMRI series in Figure 1.7.

A linear combination of values in a time series such as in (1.1) is referred to, generically, as a filtered series; hence the command `filter`. To reproduce Figure 1.8:

```
par(mfrow=2:1)
w = rnorm(500)                          # 500 N(0,1) variates
v = filter(w, sides=2, filter=rep(1/3,3))  # moving average
tsplot(w, col=4, main="white noise")
tsplot(v, ylim=c(-3,3), col=4, main="white noise")
```

◊

The SOI and Recruitment series in Figure 1.5, as well as some of the fMRI series in Figure 1.7, differ from the moving average series because they are dominated by an oscillatory behavior. A number of methods exist for generating series with this quasi-periodic behavior; we illustrate a popular one based on the autoregressive model considered in Chapter 4.

during prediction, roughly remove the error - wt, like Linear regression.

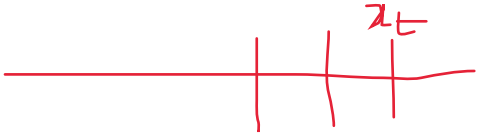
Example 1.9. Autoregressions

Suppose we consider the white noise series w_t of Example 1.7 as input and calculate the output using the second-order equation

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t \tag{1.2}$$

how many x should we bring in ? last 2 days? 3 days? 1 year?

successively for $t = 1, 2, \ldots, 250$. The resulting output series is shown in Figure 1.9. Equation (1.2) represents a regression or prediction of the current value x_t of a



coefficients ?

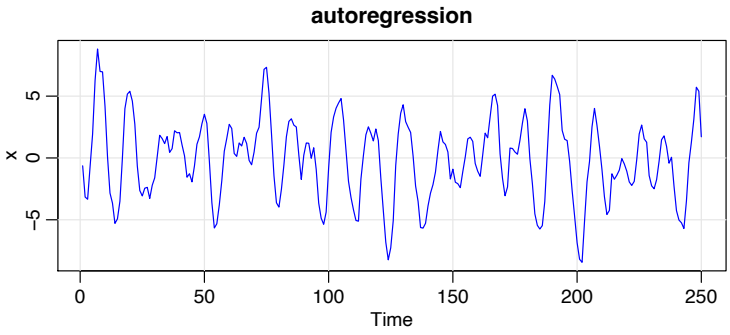


Figure 1.9: *Autoregressive series generated from model* (1.2).

time series as a function of the past two values of the series, and, hence, the term autoregression is suggested for this model. A problem with startup values exists here because (1.2) also depends on the initial conditions x_0 and x_{-1}, but for now we set them to zero. We can then generate data *recursively* by substituting into (1.2). That is, given $w_1, w_2, \ldots, w_{250}$, we could set $x_{-1} = x_0 = 0$ and then start at $t = 1$:

$$\begin{aligned} x_1 &= 1.5x_0 - .75x_{-1} + w_1 = w_1 \\ x_2 &= 1.5x_1 - .75x_0 + w_2 \quad = 1.5w_1 + w_2 \\ x_3 &= 1.5x_2 - .75x_1 + w_3 \\ x_4 &= 1.5x_3 - .75x_2 + w_4 \end{aligned}$$

and so on. We note the approximate periodic behavior of the series, which is similar to that displayed by the SOI and Recruitment in Figure 1.5 and some fMRI series in Figure 1.7. This particular model is chosen so that the data have pseudo-cyclic behavior of about 1 cycle every 12 points; thus 250 observations should contain about 20 cycles. This autoregressive model and its generalizations can be used as an underlying model for many observed series and will be studied in detail in Chapter 4.

One way to simulate and plot data from the model (1.2) in R is to use the following commands. The initial conditions are set equal to zero so we let the filter run an extra 50 values to avoid startup problems.

```
set.seed(90210)
w = rnorm(250 + 50)  # 50 extra to avoid startup problems
x = filter(w, filter=c(1.5,-.75), method="recursive")[-(1:50)]
tsplot(x, main="autoregression", col=4)
```

◊

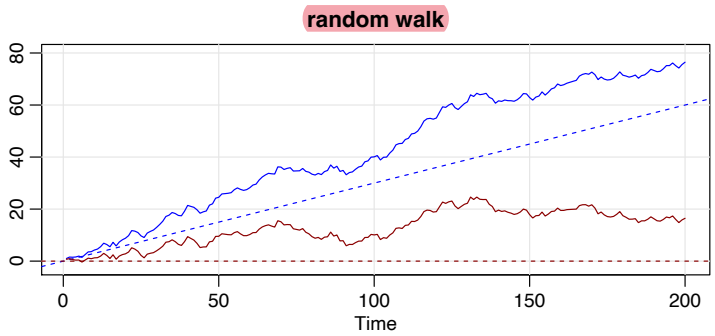
Example 1.10. Random Walk with Drift

A model for analyzing trend such as seen in the global temperature data in Figure 1.2, is the random walk with drift model given by

$$x_t = \delta + x_{t-1} + w_t \tag{1.3}$$

Delta=0 it becomes a auto regressive model.

variance increases for randomwalk , see notes.



delta is negative, trend downward. and other.

Figure 1.10 *Random walk, $\sigma_w = 1$, with drift $\delta = .3$ (upper jagged line), without drift, $\delta = 0$ (lower jagged line), and dashed lines showing the drifts.*

for $t = 1, 2, \ldots$, with initial condition $x_0 = 0$, and where w_t is white noise. The constant δ is called the drift, and when $\delta = 0$, the model is called simply a random walk because the value of the time series at time t is the value of the series at time $t-1$ plus a completely random movement determined by w_t. Note that we may rewrite (1.3) as a cumulative sum of white noise variates. That is,

random walk is cummulative summation of white noise (linear combination of white noise). + drift

$$x_t = \delta\, t + \sum_{j=1}^{t} w_j \tag{1.4}$$

for $t = 1, 2, \ldots$; either use induction, or plug (1.4) into (1.3) to verify this statement. Figure 1.10 shows 200 observations generated from the model with $\delta = 0$ and .3, and with standard normal nose. For comparison, we also superimposed the straight lines δt on the graph. To reproduce Figure 1.10 in R use the following code (notice the use of multiple commands per line using a semicolon).

```
set.seed(314159265)       # so you can reproduce the results
w  = rnorm(200);  x  = cumsum(w)   # random walk
wd = w +.3;       xd = cumsum(wd)  # random walk with drift
tsplot(xd, ylim=c(-2,80), main="random walk", ylab="", col=4)
abline(a=0, b=.3, lty=2, col=4)    # plot drift
lines(x, col="darkred")
abline(h=0, col="darkred", lty=2)
```

◊

Example 1.11. Signal Plus Noise

Many realistic models for generating time series assume an underlying signal with some consistent periodic variation contaminated by noise. For example, it is easy to detect the regular cycle fMRI series displayed on the top of Figure 1.7. Consider the model

$$x_t = 2\cos(2\pi \tfrac{t+15}{50}) + w_t \tag{1.5}$$

for $t = 1, 2, \ldots, 500$, where the first term is regarded as the signal, shown in the

non linear regression.

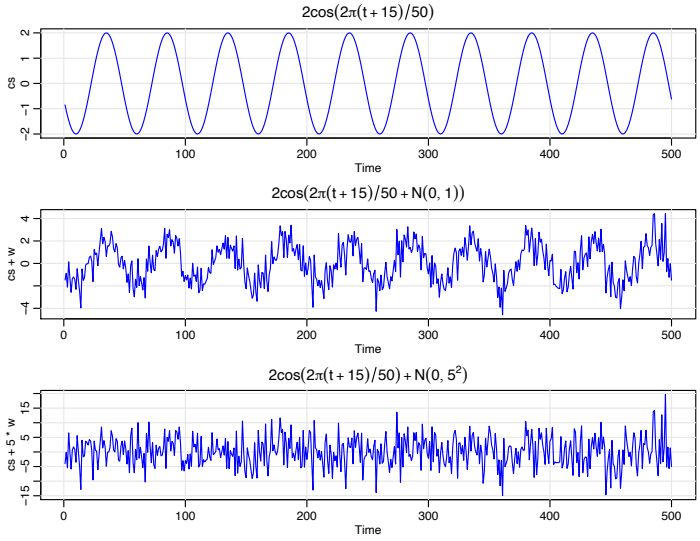


Figure 1.11 *Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise, $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel); see* (1.5).

upper panel of Figure 1.11. We note that a sinusoidal waveform can be written as

$$A\cos(2\pi\omega t + \phi), \tag{1.6}$$

where A is the amplitude, ω is the frequency of oscillation, and ϕ is a phase shift. In (1.5), $A = 2$, $\omega = 1/50$ (one cycle every 50 time points), and $\phi = .6\pi$.

An additive noise term was taken to be white noise with $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel), drawn from a normal distribution. Adding the two together obscures the signal, as shown in the lower panels of Figure 1.11. The degree to which the signal is obscured depends on the amplitude of the signal relative to the size of σ_w. The ratio of the amplitude of the signal to σ_w (or some function of the ratio) is sometimes called the *signal-to-noise ratio (SNR)*; the larger the SNR, the easier it is to detect the signal. Note that the signal is easily discernible in the middle panel, whereas the signal is obscured in the bottom panel. Typically, we will not observe the signal but the signal obscured by noise.

To reproduce Figure 1.11 in R, use the following commands:

```
t  = 1:500
cs = 2*cos(2*pi*(t+15)/50)    # signal
w  = rnorm(500)               # noise
par(mfrow=c(3,1))
tsplot(cs,    col=4, main=expression(2*cos(2*pi*(t+15)/50)))
tsplot(cs+w,  col=4, main=expression(2*cos(2*pi*(t+15)/50+N(0,1))))
tsplot(cs+5*w,col=4, main=expression(2*cos(2*pi*(t+15)/50)+N(0,5^2))) ◊
```

STAT 626: Outline of Lecture 4
Correlation and Dependence (§2.1)

1. Mean Function of a Time Series (Stochastic Process),

2. Covariance Function of a Time Series,

3. Stationary Time Series

4. NOTES: Review of Mean, Variance, Covariance of Lin Comb. of RVs.

Topics From Chapter 1

5. White Noise: The Building Blocks

6. Autoregression: The Birth of Modern Time Series Analysis

7. Random Walks: The Engine of Financial Engineering

8. Signal + Noise: For Other Engineering

DEFINITIONS

1. A Time Series $\{x_t\}$ is **stationary** if

   (a) the mean function $E(x_t)$ does not depend on the time t,

   (b) the covariance function $\mathrm{cov}(x_s, x_t)$ depends on the times s, t only through the distance $|s - t|$.

2. **Autocovariance Function** of a Stationary Time Series:

$$\gamma(h) = \mathrm{cov}(x_{t+h}, x_t), \quad h = 0, 1, \ldots.$$

   NOTE:

$$\gamma(0) = \mathrm{cov}(x_t, x_t) = \mathrm{var}(x_t).$$

3. **The Autocorrelation Function (ACF)**

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}, \quad h = 0, 1, \ldots,$$

   is a **symmetric** function of the lag h. **Correlogram** is the plot of $\rho(h)$ vs h.

4. **Multivariate Time Series:**
   Cross-Covariance Function (CCF) of Two Time Series:

$$\gamma_{xy}(h) = \mathrm{Cov}(x_{t+h}, y_t), h = 0, \pm 1, \pm 2, \ldots$$

   Why ACF is symmetric and CCF is not?
   Proof without words!

$$\gamma_{xx}(h) = \mathrm{Cov}(x_{t+h}, x_t) = \mathrm{Cov}(x_t, x_{t+h}) = \gamma_{xx}(-h), \quad h = 1, 2, \ldots.$$

Chapter 2

Correlation and Stationary Time Series

2.1 Measuring Dependence

We now discuss various measures that describe the general behavior of a process as it evolves over time. The material on probability in Appendix B may be of help with some of the content in this chapter. A rather simple descriptive measure is the mean function, such as the average monthly high temperature for your city. In this case, the mean is a *function of time*.

Definition 2.1. *The* **mean function** *is defined as*

$$\mu_{xt} = E(x_t) \tag{2.1}$$

provided it exists, where E denotes the usual expected value operator. When no confusion exists about which time series we are referring to, we will drop a subscript and write μ_{xt} *as* μ_t.

Example 2.2. Mean Function of a Moving Average Series
If w_t denotes a white noise series, then $\mu_{wt} = E(w_t) = 0$ for all t. The top series in Figure 1.8 reflects this, as the series clearly fluctuates around a mean value of zero. Smoothing the series as in Example 1.8 does not change the mean because we can write

$$\mu_{vt} = E(v_t) = \tfrac{1}{3}[E(w_{t-1}) + E(w_t) + E(w_{t+1})] = 0. \qquad \lozenge$$

Example 2.3. Mean Function of a Random Walk with Drift
Consider the random walk with drift model given in (1.4),

$$x_t = \delta\, t + \sum_{j=1}^{t} w_j, \qquad t = 1, 2, \ldots .$$

Because $E(w_t) = 0$ for all t, and δ is a constant, we have

$$\mu_{xt} = E(x_t) = \delta\, t + \sum_{j=1}^{t} E(w_j) = \delta\, t$$

which is a straight line with slope δ. A realization of a random walk with drift can be compared to its mean function in Figure 1.10. $\lozenge$

Example 2.4. Mean Function of Signal Plus Noise
A great many practical applications depend on assuming the observed data have been generated by a fixed signal waveform superimposed on a zero-mean noise process, leading to an additive signal model of the form (1.5). It is clear, because the signal in (1.5) is a fixed function of time, we will have

$$\begin{aligned}\mu_{xt} &= E\left[2\cos(2\pi\tfrac{t+15}{50}) + w_t\right] \\ &= 2\cos(2\pi\tfrac{t+15}{50}) + E(w_t) \\ &= 2\cos(2\pi\tfrac{t+15}{50}),\end{aligned}$$

and the mean function is just the cosine wave. ◊

The mean function describes only the marginal behavior of a time series. The lack of independence between two adjacent values x_s and x_t can be assessed numerically, as in classical statistics, using the notions of covariance and correlation. Assuming the variance of x_t is finite, we have the following definition.

Definition 2.5. *The* **autocovariance function** *is defined as the second moment product*

$$\gamma_x(s,t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \tag{2.2}$$

for all s and t. When no possible confusion exists about which time series we are referring to, we will drop the subscript and write $\gamma_x(s,t)$ as $\gamma(s,t)$.

Note that $\gamma_x(s,t) = \gamma_x(t,s)$ for all time points s and t. The autocovariance measures the *linear* dependence between two points on the same series observed at different times. Recall from classical statistics that if $\gamma_x(s,t) = 0$, then x_s and x_t are not linearly related, but there still may be some dependence structure between them. If, however, x_s and x_t are bivariate normal, $\gamma_x(s,t) = 0$ ensures their independence. It is clear that, for $s = t$, the autocovariance reduces to the (assumed finite) *variance*, because

$$\gamma_x(t,t) = E[(x_t - \mu_t)^2] = \text{var}(x_t). \tag{2.3}$$

Example 2.6. Autocovariance of White Noise
The white noise series w_t has $E(w_t) = 0$ and

$$\gamma_w(s,t) = \text{cov}(w_s, w_t) = \begin{cases} \sigma_w^2 & s = t, \\ 0 & s \neq t. \end{cases} \tag{2.4}$$

A realization of white noise is shown in the top panel of Figure 1.8. ◊

We often have to calculate the autocovariance between filtered series. A useful result is given in the following proposition.

Property 2.7. *If the random variables*

$$U = \sum_{j=1}^{m} a_j X_j \quad \text{and} \quad V = \sum_{k=1}^{r} b_k Y_k$$

are linear filters of (finite variance) random variables $\{X_j\}$ *and* $\{Y_k\}$*, respectively, then*

$$\text{cov}(U, V) = \sum_{j=1}^{m} \sum_{k=1}^{r} a_j b_k \text{cov}(X_j, Y_k). \tag{2.5}$$

Furthermore, $\text{var}(U) = \text{cov}(U, U)$.

An easy way to remember (2.5) is to treat it like multiplication:

$$(a_1 X_1 + a_2 X_2)(b_1 Y_1) = a_1 b_1 X_1 Y_1 + a_2 b_1 X_2 Y_1$$

Example 2.8. Autocovariance of a Moving Average
Consider applying a three-point moving average to the white noise series w_t of the previous example as in Example 1.8. In this case,

$$\gamma_v(s,t) = \text{cov}(v_s, v_t) = \text{cov}\left\{\tfrac{1}{3}\left(w_{s-1} + w_s + w_{s+1}\right), \tfrac{1}{3}\left(w_{t-1} + w_t + w_{t+1}\right)\right\}.$$

When $s = t$ we have

$$\begin{aligned}\gamma_v(t,t) &= \tfrac{1}{9}\text{cov}\{(w_{t-1} + w_t + w_{t+1}), (w_{t-1} + w_t + w_{t+1})\} \\ &= \tfrac{1}{9}[\text{cov}(w_{t-1}, w_{t-1}) + \text{cov}(w_t, w_t) + \text{cov}(w_{t+1}, w_{t+1})] \\ &= \tfrac{3}{9}\sigma_w^2.\end{aligned}$$

When $s = t + 1$,

$$\begin{aligned}\gamma_v(t+1,t) &= \tfrac{1}{9}\text{cov}\{(w_t + w_{t+1} + w_{t+2}), (w_{t-1} + w_t + w_{t+1})\} \\ &= \tfrac{1}{9}[\text{cov}(w_t, w_t) + \text{cov}(w_{t+1}, w_{t+1})] \\ &= \tfrac{2}{9}\sigma_w^2,\end{aligned}$$

using (2.4). Similar computations give $\gamma_v(t-1,t) = 2\sigma_w^2/9$, $\gamma_v(t+2,t) = \gamma_v(t-2,t) = \sigma_w^2/9$, and 0 when $|t-s| > 2$. We summarize the values for all s and t as

$$\gamma_v(s,t) = \begin{cases} \frac{3}{9}\sigma_w^2 & s = t, \\ \frac{2}{9}\sigma_w^2 & |s-t| = 1, \\ \frac{1}{9}\sigma_w^2 & |s-t| = 2, \\ 0 & |s-t| > 2. \end{cases} \tag{2.6}$$

$\Diamond$

Example 2.9. Autocovariance of a Random Walk
For the random walk model, $x_t = \sum_{j=1}^{t} w_j$, we have

$$\gamma_x(s,t) = \text{cov}(x_s, x_t) = \text{cov}\left(\sum_{j=1}^{s} w_j, \sum_{k=1}^{t} w_k\right) = \min\{s,t\}\,\sigma_w^2\,,$$

because the w_t are uncorrelated random variables. For example, with $s = 2$ and $t = 4$,

$$\text{cov}(x_2, x_4) = \text{cov}(w_1 + w_2, w_1 + w_2 + w_3 + w_4) = 2\sigma_w^2\,.$$

Note that, as opposed to the previous examples, the autocovariance function of a random walk depends on the particular time values s and t, and not on the time separation or lag. Also, notice that the variance of the random walk, $\text{var}(x_t) = \gamma_x(t,t) = t\,\sigma_w^2$, increases without bound as time t increases. The effect of this variance increase can be seen in Figure 1.10 where the processes start to move away from their mean functions $\delta\,t$ (note that $\delta = 0$ and .3 in that example). ♢

As in classical statistics, it is more convenient to deal with a measure of association between -1 and 1, and this leads to the following definition.

Definition 2.10. *The* **autocorrelation function (ACF)** *is defined as*

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}. \tag{2.7}$$

The ACF measures the linear predictability of the series at time t, say x_t, using only the value x_s. And because it is a correlation, we must have $-1 \le \rho(s,t) \le 1$. If we can predict x_t *perfectly* from x_s through a linear relationship, $x_t = \beta_0 + \beta_1 x_s$, then the correlation will be $+1$ when $\beta_1 > 0$, and -1 when $\beta_1 < 0$. Hence, we have a rough measure of the ability to forecast the series at time t from the value at time s.

Often, we would like to measure the predictability of another series y_t from the series x_s. Assuming both series have finite variances, we have the following definition.

Definition 2.11. *The* **cross-covariance function** *between two series, x_t and y_t, is*

$$\gamma_{xy}(s,t) = \text{cov}(x_s, y_t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})]. \tag{2.8}$$

We can use the cross-covariance function to develop a correlation:

Definition 2.12. *The* **cross-correlation function (CCF)** *is given by*

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}}. \tag{2.9}$$

2.2 Stationarity

Although we have previously not made any special assumptions about the behavior of the time series, many of the examples we have seen hinted that a sort of regularity may exist over time in the behavior of a time series. Stationarity requires regularity in the mean and autocorrelation functions so that these quantities (at least) may be estimated by averaging.

Definition 2.13. *A* **stationary** *time series is a finite variance process where*

(i) the mean value function, μ_t, defined in (2.1) is constant and does not depend on time t, and

(ii) the autocovariance function, $\gamma(s,t)$, defined in (2.2) depends on times s and t only through their time difference.

As an example, for a stationary hourly time series, the correlation between what happens at 1AM and 3AM is the same as between what happens at 9PM and 11PM because they are both two hours apart.

Example 2.14. A Random Walk is Not Stationary
A random walk is not stationary because its autocovariance function, $\gamma(s,t) = \min\{s,t\}\sigma_w^2$, depends on time; see Example 2.9 and Problem 2.5. Also, the random walk with drift violates both conditions of Definition 2.13 because the mean function, $\mu_{xt} = \delta t$, depends on time t as shown in Example 2.3

Because the mean function, $E(x_t) = \mu_t$, of a stationary time series is independent of time t, we will write

$$\mu_t = \mu. \tag{2.10}$$

Also, because the autocovariance function, $\gamma(s,t)$, of a stationary time series, x_t, depends on s and t only through time difference, we may simplify the notation. Let $s = t + h$, where h represents the time shift or *lag*. Then

$$\gamma(t+h,t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_h, x_0) = \gamma(h,0)$$

because the time difference between $t+h$ and t is the same as the time difference between h and 0. Thus, the autocovariance function of a stationary time series does not depend on the time argument t. Henceforth, for convenience, we will drop the second argument of $\gamma(h,0)$.

Definition 2.15. *The* **autocovariance function of a stationary time series** *will be written as*

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu)(x_t - \mu)]. \tag{2.11}$$

Definition 2.16. *The* **autocorrelation function (ACF) of a stationary time series** *will be written using (2.7) as*

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{2.12}$$

STAT 626: Review of Mean and Variance of Lin. Combination of Random Variables

Mean and Variance of Sum of Random Variables

1. If X is a random variable with mean $E(X) = \mu$ and variance

$$\sigma^2 = E(X-\mu)^2 = E(X^2) - \mu^2.$$

Then for c a constant, we have

(a) $E(cX) = cE(X) = c\mu$,

(b) $\text{Var}(cX) = c^2\sigma^2$.

2. If X, Y are two random variables with means μ_1, μ_2, variances σ_1^2, σ_2^2 and covariance

$$\sigma_{12} = \text{Cov}(X,Y) = E(X-\mu_X)(Y-\mu_Y),$$

then

(a). $E(X+Y) = \mu_1 + \mu_2$.

(b). $\text{Var}(X+Y) = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12}$.

Exercise 1 : If X, Y are random variables as above and c_1, c_2 are constants, write the formulas for

$$E(c_1X + c_2Y), \quad \text{Var}(c_1X + c_2Y),$$

and $Q(b) = \text{Var}(Y - bX)^2$, where b is a scalar.

Exercise 2 : If X, Y are random variables with

$$\mu_1 = 2, \quad \mu_2 = -5, \quad \sigma_1^2 = 4, \quad \sigma_2^2 = 17, \quad \sigma_{12} = -3,$$

and $c_1 = -2, \quad c_2 = 1$, find the numerical values of

$$E(c_1X + c_2Y), \quad \text{Var}(c_1X + c_2Y), \quad Q(b).$$

Compute the correlation coefficient ρ between X and Y.

Exercise 3: If $X_1, \ldots, X_n$ are independent random variables with $E(X_i) = \mu_i, \text{Var}(X_i) = \sigma_i^2$, and c_1, c_2 are constant scalars, find

$$E(c_1X_1 + c_2X_2), \quad \text{Var}(c_1X_1 + c_2X_2),$$

and $\text{Var}(\bar{X})$, where $\bar{X}$ is the sample mean.

Exercise 4: Show that for any scalars a_1, a_2, a_3 and random variables X_1, X_2, X_3: $\text{Var}(a_1X_1 + a_2X_2 + a_3X_3) =$

$$a_1^2\text{Var}(X_1)+a_2^2\text{Var}(X_2)+a_3^2\text{Var}(X_3)+2a_1a_2\text{Cov}(X_1, X_2)+2a_1a_3\text{Cov}(X_1, X_3)+2a_2a_3\text{Cov}(X_2, X_3).$$

(a) If Cov (X_1, X_2)= Cov (X_2, X_3)=ρ, Cov $(X_1, X_3) = \rho^2$ and $\text{Var}(X_i) = 1, i = 1, 2, 3$, then write the 3×3 covariance matrix of the random vector $X = (X_1, X_2, X_3)$.

(b) Compute $\text{Var}(X_1 + X_2 + X_3)$ when $\rho = 0.6$.

(c) Mark T is interested in forecasting X_3 using the linear predictor $\widehat{X}_3 = b_2X_2 + b_1X_1$. He realizes the forecast error is $X_3 - \widehat{X}_3 = X_3 - b_2X_2 - b_1X_1$ and a great way to find the predictor coefficients b_1, b_2, is by *minimizing the variance of forecast error*

$$Q(b_1, b_2) = \text{Var}(X_3 - b_2X_2 - b_1X_1),$$

which turns out to be a quadratic function of b_1, b_2 (as in least-squares estimation in regression). Help Mark to minimize this function or derive the normal equations.

(d) Solve the normal equations and observe that $b_1 = 0$, regardless of the value of ρ.

Exercise 5: Now consider the random variables X_1, X_2, X_3.

(a) If Cov (X_1, X_2)= Cov (X_2, X_3)= Cov $(X_1, X_3) = \rho$ and $\text{Var}(X_i) = 1, i = 1, 2, 3$, write the 3×3 covariance matrix of the random vector $X = (X_1, X_2, X_3)$.

(b) Express $\text{Var}(X_1 + X_2 + X_3)$ in terms of ρ.

(c) In forecasting X_3 using the linear predictor $\widehat{X}_3 = b_2X_2 + b_1X_1$, the forecast error is $X_3 - \widehat{X}_3 = X_3 - b_2X_2 - b_1X_1$, find the predictor coefficients b_1, b_2, by *minimizing the variance of forecast error*

$$Q(b_1, b_2) = \text{Var}(X_3 - b_2X_2 - b_1X_1),$$

which turns out to be a quadratic function of b_1, b_2. Minimize this function or derive the normal equations.

(d) Solve the normal equations and express b_1 and b_2 in terms of ρ. Compare these predictor coefficients with those in the previous exercise.

Exercise 6: Suppose all pairwise covariances between the (past) random variables $X_1, X_2, \ldots, X_p$ and a (future) random variable X_{p+1} are known and given by

$$\text{Cov}(X_i, X_j) = \rho^{|i-j|}.$$

(a) Organize the above pairwise covariance information in a $(p+1) \times (p+1)$ matrix. For $p = 4, \rho = 0.5$, write the form of this matrix.

(b) Explain in words why you might be interested in choosing the b_i's so that

$$Q(b_1, b_2, \ldots, b_p) = \text{Var}(X_{p+1} - b_1X_1 - b_2X_2 - \ldots - b_pX_p),$$

is *minimized*. (Hint: Think in terms of prediction or forecasting).

(c) Derive the *normal equations* for minimizing $Q(b_1, b_2, \ldots, b_p)$.

(d) Write the equations in (c) in matrix form. What does it take to solve it for the b_i's.

Exercise 7: Suppose all pairwise covariances between the (past) random variables $X_1, X_2, \ldots, X_p$ and a (future) random variable Y are known and given by Cov $(X_i, X_j) = \rho, \quad i \neq j$, and Cov $(X_i, Y) = 0$, for all i.

(a) Organize the above pairwise covariance information in a $(p+1) \times (p+1)$ matrix. For $p = 4, \rho = 0.5$, write the form of this matrix.

(b) One is interested in choosing the b_i's so that

$$Q(b_1, b_2, \ldots, b_p) = \mathrm{Var}(Y - b_1X_1 - b_2X_2 - \ldots - b_pX_p),$$

is *minimized*. Derive the *normal equations* for minimizing $Q(b_1, b_2, \ldots, b_p)$.

(c) Write the equations in (b) in matrix form and solve it to find the b_i's.

STAT 626: Outline of Lectures 5-6
Stationary Time Series (§2.2)

1. Autocovariance of a Time Series (TS),

2. Autocorrelation Function (ACF) of a Stationary TS and Correlogram:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}, \quad h = 0, 1, \ldots.$$

3. Important Example of Stationary Time Series: Moving Average (MA), Autoregressive (AR) Models,...

4. Wold Decomposition: A good stationary TS is linear in the WN.

5. Bivariate TS and Stationarity.

Topics From Chapter 1

6. White Noise: The Building Blocks

7. Autoregression: The Birth of Modern Time Series Analysis

8. Random Walks: The Engine of Financial Engineering

9. Signal + Noise: For Other Engineering

DEFINITIONS

1. A Time Series $\{x_t\}$ is **stationary** if

   (a) the mean function $E(x_t)$ does not depend on the time t,

   (b) the covariance function $\mathrm{cov}(x_s, x_t)$ depends on the times s, t only through the distance $|s - t|$.

2. **Autocovariance Function** of a Stationary Time Series:

$$\gamma(h) = \mathrm{cov}(x_{t+h}, x_t), \quad h = 0, 1, \ldots.$$

   NOTE:

$$\gamma(0) = \mathrm{cov}(x_t, x_t) = \mathrm{var}(x_t).$$

3. **The Autocorrelation Function (ACF)**

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}, \quad h = 0, 1, \ldots,$$

   is a **symmetric** function of the lag h. **Correlogram** is the plot of $\rho(h)$ vs h.

4. **Multivariate Time Series:**
   Cross-Covariance Function (CCF) of Two Time Series:

$$\gamma_{xy}(h) = \mathrm{Cov}(x_{t+h}, y_t), h = 0, \pm 1, \pm 2, \ldots$$

   Why ACF is symmetric and CCF is not?
   Proof without words!

$$\gamma_{xx}(h) = \mathrm{Cov}(x_{t+h}, x_t) = \mathrm{Cov}(x_t, x_{t+h}) = \gamma_{xx}(-h), \quad h = 1, 2, \ldots.$$

2.2 Stationarity

Although we have previously not made any special assumptions about the behavior of the time series, many of the examples we have seen hinted that a sort of regularity may exist over time in the behavior of a time series. Stationarity requires regularity in the mean and autocorrelation functions so that these quantities (at least) may be estimated by averaging.

Definition 2.13. *A* **stationary** *time series is a finite variance process where*

(i) the mean value function, μ_t, defined in (2.1) is constant and does not depend on time t, and

(ii) the autocovariance function, $\gamma(s,t)$, defined in (2.2) depends on times s and t only through their time difference.

As an example, for a stationary hourly time series, the correlation between what happens at 1AM and 3AM is the same as between what happens at 9PM and 11PM because they are both two hours apart.

Example 2.14. A Random Walk is Not Stationary
A random walk is not stationary because its autocovariance function, $\gamma(s,t) = \min\{s,t\}\sigma_w^2$, depends on time; see Example 2.9 and Problem 2.5. Also, the random walk with drift violates both conditions of Definition 2.13 because the mean function, $\mu_{xt} = \delta t$, depends on time t as shown in Example 2.3

Because the mean function, $E(x_t) = \mu_t$, of a stationary time series is independent of time t, we will write

$$\mu_t = \mu. \tag{2.10}$$

Also, because the autocovariance function, $\gamma(s,t)$, of a stationary time series, x_t, depends on s and t only through time difference, we may simplify the notation. Let $s = t+h$, where h represents the time shift or *lag*. Then

$$\gamma(t+h,t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_h, x_0) = \gamma(h,0)$$

because the time difference between $t+h$ and t is the same as the time difference between h and 0. Thus, the autocovariance function of a stationary time series does not depend on the time argument t. Henceforth, for convenience, we will drop the second argument of $\gamma(h,0)$.

Definition 2.15. *The* **autocovariance function of a stationary time series** *will be written as*

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = E[(x_{t+h}-\mu)(x_t-\mu)]. \tag{2.11}$$

Definition 2.16. *The* **autocorrelation function (ACF) of a stationary time series** *will be written using (2.7) as*

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{2.12}$$

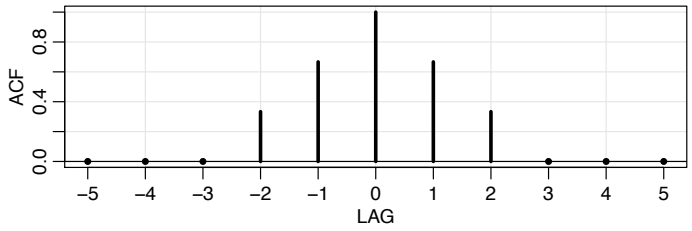


Figure 2.1: *Autocorrelation function of a three-point moving average.*

Because it is a correlation, we have $-1 \leq \rho(h) \leq 1$ for all h, enabling one to assess the relative importance of a given autocorrelation value by comparing with the extreme values -1 and 1.

Example 2.17. Stationarity of White Noise

The mean and autocovariance functions of the white noise series discussed in Example 1.7 and Example 2.6 are easily evaluated as $\mu_{wt} = 0$ and

$$\gamma_w(h) = \text{cov}(w_{t+h}, w_t) = \begin{cases} \sigma_w^2 & h = 0, \\ 0 & h \neq 0. \end{cases}$$

Thus, white noise satistfies Definition 2.13 and is stationary. ◊

Example 2.18. Stationarity of a Moving Average

The three-point moving average process of Example 1.8 is stationary because, from Example 2.2 and Example 2.8, the mean and autocovariance functions $\mu_{vt} = 0$, and

$$\gamma_v(h) = \begin{cases} \frac{3}{9}\sigma_w^2 & h = 0, \\ \frac{2}{9}\sigma_w^2 & h = \pm 1, \\ \frac{1}{9}\sigma_w^2 & h = \pm 2, \\ 0 & |h| > 2 \end{cases}$$

are independent of time t, satisfying the conditions of Definition 2.13. Note that the ACF, $\rho(h) = \gamma(h)/\gamma(0)$, is given by

$$\rho_v(h) = \begin{cases} 1 & h = 0, \\ 2/3 & h = \pm 1, \\ 1/3 & h = \pm 2, \\ 0 & |h| > 2 \end{cases}.$$

Figure 2.1 shows a plot of the autocorrelation as a function of lag h. Note that the autocorrelation function is symmetric about lag zero.

```
ACF = c(0,0,0,1,2,3,2,1,0,0,0)/3
LAG = -5:5
tsplot(LAG, ACF, type="h", lwd=3, xlab="LAG")
```

```
abline(h=0)
points(LAG[-(4:8)], ACF[-(4:8)], pch=20)
axis(1, at=seq(-5, 5, by=2))
```

Example 2.19. Trend Stationarity
A time series can have stationary behavior around a trend. For example, if

$$x_t = \beta t + y_t\,,$$

where y_t is stationary with mean and autocovariance functions μ_y and $\gamma_y(h)$, respectively. Then the mean function of x_t is

$$\mu_{x,t} = E(x_t) = \beta t + \mu_y\,,$$

which is not independent of time. Therefore, the process is not stationary. The autocovariance function, however, is independent of time, because

$$\begin{aligned}\gamma_x(h) = \mathrm{cov}(x_{t+h}, x_t) &= E[(x_{t+h} - \mu_{x,t+h})(x_t - \mu_{x,t})] \\ &= E[(y_{t+h} - \mu_y)(y_t - \mu_y)] = \gamma_y(h).\end{aligned}$$

This behavior is sometimes called *trend stationarity*. An example of such a process is the export price of salmon series displayed in Figure 3.1. ◊

The autocovariance function of a stationary process has several useful properties. First, the value at $h = 0$ is the variance of the series,

$$\gamma(0) = E[(x_t - \mu)^2] = \mathrm{var}(x_t)\,. \tag{2.13}$$

Another useful property is that the autocovariance function of a stationary series is symmetric around the origin,

$$\gamma(h) = \gamma(-h) \tag{2.14}$$

for all h. This property follows because

$$\begin{aligned}\gamma(h) &= \gamma((t+h) - t) = E[(x_{t+h} - \mu)(x_t - \mu)] \\ &= E[(x_t - \mu)(x_{t+h} - \mu)] = \gamma(t - (t+h)) = \gamma(-h),\end{aligned}$$

which shows how to use the notation as well as proving the result.

Example 2.20. Autoregressive Models
The stationarity of AR models is a little more complex and is dealt with in Chapter 4. We'll use an AR(1) to examine some aspects of the model,

$$x_t = \phi x_{t-1} + w_t.$$

Since the mean must be constant, if x_t is stationary the mean function $\mu_t = E(x_t) = \mu$ is constant so

$$E(x_t) = \phi E(x_{t-1}) + E(w_t)$$

implies $\mu = \phi\mu + 0$; thus $\mu = 0$. In addition, assuming x_{t-1} and w_t are uncorrelated,

$$\begin{aligned}\text{var}(x_t) &= \text{var}(\phi x_{t-1} + w_t)\\ &= \text{var}(\phi x_{t-1}) + \text{var}(w_t) + 2\text{cov}(\phi x_{t-1}, w_t)\\ &= \phi^2\text{var}(x_{t-1}) + \text{var}(w_t)\,.\end{aligned}$$

If x_t is stationary, the variance, $\text{var}(x_t) = \gamma_x(0)$, is constant, so

$$\gamma_x(0) = \phi^2\gamma_x(0) + \sigma_w^2\,.$$

Thus

$$\gamma_x(0) = \sigma_w^2\,\frac{1}{(1-\phi^2)}\,.$$

Note that for the process to have a positive, finite variance, we should require $|\phi| < 1$. Similarly,

$$\begin{aligned}\gamma_x(1) &= \text{cov}(x_t, x_{t-1}) = \text{cov}(\phi x_{t-1} + w_t, x_{t-1})\\ &= \text{cov}(\phi x_{t-1}, x_{t-1}) = \phi\gamma_x(0).\end{aligned}$$

Thus,

$$\rho_x(1) = \frac{\gamma_x(1)}{\gamma_x(0)} = \phi\,,$$

and we see that ϕ is in fact a correlation, $\phi = \text{corr}(x_t, x_{t-1})$.

It should be evident that we have to be careful when working with AR models. It should also be evident that, as in Example 1.9, simply setting the initial conditions equal to zero does not meet the stationary criteria because x_0 is not a constant, but a random variable with mean μ and variance $\sigma_w^2/(1-\phi^2)$. ◊

In Section 1.3, we discussed the notion that it is possible to generate realistic time series models by filtering white noise. In fact, there is a result by Wold (1954) that states that any (non-deterministic[1]) stationary time series is in fact a filter of white noise.

Property 2.21 (Wold Decomposition). *Any stationary time series, x_t, can be written as linear combination (filter) of white noise terms; that is,*

$$x_t = \mu + \sum_{j=0}^{\infty}\psi_j w_{t-j}, \tag{2.15}$$

where the ψs are numbers satisfying $\sum_{j=0}^{\infty}\psi_j^2 < \infty$ and $\psi_0 = 1$. We call these **linear processes**.

[1]This means that no part of the series is deterministic, meaning one where the future is perfectly predictable from the past; e.g., model (1.6).

Remark. Property 2.21 is important in the following ways:

- As previously suggested, stationary time series can be thought of as filters of white noise. It may not always be the best model, but models of this form are viable in many situations.
- Any stationary time series can be represented as a model that does not depend on the future. That is, x_t in (2.15) depends only on the present w_t and the past $w_{t-1}, w_{t-2}, \ldots$.
- Because the coefficients satisfy $\psi_j^2 \to 0$ as $j \to \infty$, the dependence on the distant past is negligible. Many of the models we will encounter satisfy the much stronger condition $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

The models we will encounter in Chapter 4 are linear processes. For the linear process, we may show that the mean function is $E(x_t) = \mu$, and the autocovariance function is given by

$$\gamma(h) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_{j+h}\psi_j \tag{2.16}$$

for $h \geq 0$; recall that $\gamma(-h) = \gamma(h)$. To see (2.16), note that

$$\begin{aligned}
\gamma(h) &= \operatorname{cov}(x_{t+h}, x_t) = \operatorname{cov}\Big(\sum_{j=0}^{\infty} \psi_j w_{t+h-j},\ \sum_{k=0}^{\infty} \psi_k w_{t-k}\Big) \\
&= \operatorname{cov}\big[w_{t+h} + \cdots + \psi_h w_t + \psi_{h+1} w_{t-1} + \cdots,\ \psi_0 w_t + \psi_1 w_{t-1} + \cdots\big] \\
&= \sigma_w^2 \sum_{j=0}^{\infty} \psi_{h+j}\psi_j\,.
\end{aligned}$$

The moving average model is already in the form of a linear process. The autoregressive model such as the one in Example 1.9 can also be put in this form as we suggested in that example.

When several series are available, a notion of stationarity still applies with additional conditions.

Definition 2.22. *Two time series, say, x_t and y_t, are* **jointly stationary** *if they are each stationary, and the* cross-covariance function

$$\gamma_{xy}(h) = \operatorname{cov}(x_{t+h}, y_t) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)] \tag{2.17}$$

is a function only of lag h.

Definition 2.23. *The* **cross-correlation function (CCF)** *of jointly stationary time series x_t and y_t is defined as*

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}. \tag{2.18}$$

As usual, we have the result $-1 \leq \rho_{xy}(h) \leq 1$ which enables comparison with the extreme values -1 and 1 when looking at the relation between x_{t+h} and y_t. The cross-correlation function is *not* generally symmetric about zero because when $h > 0$, y_t happens before x_{t+h} whereas when $h < 0$, y_t happens after x_{t+h}.

Example 2.24. Joint Stationarity
Consider the two series, x_t and y_t, formed from the sum and difference of two successive values of a white noise process, say,

$$x_t = w_t + w_{t-1} \quad \text{and} \quad y_t = w_t - w_{t-1},$$

where w_t is white noise with variance σ_w^2. It is easy to show that $\gamma_x(0) = \gamma_y(0) = 2\sigma_w^2$ because the w_ts are uncorrelated. In addition,

$$\gamma_x(1) = \text{cov}(x_{t+1},\, x_t) = \text{cov}(w_{t+1} + w_t,\, w_t + w_{t-1}) = \sigma_w^2$$

and $\gamma_x(-1) = \gamma_x(1)$; similarly $\gamma_y(1) = \gamma_y(-1) = -\sigma_w^2$. Also,

$$\begin{aligned}
\gamma_{xy}(0) &= \text{cov}(x_t, y_t) = \text{cov}(w_{t+1} + w_t, w_{t+1} - w_t) = \sigma_w^2 - \sigma_w^2 = 0\,; \\
\gamma_{xy}(1) &= \text{cov}(x_{t+1}, y_t) = \text{cov}(w_{t+1} + w_t, w_t - w_{t-1}) = \sigma_w^2\,; \\
\gamma_{xy}(-1) &= \text{cov}(x_{t-1}, y_t) = \text{cov}(w_{t-1} + w_{t-2}, w_t - w_{t-1}) = -\sigma_w^2\,.
\end{aligned}$$

Noting that $\text{cov}(x_{t+h}, y_t) = 0$ for $|h| > 2$, using (2.18) we have,

$$\rho_{xy}(h) = \begin{cases} 0 & h = 0, \\ \frac{1}{2} & h = 1, \\ -\frac{1}{2} & h = -1, \\ 0 & |h| \geq 2. \end{cases}$$

Clearly, the autocovariance and cross-covariance functions depend only on the lag separation, h, so the series are jointly stationary. ◊

Example 2.25. Prediction via Cross-Correlation
Consider the problem of determining leading or lagging relations between two stationary series x_t and y_t. If for some unknown integer ℓ, the model

$$y_t = Ax_{t-\ell} + w_t$$

holds, the series x_t is said to **lead** y_t for $\ell > 0$ and is said to **lag** y_t for $\ell < 0$. Estimating the lead or lag relations might be important in predicting the value of y_t from x_t. Assuming that the noise w_t is uncorrelated with the x_t series, the cross-covariance function can be computed as

$$\begin{aligned}
\gamma_{yx}(h) &= \text{cov}(y_{t+h}, x_t) = \text{cov}(Ax_{t+h-\ell} + w_{t+h},\, x_t) \\
&= \text{cov}(Ax_{t+h-\ell},\, x_t) = A\gamma_x(h - \ell)\,.
\end{aligned}$$

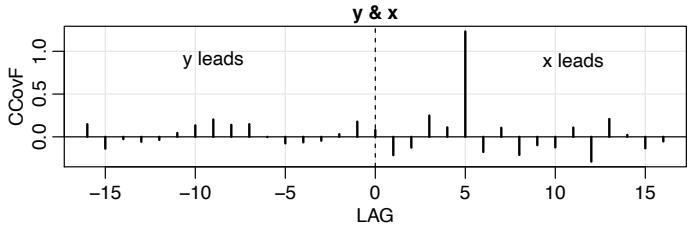


Figure 2.2 *Demonstration of the results of Example 2.25 when $\ell = 5$. The title indicates which series is leading.*

Since the largest value of $|\gamma_x(h-\ell)|$ is $\gamma_x(0)$, i.e., when $h = \ell$, the cross-covariance function will look like the autocovariance of the input series x_t, and it will have an extremum on the positive side if x_t leads y_t and an extremum on the negative side if x_t lags y_t. Below is the R code of an example with a delay of $\ell = 5$ and $\hat{\gamma}_{yx}(h)$, which is defined in Definition 2.30, shown in Figure 2.2.

```
x = rnorm(100)
y = lag(x,-5) + rnorm(100)
ccf(y, x, ylab="CCovF", type="covariance", panel.first=grid())
```

◊

2.3 Estimation of Correlation

For data analysis, only the sample values, $x_1, x_2, \ldots, x_n$, are available for estimating the mean, autocovariance, and autocorrelation functions. *In this case, the assumption of stationarity becomes critical and allows the use of averaging to estimate the population mean and covariance functions.*

Accordingly, if a time series is stationary, the mean function (2.10) $\mu_t = \mu$ is constant so we can estimate it by the *sample mean*,

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t. \tag{2.19}$$

The estimate is unbiased, $E(\bar{x}) = \mu$, and its standard error is the square root of $\mathrm{var}(\bar{x})$, which can be computed using first principles (Property 2.7), and is given by

$$\mathrm{var}(\bar{x}) = \frac{1}{n}\sum_{h=-n}^{n}\Big(1 - \frac{|h|}{n}\Big)\gamma_x(h)\,. \tag{2.20}$$

If the process is white noise, (2.20) reduces to the familiar σ_x^2/n recalling that $\gamma_x(0) = \sigma_x^2$. Note that in the case of dependence, the standard error of $\bar{x}$ may be smaller or larger than the white noise case depending on the nature of the correlation structure (see Problem 2.10).

The theoretical autocorrelation function, (2.12), is estimated by the sample ACF as follows.

STAT 626: Outline of Lecture 8
Estimation of the Mean, Correlation and ACF (§2.3)

1. A Quick Review of How Statistics Works: **Sample** $\Longrightarrow$ **Population**.

   Estimation of μ, $\gamma(1)$, $\rho(1), \ldots$

2. The Sample Mean: $\bar{x}$

3. Sample Autocovariance Function: $\widehat{\gamma}(1)$

4. Distribution of the Sample Autocorrelation Function (ACF)

5. **The Sample Correlogram and Confidence Interval**

6. Bivariate Time Series: Estimation of Cross-Correlations.

Review of the DEFINITIONS

1. A Time Series $\{x_t\}$ is **stationary** if

   (a) the mean function $E(x_t)$ does not depend on the time t,

   (b) the covariance function $\text{cov}(x_s, x_t)$ depends on the times s, t only through the (time-)lag $|s-t|$.

2. **Autocovariance Function** of a Stationary Time Series:

$$\gamma(h) = \text{cov}(x_{t+h}, x_t), \quad h = 0, 1, \ldots.$$

   NOTE: Setting $h = 0$ it follows that

$$\gamma(0) = \text{cov}(x_t, x_t) = \text{var}(x_t),$$

   so that the variance of the series, just like its mean, is not time-varying.

3. **The Autocorrelation Function (ACF)**

$$\rho(h) = \frac{\text{cov}(x_{t+h}, x_t)}{\sqrt{\text{var}(x_{t+h})\text{var}(x_t)}} = \frac{\gamma(h)}{\gamma(0)}, \quad h = 0, 1, \ldots.$$

4. **Correlogram** is the plot of $\rho(h)$ vs h.
   Its role in identifying TS models is just like that of the histogram in basic statistics.

LINEAR PROCESSES are the most general form of stationary processes, they are formed as linear combinations of a **white noise** $\{w_t\} \sim \text{WN}(0, \sigma_w^2)$.

Moving Average of order q or MA(q) Models:

$$x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q},$$

where $\theta = (\theta_1, \ldots, \theta_q)$ is the vector of parameters.

What happens when $q = \infty$?
MA(∞) Models or Processes.

Example: Compute the ACF of MA(∞) when $\theta_i = \phi^i,\ i = 1, \ldots,$ for a $|\phi| < 1$.

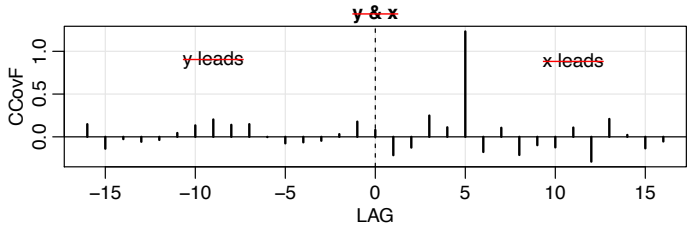


Figure 2.2 *Demonstration of the results of Example 2.25 when $\ell = 5$. The title indicates which series is leading.*

Since the largest value of $|\gamma_{x}(h-\ell)|$ is $\gamma_x(0)$, i.e., when $h = \ell$, the cross-covariance function will look like the autocovariance of the input series x_t, and it will have an extremum on the positive side if x_t leads y_t and an extremum on the negative side if x_t lags y_t. Below is the R code of an example with a delay of $\ell = 5$ and $\hat{\gamma}_{yx}(h)$, which is defined in Definition 2.30, shown in Figure 2.2.

```
x = rnorm(100)
y = lag(x,-5) + rnorm(100)
ccf(y, x, ylab="CCovF", type="covariance", panel.first=grid())
```

◊

2.3 Estimation of Correlation

For data analysis, only the sample values, $x_1, x_2, \ldots, x_n$, are available for estimating the mean, autocovariance, and autocorrelation functions. *In this case, the assumption of stationarity becomes critical and allows the use of averaging to estimate the population mean and covariance functions.*

Accordingly, if a time series is stationary, the mean function (2.10) $\mu_t = \mu$ is constant so we can estimate it by the *sample mean*,

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t. \tag{2.19}$$

The estimate is unbiased, $E(\bar{x}) = \mu$, and its standard error is the square root of $\mathrm{var}(\bar{x})$, which can be computed using first principles (Property 2.7), and is given by

$$\mathrm{var}(\bar{x}) = \frac{1}{n}\sum_{h=-n}^{n}\Big(1 - \frac{|h|}{n}\Big)\gamma_x(h)\,. \tag{2.20}$$

If the process is white noise, (2.20) reduces to the familiar σ_x^2/n recalling that $\gamma_x(0) = \sigma_x^2$. Note that in the case of dependence, the standard error of $\bar{x}$ may be smaller or larger than the white noise case depending on the nature of the correlation structure (see Problem 2.10).

The theoretical autocorrelation function, (2.12), is estimated by the sample ACF as follows.

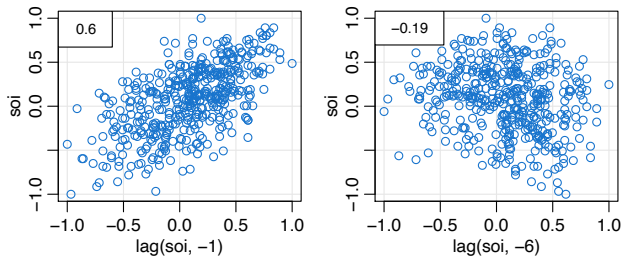


Figure 2.3 *Display for Example 2.27. For the SOI series, we have a scatterplot of pairs of values one month apart (left) and six months apart (right). The estimated autocorrelation is displayed in the box.*

Definition 2.26. *The* **sample autocorrelation function (ACF)** *is defined as*

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)} = \frac{\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \tag{2.21}$$

for $h = 0, 1, \ldots, n-1$.

The sum in the numerator of (2.21) runs over a restricted range because x_{t+h} is not available for $t + h > n$. Note that we are in fact estimating the autocovariance function by

$$\widehat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}), \tag{2.22}$$

with $\widehat{\gamma}(-h) = \widehat{\gamma}(h)$ for $h = 0, 1, \ldots, n-1$. That is, we divide by n even though there are only $n - h$ pairs of observations at lag h,

$$\{(x_{t+h},\, x_t);\ t = 1, \ldots, n-h\}\,. \tag{2.23}$$

This assures that the sample autocovariance function will behave as a true autocovariance function, and for example, will not give negative values when estimating $\mathrm{var}(\bar{x})$ by replacing $\gamma_x(h)$ with $\widehat{\gamma}_x(h)$ in (2.20).

Example 2.27. Sample ACF and Scatterplots

Estimating autocorrelation is similar to estimating of correlation in the classical case, but we use (2.21) instead of the sample correlation coefficient you learned in a course on regression. Figure 2.3 shows an example using the SOI series where $\widehat{\rho}(1) = .60$ and $\widehat{\rho}(6) = -.19$. The following code was used for Figure 2.3.

```
(r = acf1(soi, 6, plot=FALSE))  # sample acf values
  [1]  0.60  0.37  0.21  0.05 -0.11 -0.19
par(mfrow=c(1,2), mar=c(2.5,2.5,0,0)+.5, mgp=c(1.6,.6,0))
plot(lag(soi,-1), soi, col="dodgerblue3", panel.first=grid(lty=1))
legend("topleft", legend=r[1], bg="white", adj=.45, cex = 0.85)
plot(lag(soi,-6), soi, col="dodgerblue3", panel.first=grid(lty=1))
legend("topleft", legend=r[6], bg="white", adj=.25, cex = 0.8)
```

◊

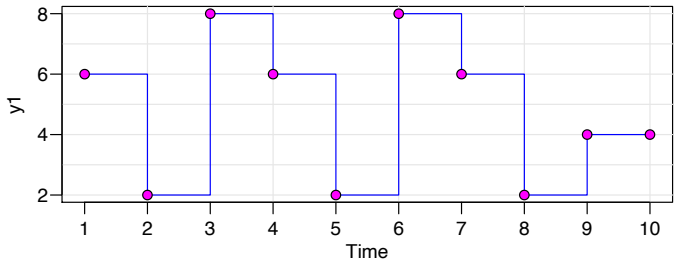


Figure 2.4: *Realization of* (2.24), $n = 10$.

Remark. It is important to note that this approach to estimating correlation ***makes sense only if the data are stationary.*** If the data were not stationary, each point in the graph could be an observation from a different correlation structure.

The sample autocorrelation function has a sampling distribution that allows us to assess whether the data comes from a completely random or white series or whether correlations are statistically significant at some lags.

Property 2.28 (Large-Sample Distribution of the ACF). *If x_t is white noise, then for n large and under mild conditions, the sample ACF, $\widehat{\rho}_x(h)$, for $h = 1, 2, \ldots, H$, where H is fixed but arbitrary, is approximately normal with zero mean and standard deviation given by of $1/\sqrt{n}$.*

Based on Property 2.28, we obtain a rough method for assessing whether a series is white noise by determining how many values of $\widehat{\rho}(h)$ are outside the interval $\pm 2/\sqrt{n}$ (two standard errors); for white noise, approximately 95% of the sample ACFs should be within these limits.[2]

Example 2.29. A Simulated Time Series

To compare the sample ACF for various sample sizes to the theoretical ACF, consider a contrived set of data generated by tossing a fair coin, letting $x_t = 2$ when a head is obtained and $x_t = -2$ when a tail is obtained. Then, because we can only appreciate 2, 4, 6, or 8, we let

$$y_t = 5 + x_t - .5x_{t-1}\,. \tag{2.24}$$

We consider two cases, one with a small sample size ($n = 10$; see Figure 2.4) and another with a moderate sample size ($n = 100$).

```
set.seed(101011)
x1 = sample(c(-2,2),  11, replace=TRUE) # simulated coin tosses
x2 = sample(c(-2,2), 101, replace=TRUE)
y1 = 5 + filter(x1, sides=1, filter=c(1,-.5))[-1]
y2 = 5 + filter(x2, sides=1, filter=c(1,-.5))[-1]
tsplot(y1, type="s", col=4, xaxt="n", yaxt="n")  # y2 not shown
axis(1, 1:10); axis(2, seq(2,8,2), las=1)
```

[2] In this text, $z_{.025} = 1.95996398454\ldots$ of normal fame, often rounded to 1.96, is rounded to 2.

```
points(y1, pch=21, cex=1.1, bg=6)
acf(y1, lag.max=4, plot=FALSE)  # 1/√10 =.32
      0      1      2      3      4
  1.000 -0.352 -0.316  0.510 -0.245
acf(y2, lag.max=4, plot=FALSE)  # 1/√100 =.1
      0      1      2      3      4
  1.000 -0.496  0.067  0.087  0.063
```

The theoretical ACF can be obtained from the model (2.24) using first principles so that

$$\rho_y(1) = \frac{-.5}{1+.5^2} = -.4$$

and $\rho_y(h) = 0$ for $|h| > 1$ (do Problem 2.15 now). It is interesting to compare the theoretical ACF with sample ACFs for the realization where $n = 10$ and where $n = 100$; note that small sample size means increased variability. ◊

Definition 2.30. *The estimators for the cross-covariance function, $\gamma_{xy}(h)$, as given in (2.17) and the cross-correlation, $\rho_{xy}(h)$, in (2.18) are given, respectively, by the* **sample cross-covariance function**

$$\widehat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}), \tag{2.25}$$

where $\widehat{\gamma}_{xy}(-h) = \widehat{\gamma}_{yx}(h)$ determines the function for negative lags, and the **sample cross-correlation function**

$$\widehat{\rho}_{xy}(h) = \frac{\widehat{\gamma}_{xy}(h)}{\sqrt{\widehat{\gamma}_x(0)\widehat{\gamma}_y(0)}}. \tag{2.26}$$

The sample cross-correlation function can be examined graphically as a function of lag h to search for leading or lagging relations in the data using the property mentioned in Example 2.25 for the theoretical cross-covariance function. Because $-1 \leq \widehat{\rho}_{xy}(h) \leq 1$, the practical importance of peaks can be assessed by comparing their magnitudes with their theoretical maximum values.

Property 2.31 (Large-Sample Distribution of Cross-Correlation). *If x_t and y_t are independent processes, then under mild conditions, the large sample distribution of $\widehat{\rho}_{xy}(h)$ is normal with mean zero and standard deviation $1/\sqrt{n}$ if at least one of the processes is independent white noise.*

Example 2.32. SOI and Recruitment Correlation Analysis

The autocorrelation and cross-correlation functions are also useful for analyzing the joint behavior of two stationary series whose behavior may be related in some unspecified way. In Example 1.4 (see Figure 1.5), we have considered simultaneous monthly readings of the SOI and an index for the number of new fish (Recruitment).

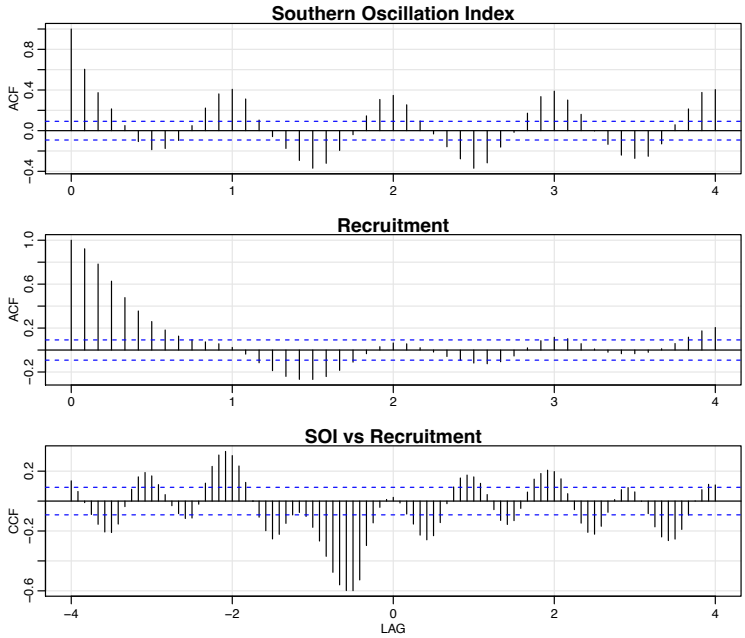


Figure 2.5 *Sample ACFs of the SOI series (top) and of the Recruitment series (middle), and the sample CCF of the two series (bottom); negative lags indicate SOI leads Recruitment. The lag axes are in terms of seasons (12 months).*

Figure 2.5 shows the sample autocorrelation and cross-correlation functions (ACFs and CCF) for these two series.

Both of the ACFs exhibit periodicities corresponding to the correlation between values separated by 12 units. Observations 12 months or one year apart are strongly positively correlated, as are observations at multiples such as $24, 36, 48, \ldots$ Observations separated by six months are negatively correlated, showing that positive excursions tend to be associated with negative excursions six months removed. This appearance is rather characteristic of the pattern that would be produced by a sinusoidal component with a period of 12 months; see Example 2.33. The cross-correlation function peaks at $h = -6$, showing that the SOI measured at time $t - 6$ months is associated with the Recruitment series at time t. We could say the SOI leads the Recruitment series by six months. The sign of the CCF at $h = -6$ is negative, leading to the conclusion that the two series move in different directions; that is, increases in SOI lead to decreases in Recruitment and vice versa. Again, note the periodicity of 12 months in the CCF.

The flat lines shown on the plots indicate $\pm 2/\sqrt{453}$, so that upper values would be exceeded about 2.5% of the time if the noise were white as specified in Property 2.28 and Property 2.31. Of course, neither series is noise, so we can ignore these lines. To reproduce Figure 2.5 in R, use the following commands:

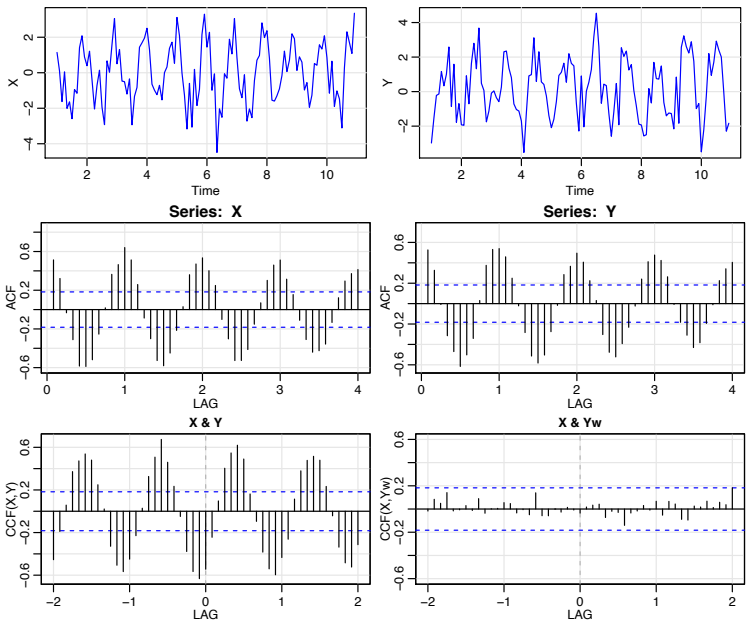


Figure 2.6: *Display for Example 2.33*

```
par(mfrow=c(3,1))
acf1(soi, 48, main="Southern Oscillation Index")
acf1(rec, 48, main="Recruitment")
ccf2(soi, rec, 48, main="SOI vs Recruitment")
```

◊

Example 2.33. Prewhitening and Cross Correlation Analysis *
Although we do not have all the tools necessary yet, it is worthwhile discussing the idea of prewhitening a series prior to a cross-correlation analysis. The basic idea is simple, to use Property 2.31, at least one of the series must be white noise. If this is not the case, there is no simple way of telling if a cross-correlation estimate is significantly different from zero. Hence, in Example 2.32, we were only guessing at the linear dependence relationship between SOI and Recruitment. The preferred method of prewhitening a time series is discussed in Section 8.5.

For example, in Figure 2.6 we generated two series, x_t and y_t, for $t = 1, \ldots, 120$ independently as

$$x_t = 2\cos(2\pi\, t\, \tfrac{1}{12}) + w_{t1} \quad \text{and} \quad y_t = 2\cos(2\pi\, [t+5]\, \tfrac{1}{12}) + w_{t2}$$

where $\{w_{t1}, w_{t2};\ t = 1, \ldots, 120\}$ are all independent standard normals. The series are made to resemble SOI and Recruitment. The generated data are shown in the top row of the figure. The middle row of Figure 2.6 shows the sample ACF of each series, each of which exhibits the cyclic nature of each series. The bottom row (left) of Figure 2.6 shows the sample CCF between x_t and y_t, which appears to show

cross-correlation even though the series are independent. The bottom row (right) also displays the sample CCF between x_t and the prewhitened y_t, which shows that the two sequences are uncorrelated. By prewhtiening y_t, we mean that the signal has been removed from the data by running a regression of y_t on $\cos(2\pi t/12)$ and $\sin(2\pi t/12)$ (both are needed to capture the phase; see Example 3.15) and then putting $\tilde{y}_t = y_t - \hat{y}_t$, where $\hat{y}_t$ are the predicted values from the regression.

The following code will reproduce Figure 2.6.

```
set.seed(1492)
num  = 120
t    = 1:num
X    = ts( 2*cos(2*pi*t/12)     + rnorm(num), freq=12 )
Y    = ts( 2*cos(2*pi*(t+5)/12) + rnorm(num), freq=12 )
Yw   = resid(lm(Y~ cos(2*pi*t/12) + sin(2*pi*t/12), na.action=NULL))
par(mfrow=c(3,2))
tsplot(X, col=4);  tsplot(Y, col=4)
acf1(X, 48); acf1(Y, 48)
ccf2(X, Y, 24);  ccf2(X, Yw, 24, ylim=c(-.6,.6))
```

◊

Problems

2.1. In 25 words or less, and without using symbols, why is stationarity important?

2.2. Consider the time series

$$x_t = \beta_0 + \beta_1 t + w_t,$$

where β_0 and β_1 are regression coefficients, and w_t is a white noise process with variance σ_w^2.

(a) Determine whether x_t is stationary.

(b) Show that the process $y_t = x_t - x_{t-1}$ is stationary.

(c) Show that the mean of the two-sided moving average

$$v_t = \tfrac{1}{3}(x_{t-1} + x_t + x_{t+1})$$

is $\beta_0 + \beta_1 t$.

2.3. When smoothing time series data, it is sometimes advantageous to give decreasing amounts of weights to values farther away from the center. Consider the simple two-sided moving average smoother of the form

$$x_t = \tfrac{1}{4}(w_{t-1} + 2w_t + w_{t+1}),$$

where w_t are independent with zero mean and variance σ_w^2. Determine the autocovariance and autocorrelation functions as a function of lag h and sketch the ACF as a function of h.

2.4. We have not discussed the stationarity of autoregressive models, and we will do that in Chapter 4. But for now, let $x_t = \phi x_{t-1} + w_t$ where $w_t \sim wn(0,1)$ and ϕ is a constant. Assume x_t is stationary and x_{t-1} is uncorrelated with the noise term w_t.

(a) Show that mean function of x_t is $\mu_{xt} = 0$.

(b) Show $\gamma_x(0) = \text{var}(x_t) = 1/(1-\phi^2)$.

(c) For which values of ϕ does the solution to part (b) make sense?

(d) Find the lag-one autocorrelation, $\rho_x(1)$.

2.5. Consider the random walk with drift model

$$x_t = \delta + x_{t-1} + w_t,$$

for $t = 1, 2, \ldots$, with $x_0 = 0$, where w_t is white noise with variance σ_w^2.

(a) Show that the model can be written as $x_t = \delta t + \sum_{k=1}^{t} w_k$.

(b) Find the mean function and the autocovariance function of x_t.

(c) Argue that x_t is not stationary.

(d) Show $\rho_x(t-1,t) = \sqrt{\frac{t-1}{t}} \to 1$ as $t \to \infty$. What is the implication of this result?

(e) Suggest a transformation to make the series stationary, and prove that the transformed series is stationary.

2.6. Would you treat the global temperature data discussed in Example 1.2 and shown in Figure 1.2 as stationary or non-stationary? Support your answer.

2.7. A time series with a periodic component can be constructed from

$$x_t = U_1 \sin(2\pi\omega_0 t) + U_2 \cos(2\pi\omega_0 t),$$

where U_1 and U_2 are independent random variables with zero means and $E(U_1^2) = E(U_2^2) = \sigma^2$. The constant ω_0 determines the period or time it takes the process to make one complete cycle. Show that this series is weakly stationary with autocovariance function

$$\gamma(h) = \sigma^2 \cos(2\pi\omega_0 h).$$

2.8. Consider the two series

$$x_t = w_t$$

$$y_t = w_t - \theta w_{t-1} + u_t,$$

where w_t and u_t are independent white noise series with variances σ_w^2 and σ_u^2, respectively, and θ is an unspecified constant.

(a) Express the ACF, $\rho_y(h)$, for $h = 0, \pm 1, \pm 2, \ldots$ of the series y_t as a function of σ_w^2, σ_u^2, and θ.

(b) Determine the CCF, $\rho_{xy}(h)$ relating x_t and y_t.

(c) Show that x_t and y_t are jointly stationary.

2.9. Let w_t, for $t = 0, \pm 1, \pm 2, \ldots$ be a normal white noise process, and consider the series

$$x_t = w_t w_{t-1}.$$

Determine the mean and autocovariance function of x_t, and state whether it is stationary.

2.10. Suppose $x_t = \mu + w_t + \theta w_{t-1}$, where $w_t \sim wn(0, \sigma_w^2)$.

(a) Show that mean function is $E(x_t) = \mu$.

(b) Show that the autocovariance function of x_t is given by $\gamma_x(0) = \sigma_w^2(1+\theta^2)$, $\gamma_x(\pm 1) = \sigma_w^2\theta$, and $\gamma_x(h) = 0$ otherwise.

(c) Show that x_t is stationary for all values of $\theta \in \mathbb{R}$.

(d) Use (2.20) to calculate $\mathrm{var}(\bar{x})$ for estimating μ when (i) $\theta = 1$, (ii) $\theta = 0$, and (iii) $\theta = -1$

(e) In time series, the sample size n is typically large, so that $\frac{(n-1)}{n} \approx 1$. With this as a consideration, comment on the results of part (d); in particular, how does the accuracy in the estimate of the mean μ change for the three different cases?

2.11.(a) Simulate a series of $n = 500$ Gaussian white noise observations as in Example 1.7 and compute the sample ACF, $\widehat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. [Recall Example 2.17.]

(b) Repeat part (a) using only $n = 50$. How does changing n affect the results?

2.12.(a) Simulate a series of $n = 500$ moving average observations as in Example 1.8 and compute the sample ACF, $\widehat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. [Recall Example 2.18.]

(b) Repeat part (a) using only $n = 50$. How does changing n affect the results?

2.13. Simulate 500 observations from the AR model specified in Example 1.9 and then plot the sample ACF to lag 50. What does the sample ACF tell you about the approximate cyclic behavior of the data? Hint: Recall Example 2.32.

2.14. Simulate a series of $n = 500$ observations from the signal-plus-noise model presented in Example 1.11 with (a) $\sigma_w = 0$, (b) $\sigma_w = 1$ and (c) $\sigma_w = 5$. Compute the sample ACF to lag 100 of the three series you generated and comment.

2.15. For the time series y_t described in Example 2.29, verify the stated result that $\rho_y(1) = -.4$ and $\rho_y(h) = 0$ for $h > 1$.

STAT 626: Outline of Lectures 10-11
Time Series Regression and EDA (§3.1, 3.2)

1. **Review of Stationarity, Preview of TS Models**

2. **Example 3.1: Estimating the Linear Trend**

3. **A Quick Review of Multiple Regression**

$$x = Z\beta + w,$$

LSE of β :

$$\widehat{\beta} = (Z'Z)^{-1}Z'x.$$

4. **Tests, CIs and Variable Selection**

5. **AIC (Akaike Information Criterion)**, BIC

6. **Example 3.5: Pollution, Temperature and Mortality (PTM) Data**

7. **Example 3.6: Regression with Lagged Variables**

8. **Example 3.7: Detrending and Differencing**

9. **The Backshift Operator B:** $Bx_t = x_{t-1}$.

Regression and Forecasting:

PROBLEM (POPULATION INFORMATION): Given the value of a random variable X, **find β to minimize the mean-square error (MSE) of predicting Y by $\widehat{Y} = \beta X$**:

$$\mathrm{MSE}(\beta) = E(Y - \beta X)^2.$$

SOLUTION: The minimizer satisfies the *normal equation*:

$$\mathrm{Var}(X) \ \ \widehat{\beta} = \mathrm{Cov}(X, Y)$$

or

$$\widehat{\beta} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)},$$

and

$$\mathrm{MSE}(\widehat{\beta}) = E(Y - \widehat{Y})^2 = (1 - \rho^2)\mathrm{Var}(Y).$$

Time Series Prediction: Given the time series data $x_1, \ldots, x_n$ from a zero-mean stationary process $\{x_t\}$ with **known** autocovariance function,$\gamma(h)$ find the forecast value of the process at the next time point, x_{n+1}. More precisely, find $\phi_{n1}, \ldots, \phi_n$ to minimize the MSE of forecasting:

$$E(x_{n+1} - \phi_{n1}x_n - \ldots - \phi_{nn}x_1)^2.$$

Review of Stationarity, Preview of TS Models (Chapter 4)

1. **Linear Processes:** $x_t = \mu + \sum_{j=-\infty}^{+\infty} \psi_j w_{t-j}$ is stationary with the *autocovariance function*

$$\gamma(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j.$$

2. **MA(q) Models**: $x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}, \quad \theta_q \neq 0$, is stationary, its autocovariance is zero at lags greater than q.

3. **Autoregressive Models of order p or AR(p) Models:**

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t, \quad \phi_p \neq 0.$$

QUESTION: Is a time series $\{x_t\}$ defined via an AR(p) model always stationary? If so, what is its autocovariance function?
To get a feel for the answer consider the AR(1):

$$x_t = \phi x_{t-1} + w_t,$$

what happens when $\phi = 1$?

4. **The Backshift Operator B:** $Bx_t = x_{t-1}$.

5. **MA(q) and B:**

$$x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q} = (1 + \theta_1 B + \ldots + \theta_q B^q) w_t = \theta(B) w_t.$$

6. **AR(p) and B:**

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = w_t, \quad (1 - \phi_1 B - \ldots - \phi_p B^p) x_t = \phi(B) x_t = w_t.$$

7. **The ROOTS of the polynomial equation**

$$\phi(B) = 0,$$

hold the key to the question of stationarity of the solutions of AR models.

Chapter 3

Time Series Regression and EDA

3.1 Ordinary Least Squares for Time Series

We first consider the problem where a time series, say, x_t, for $t = 1, \ldots, n$, is possibly being influenced by a collection of fixed series, say, $z_{t1}, z_{t2}, \ldots, z_{tq}$. The data collection with $q = 3$ exogenous variables is as follows:

Time	Dependent Variable	Independent Variables		
1	x_1	z_{11}	z_{12}	z_{13}
2	x_2	z_{21}	z_{22}	z_{23}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	z_{n1}	z_{n2}	z_{n3}

We express the general relation through the *linear regression model*

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t, \tag{3.1}$$

where $\beta_0, \beta_1, \ldots, \beta_q$ are unknown fixed regression coefficients, and $\{w_t\}$ is white normal noise with variance σ_w^2; we will relax this assumption later.

Example 3.1. Estimating the Linear Trend of a Commodity

Consider the monthly export price of Norwegian salmon per kilogram from September 2003 to June 2017 shown in Figure 3.1. There is an obvious upward trend in the series, and we might use simple linear regression to estimate that trend by fitting the model,

$$x_t = \beta_0 + \beta_1 z_t + w_t, \quad z_t = 2003\tfrac{8}{12}, 2001\tfrac{8}{12}, \ldots, 2017\tfrac{5}{12}.$$

This is in the form of the regression model (3.1) with $q = 1$. The data x_t are in `salmon` and z_t is month, with values in `time(salmon)`. Our assumption that the error, w_t, is white noise is probably not true, but we will assume it is true for now. The problem of autocorrelated errors will be discussed in detail in Section 5.4.

In ordinary least squares (OLS), we minimize the error sum of squares

$$S = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - [\beta_0 + \beta_1 z_t])^2$$

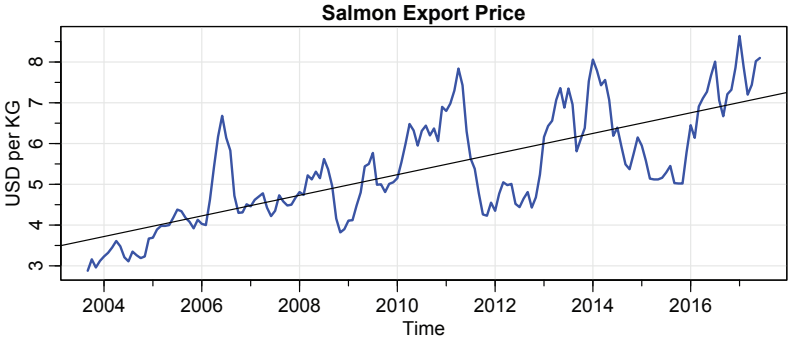


Figure 3.1 *The monthly export price of Norwegian salmon per kilogram from September 2003 to June 2017, with fitted linear trend line.*

with respect to β_i for $i = 0, 1$. In this case we can use simple calculus to evaluate $\partial S/\partial \beta_i = 0$ for $i = 0, 1$, to obtain two equations to solve for the βs. The OLS estimates of the coefficients are explicit and given by

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(z_t - \bar{z})}{\sum_{t=1}^{n}(z_t - \bar{z})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{z},$$

where $\bar{x} = \sum_t x_t/n$ and $\bar{z} = \sum_t z_t/n$ are the respective sample means.

Using R, we obtained the estimated slope coefficient of $\hat{\beta}_1 = .25$ (with a standard error of .02) yielding a highly significant estimated increase of about 25 cents *per year*.[1] Finally, Figure 3.1 shows the data with the estimated trend line superimposed. To perform this analysis in R, use the following commands:

```
summary(fit <- lm(salmon~time(salmon), na.action=NULL))
 Coefficients:
                Estimate Std. Error t value Pr(>|t|)
  (Intercept)  -503.08947   34.44164  -14.61   <2e-16
  time(salmon)    0.25290    0.01713   14.76   <2e-16
  ---
  Residual standard error: 0.8814 on 164 degrees of freedom
  Multiple R-squared:  0.5706,    Adjusted R-squared:  0.568
  F-statistic: 217.9 on 1 and 164 DF,  p-value: < 2.2e-16
tsplot(salmon, col=4, ylab="USD per KG", main="Salmon Export Price")
abline(fit)
```

◊

Simple linear regression extends to multiple linear regression in a fairly straightforward manner. As in the previous example, OLS estimation minimizes the error sum of squares

$$S = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n}(x_t - [\beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq}])^2, \qquad (3.2)$$

[1] The unit of time here is one year, $z_t - z_{t-12} = 1$. Thus $\hat{x}_t - \hat{x}_{t-12} = \hat{\beta}_1(z_t - z_{t-12}) = \hat{\beta}_1$.

with respect to $\beta_0, \beta_1, \ldots, \beta_q$. This minimization can be accomplished by solving $\partial S/\partial \beta_i = 0$ for $i = 0, 1, \ldots, q$, which yields $q+1$ equations with $q+1$ unknowns. These equations are typically called the *normal equations*. The minimized error sum of squares (3.2), denoted SSE, can be written as

$$SSE = \sum_{t=1}^{n} (x_t - \widehat{x}_t)^2, \tag{3.3}$$

where

$$\widehat{x}_t = \hat{\beta}_0 + \hat{\beta}_1 z_{t1} + \hat{\beta}_2 z_{t2} + \cdots + \hat{\beta}_q z_{tq},$$

and $\hat{\beta}_i$ denotes the OLS estimate of β_i for $i = 0, 1, \ldots, q$. The ordinary least squares estimators of the βs are unbiased and have the smallest variance within the class of linear unbiased estimators. An unbiased estimator for the variance σ_w^2 is

$$s_w^2 = MSE = \frac{SSE}{n-(q+1)}, \tag{3.4}$$

where MSE denotes the *mean squared error*. Because the errors are normal, if $\mathrm{se}(\hat{\beta}_i)$ represents the estimated standard error of the estimate of β_i, then

$$t = \frac{(\widehat{\beta}_i - \beta_i)}{\mathrm{se}(\hat{\beta}_i)} \tag{3.5}$$

has the t-distribution with $n-(q+1)$ degrees of freedom. This result is often used for individual tests of the null hypothesis $\mathrm{H}_0\colon \beta_i = 0$ for $i = 1, \ldots, q$.

Various competing models are often of interest to isolate or select the best subset of independent variables. Suppose a proposed model specifies that only a subset $r < q$ independent variables, say, $z_{t,1:r} = \{z_{t1}, z_{t2}, \ldots, z_{tr}\}$ is influencing the dependent variable x_t. The reduced model is

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_r z_{tr} + w_t \tag{3.6}$$

where $\beta_1, \beta_2, \ldots, \beta_r$ are a subset of coefficients of the original q variables.

The null hypothesis in this case is $\mathrm{H}_0\colon \beta_{r+1} = \cdots = \beta_q = 0$. We can test the reduced model (3.6) against the full model (3.1) by comparing the error sums of squares under the two models using the F-statistic

$$F = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-q-1)} = \frac{MSR}{MSE}, \tag{3.7}$$

where SSE_r is the error sum of squares under the reduced model (3.6). Note that $SSE_r \geq SSE$ because the reduced model has fewer parameters. If $\mathrm{H}_0\colon \beta_{r+1} = \cdots = \beta_q = 0$ is true, then $SSE_r \approx SSE$ because the estimates of those βs will be close to 0. Hence, we do not believe H_0 if $SSR = SSE_r - SSE$ is big. Under the null hypothesis, (3.7) has a central F-distribution with $q-r$ and $n-q-1$ degrees of freedom when (3.6) is the correct model.

Table 3.1 *Analysis of Variance for Regression*

Source	df	Sum of Squares	Mean Square	F
$z_{t,r+1:q}$	$q-r$	$SSR = SSE_r - SSE$	$MSR = SSR/(q-r)$	$F = \frac{MSR}{MSE}$
Error	$n-(q+1)$	SSE	$MSE = SSE/(n-q-1)$	

These results are often summarized in an ANOVA table as given in Table 3.1 for this particular case. The difference in the numerator is often called the regression sum of squares (SSR). The null hypothesis is rejected at level α if $F > F^{q-r}_{n-q-1}(\alpha)$, the $1-\alpha$ percentile of the F distribution with $q-r$ numerator and $n-q-1$ denominator degrees of freedom.

A special case of interest is $H_0\colon \beta_1 = \cdots = \beta_q = 0$. In this case $r = 0$, and the model in (3.6) becomes

$$x_t = \beta_0 + w_t\,.$$

The residual sum of squares under this reduced model is

$$SSE_0 = \sum_{t=1}^{n}(x_t - \bar{x})^2\,, \tag{3.8}$$

and SSE_0 is often called the *adjusted total sum of squares*, or SST (i.e., $SST = SSE_0$). In this case,

$$SST = SSR + SSE\,,$$

and we may measure the proportion of variation accounted for by all the variables using

$$R^2 = \frac{SSR}{SST}\,. \tag{3.9}$$

The measure R^2 is called the *coefficient of determination*.

The techniques discussed in the previous paragraph can be used for model selection; e.g., stepwise regression. Another approach is based on *parsimony* (also called *Occam's razor*) where we try to find the most *accurate* model with the least amount of *complexity*. For regression models, this means that we find the model that has the best fit with the fewest number of parameters. You may have been introduced to parsimony and model choice via Mallows C_p in a course on regression.

To measure accuracy, we use the error sum of squares, $SSE = \sum_{t=1}^{n}(x_t - \widehat{x}_t)^2$, because it measures how close the fitted values ($\widehat{x}_t$) are to the actual data (x_t). In particular, for a normal regression model with k coefficients, consider the (maximum likelihood) estimator for the variance as

$$\widehat{\sigma}^2_k = \frac{SSE(k)}{n}, \tag{3.10}$$

where by $SSE(k)$, we mean the residual sum of squares under the model with k regression coefficients. The complexity of the model can be characterized by k, the number of parameters in the model. Akaike (1974) suggested balancing the accuracy of the fit against the number of parameters in the model.

Definition 3.2. Akaike's Information Criterion (AIC)

$$\text{AIC} = \log \widehat{\sigma}_k^2 + \frac{n+2k}{n}, \tag{3.11}$$

where $\widehat{\sigma}_k^2$ *is given by* (3.10) *and* k *is the number of parameters in the model.*[2]

Thus, the parsimonious model will be an accurate one (with small error $\widehat{\sigma}_k$) that is not overly complex (small k). Hence, the model yielding the minimum AIC specifies the best model.

The choice for the penalty term given by (3.11) is not the only one, and a considerable literature is available advocating different penalty terms. A corrected form, suggested by Sugiura (1978), and expanded by Hurvich and Tsai (1989), can be based on small-sample distributional results for the linear regression model. The corrected form is defined as follows.

Definition 3.3. AIC, Bias Corrected (AICc)

$$\text{AICc} = \log \widehat{\sigma}_k^2 + \frac{n+k}{n-k-2}, \tag{3.12}$$

where $\widehat{\sigma}_k^2$ *is given by* (3.10)*,* k *is the number of parameters in the model.*

We may also derive a penalty term based on Bayesian arguments, as in Schwarz (1978), which leads to the following.

Definition 3.4. Bayesian Information Criterion (BIC)

$$\text{BIC} = \log \widehat{\sigma}_k^2 + \frac{k \log n}{n}, \tag{3.13}$$

using the same notation as in Definition 3.2.

BIC is also called the Schwarz Information Criterion (SIC). Various simulation studies have tended to verify that BIC does well at getting the correct order in large samples, whereas AICc tends to be superior in smaller samples where the relative number of parameters is large; see McQuarrie and Tsai (1998) for detailed comparisons.

Example 3.5. Pollution, Temperature, and Mortality

The data shown in Figure 3.2 are extracted series from a study by Shumway et al. (1988) of the possible effects of temperature and pollution on weekly mortality in Los Angeles County. Note the strong seasonal components in all of the series, corresponding to winter-summer variations and the downward trend in the cardiovascular mortality over the 10-year period.

Notice the inverse relationship between mortality and temperature; the mortality

[2]Formally, AIC is defined as $-2\log L_k + 2k$ where L_k is the maximum value of the likelihood and k is the number of parameters in the model. For the normal regression problem, AIC can be reduced to the form given by (3.11). For comparison, BIC is defined as $-2\log L_k + k\log n$, so complexity has a much larger penalty.

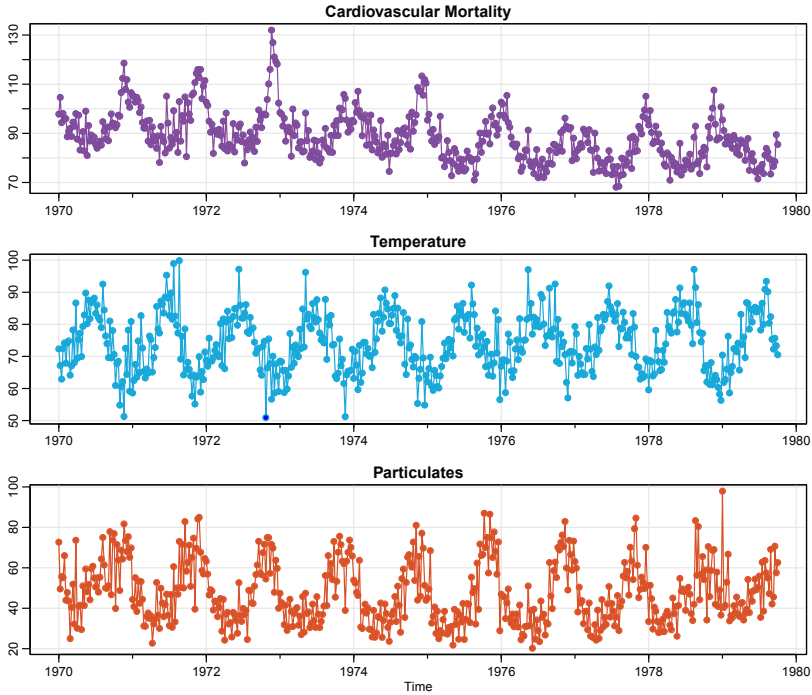


Figure 3.2 *Average weekly cardiovascular mortality (top), temperature (middle), and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10-year period 1970–1979.*

rate is higher for cooler temperatures. In addition, it appears that particulate pollution is directly related to mortality; the mortality rate increases for higher levels of pollution. These relationships can be better seen in Figure 3.3, where the data are plotted together. The time series plots were produced using the following R code:

```
##--  Figure 3.2  --##
culer = c(rgb(.66,.12,.85), rgb(.12,.66,.85), rgb(.85,.30,.12))
par(mfrow=c(3,1))
tsplot(cmort, main="Cardiovascular Mortality", col=culer[1],
           type="o", pch=19, ylab="")
tsplot(tempr, main="Temperature", col=culer[2], type="o", pch=19,
           ylab="")
tsplot(part, main="Particulates", col=culer[3], type="o", pch=19,
           ylab="")
##--  Figure 3.3  --##
tsplot(cmort, main="", ylab="", ylim=c(20,130), col=culer[1])
lines(tempr, col=culer[2])
lines(part, col=culer[3])
legend("topright", legend=c("Mortality", "Temperature", "Pollution"),
           lty=1, lwd=2, col=culer, bg="white")
```

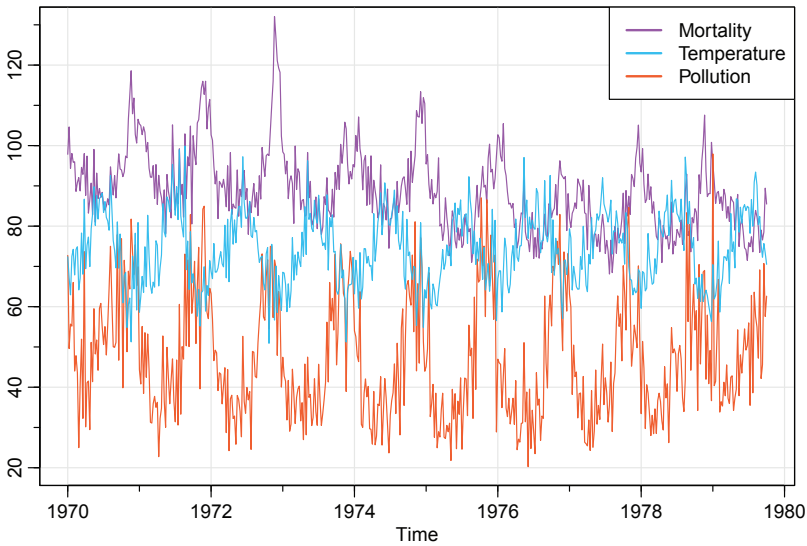


Figure 3.3 *Mortality data on same plot.*

To investigate these relationships further, a scatterplot matrix is shown in Figure 3.4 and indicates that cardiovascular mortality is linearly related to pollutant particulates, but is nonlinearly related to temperature. We note that the curvilinear shape of the temperature–mortality curve indicates that higher temperatures as well as lower temperatures are associated with increases in cardiovascular mortality. The scatterplot matrix shown in Figure 3.4 was generated in R as follows. The script `panel.cor` calculates the correlations between all the variables, and when called in `pairs`, inserts the corresponding correlation value.

```
panel.cor <- function(x, y, ...){
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- round(cor(x, y), 2)
    text(0.5, 0.5, r, cex = 1.75)
}
pairs(cbind(Mortality=cmort, Temperature=tempr, Particulates=part),
           col="dodgerblue3", lower.panel=panel.cor)
```

It is important that temperature and particulate pollution are nearly uncorrelated. If these two independent variables were highly correlated (i.e., collinear), then it would be difficult to distinguish between the effects of each on mortality.

For ease, let M_t denote cardiovascular mortality, T_t denote temperature, and P_t denote the particulate levels. Based on the scatterplot matrix, it seems clear that both T_t and P_t should be in the model, but for demonstration purposes, we entertain four

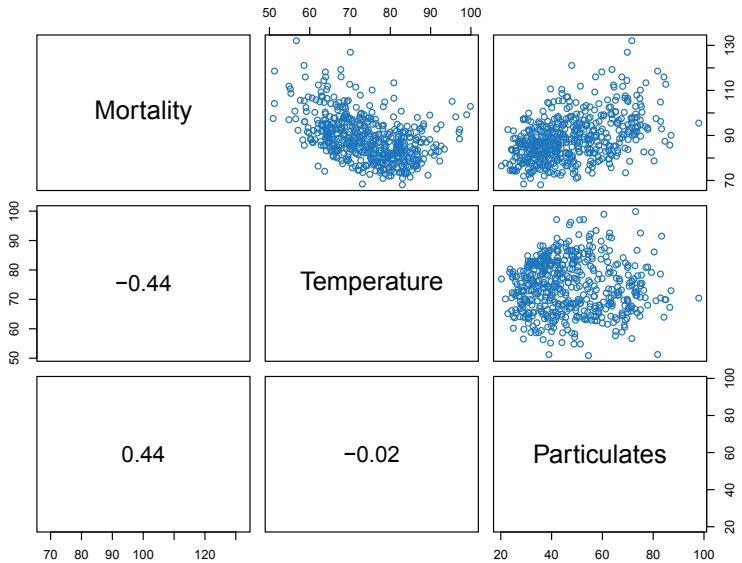


Figure 3.4 *Scatterplot matrix showing relations between mortality, temperature, and pollution. The lower panels display the correlations.*

models. They are

$$M_t = \beta_0 + \beta_1 t + w_t \tag{3.14}$$
$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T.) + w_t \tag{3.15}$$
$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T.) + \beta_3(T_t - T.)^2 + w_t \tag{3.16}$$
$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T.) + \beta_3(T_t - T.)^2 + \beta_4 P_t + w_t \tag{3.17}$$

where we adjust temperature for its mean, $T. = 74.26$, to avoid collinearity problems. For this range of temperatures, T_t and T_t^2 are highly collinear, but $T_t - T.$ and $(T_t - T.)^2$ are not. To see this, run this simple R code:

```
par(mfrow = 2:1)
plot(tempr, tempr^2)  # collinear
cor(tempr, tempr^2)
 [1] 0.9972099
temp = tempr - mean(tempr)
plot(temp, temp^2)    # not collinear
cor(temp, temp^2)
 [1] 0.07617904
```

Note that (3.14) is a trend only model, (3.15) adds a linear temperature term, (3.16) adds a curvilinear temperature term and (3.17) adds a pollution term. We summarize some of the statistics given for this particular case in Table 3.2.

We note that each model does substantially better than the one before it and

Table 3.2 *Summary Statistics for Mortality Models*

Model	k	SSE	df	MSE	R^2	AIC	BIC
(3.14)	2	40,020	506	79.0	.21	5.38	5.40
(3.15)	3	31,413	505	62.2	.38	5.14	5.17
(3.16)	4	27,985	504	55.5	.45	5.03	5.07
(3.17)	5	20,508	503	40.8	.60	4.72	4.77

that the model including temperature, temperature squared, and particulates does the best, accounting for some 60% of the variability and with the best value for AIC and BIC (because of the large sample size, AIC and AICc are nearly the same). Note that one can compare any two models using the residual sums of squares and (3.7). Hence, a model with only trend could be compared to the full model using $q = 4, r = 1, n = 508$, so

$$F_{3,503} = \frac{(40,020 - 20,508)/3}{20,508/503} = 160,$$

which exceeds $F_{3,503}(.001) = 5.51$. We obtain the best prediction model,

$$\begin{aligned}\widehat{M}_t = 2831.5 &- 1.396_{(.10)}\text{trend} - .472_{(.032)}(T_t - 74.26) \\ &+ .023_{(.003)}(T_t - 74.26)^2 + .255_{(.019)}P_t,\end{aligned}$$

for mortality, where the standard errors are given in parentheses.

As expected, a negative trend is present over time as well as a negative coefficient for adjusted temperature. Pollution weights positively and can be interpreted as the incremental contribution to daily deaths per unit of particulate pollution. It would still be essential to check the residuals $\hat{w}_t = M_t - \hat{M}_t$ for autocorrelation (of which there is a substantial amount), but we defer this question to Section 5.4 when we discuss regression with correlated errors.

Below is the R code to fit the final regression model (3.17), and compute the corresponding values of AIC and BIC.[3] Our definitions differ from R by terms that do not change from model to model. In the example, we show how to obtain (3.11) and (3.13) from the R output. Finally, the use of `na.action` in `lm()` is to retain the time series attributes for the residuals and fitted values.

```
temp   = tempr - mean(tempr) # center temperature
temp2  = temp^2
trend  = time(cmort)         # time is trend
fit    = lm(cmort~ trend + temp + temp2 + part, na.action=NULL)
summary(fit)                 # regression results
summary(aov(fit))            # ANOVA table   (compare to next line)
```

[3]The easiest way to extract AIC and BIC from an `lm()` run in R is to use the command `AIC()` or `BIC()`.

```
summary(aov(lm(cmort~cbind(trend, temp, temp2, part)))) # Table 3.1
num = length(cmort)          # sample size
AIC(fit)/num - log(2*pi)     # AIC
BIC(fit)/num - log(2*pi)     # BIC
```

Finally, in Figure 3.3 it appears that mortality may peak a few weeks after pollution peaks. In this case, we may want to include a lagged value of pollution into the model. This concept is explored further in Problem 3.2. ◊

It is possible to include lagged variables in time series regression models with some care. We will continue to discuss this type of problem throughout the text. To first address this problem, we consider a simple example of lagged regression.

Example 3.6. Regression with Lagged Variables
In Example 2.32, we discovered that the Southern Oscillation Index (SOI) measured at time $t-6$ months is associated with the Recruitment series at time t, indicating that the SOI leads the Recruitment series by six months. Although there is strong evidence that the relationship is NOT linear (this is discussed further in Example 3.13), *for demonstration purposes only*, we consider the following regression,

$$R_t = \beta_0 + \beta_1 S_{t-6} + w_t, \tag{3.18}$$

where R_t denotes Recruitment for month t and S_{t-6} denotes SOI six months prior. Assuming the w_t sequence is white, the fitted model is

$$\widehat{R}_t = 65.79 - 44.28_{(2.78)} S_{t-6} \tag{3.19}$$

with $\widehat{\sigma}_w = 22.5$ on 445 degrees of freedom. Of course, it is essential to check the model assumptions before making any conclusions, but we defer most of this discussion until later. We do, however, display a time series plot of the regression residuals in Figure 3.5, which clearly demonstrates a pattern and contradicts the assumption that w_t is white noise.

Performing lagged regression in R is a little difficult because the series must be aligned prior to running the regression. The easiest way to do this is to create an object (that we call `fish`) using `ts.intersect`, which aligns the lagged series.

```
fish = ts.intersect( rec, soiL6=lag(soi,-6) )
summary(fit1 <- lm(rec~ soiL6, data=fish, na.action=NULL))
  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept)   65.790      1.088   60.47   <2e-16
  soiL6        -44.283      2.781  -15.92   <2e-16
  ---
  Residual standard error: 22.5 on 445 degrees of freedom
  Multiple R-squared:  0.3629,    Adjusted R-squared:  0.3615
  F-statistic: 253.5 on 1 and 445 DF,  p-value: < 2.2e-16
tsplot(resid(fit1), col=4)    # residual time plot
```

The headache of aligning the lagged series can be avoided by using the R package `dynlm`. The setup is easier and the results are identical.

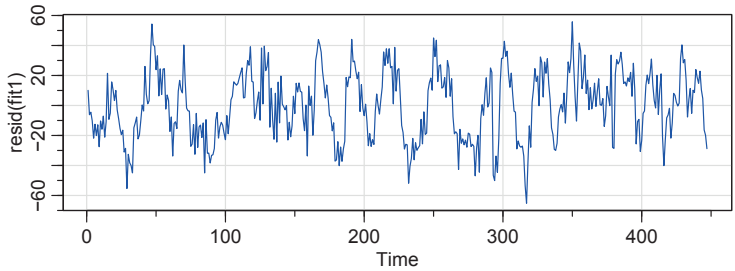


Figure 3.5 *Residual plot for Example 3.6.*

```
library(dynlm)
summary(fit2 <- dynlm(rec~ L(soi,6)))
```

◊

3.2 Exploratory Data Analysis

For time series, it is the dependence between the values of the series that is important to measure; we must, at least, be able to estimate autocorrelations with precision. It would be difficult to measure correlation between contiguous time points if the correlation were different for every pair of observations. Hence, it is crucial that a time series satisfies the conditions of stationarity stated in Definition 2.13 for at least some reasonable stretch of time. Often, this is not the case, and in this section we discuss some methods for coercing nonstationary data to stationarity.

A number of our examples came from clearly nonstationary series. The Johnson & Johnson series in Figure 1.1 has a mean that increases exponentially over time, and the increase in the magnitude of the fluctuations around this trend causes changes in the covariance function; the variance of the process, for example, clearly increases as one progresses over the length of the series. Also, the global temperature series shown in Figure 1.2 contain clear evidence of an increasing trend over time.

Perhaps the easiest form of nonstationarity to work with is the *trend stationary* model wherein the process has stationary behavior around a trend. We may write this type of model as

$$x_t = \mu_t + y_t \tag{3.20}$$

where x_t are the observations, μ_t denotes the trend, and y_t is a stationary process. Quite often, strong trend will obscure the behavior of the stationary process, y_t, as we shall see in numerous examples. Hence, there is some advantage to removing the trend as a first step in an exploratory analysis of such time series. The steps involved are to obtain a reasonable estimate of the trend component, say $\widehat{\mu}_t$, and then work with the residuals

$$\widehat{y}_t = x_t - \widehat{\mu}_t. \tag{3.21}$$

Consider the following example.

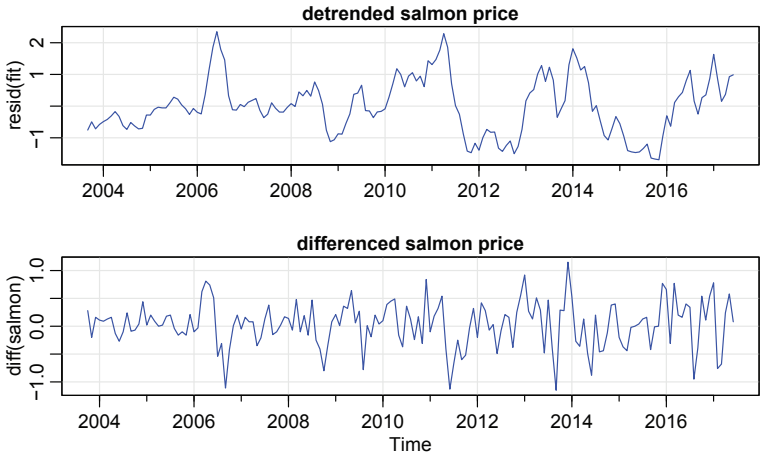


Figure 3.6 *Detrended (top) and differenced (bottom) salmon price series. The original data are shown in Figure 3.1.*

Example 3.7. Detrending a Commodity

Let x_t represent the salmon price data presented in Example 3.1. Here we suppose the model is of the form of (3.20),

$$x_t = \mu_t + y_t\,,$$

where, as we suggested in Example 3.1, a straight line might be useful for detrending the data; i.e.,

$$\mu_t = \beta_0 + \beta_1\, t\,,$$

where the time indices are the values in `time(salmon)`. In that example, we estimated the trend using ordinary least squares and found

$$\hat{\mu}_t = -503 + .25\, t.$$

Figure 3.1 (top) shows the data with the estimated trend line superimposed. To obtain the detrended series we simply subtract $\hat{\mu}_t$ from the observations, x_t, to obtain the detrended series[4]

$$\hat{y}_t = x_t + 503 - .25\, t.$$

The top graph of Figure 3.6 shows the detrended series. Figure 3.7 shows the ACF of the detrended data (top panel). ◊

In Example 1.10 we saw that a random walk might also be a good model for trend.

[4]Because the error term, y_t, is not assumed to be white noise, the reader may feel that weighted least squares is called for in this case. The problem is, we do not know the behavior of y_t and that is precisely what we are trying to assess at this stage. A notable result by Grenander and Rosenblatt (2008, Ch 7) is that under mild conditions on y_t, for polynomial regression or periodic regression, ordinary least squares is equivalent to weighted least squares with regard to efficiency for large samples.

That is, rather than modeling trend as fixed (as in Example 3.7), we might model trend as a stochastic component using the random walk with drift model,

$$\mu_t = \delta + \mu_{t-1} + w_t, \tag{3.22}$$

where w_t is white noise and is independent of y_t. If the appropriate model is (3.20), then *differencing* the data, x_t, yields a stationary process; that is,

$$\begin{aligned} x_t - x_{t-1} &= (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) \\ &= \delta + w_t + y_t - y_{t-1}. \end{aligned} \tag{3.23}$$

It is easy to show $z_t = y_t - y_{t-1}$ is stationary using Property 2.7. That is, because y_t is stationary,

$$\begin{aligned} \gamma_z(h) &= \text{cov}(z_{t+h}, z_t) = \text{cov}(y_{t+h} - y_{t+h-1}, y_t - y_{t-1}) \\ &= 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1) \end{aligned} \tag{3.24}$$

is independent of time; we leave it as an exercise (Problem 3.5) to show that $x_t - x_{t-1}$ in (3.23) is stationary.

One advantage of differencing over detrending to remove trend is that no parameters are estimated in the differencing operation. One disadvantage, however, is that differencing does not yield an estimate of the stationary process y_t as can be seen in (3.23). If an estimate of y_t is essential, then detrending may be more appropriate. This would be the case, for example, if we were interested in the business cycle of commodities. The salmon prices appear to have a 3- to 4-year business cycle, which is known as the Kitchin cycle (Kitchin, 1923) and is seen in many commodity series.

If the goal is to coerce the data to stationarity, then differencing may be more appropriate. Differencing is also a viable tool if the trend is fixed, as in Example 3.7. That is, e.g., if $\mu_t = \beta_0 + \beta_1 t$ in the model (3.20), differencing the data produces stationarity (see Problem 3.4):

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_1 + y_t - y_{t-1}.$$

Because differencing plays a central role in time series analysis, it receives its own notation. The first difference is denoted as

$$\nabla x_t = x_t - x_{t-1}. \tag{3.25}$$

As we have seen, the first difference eliminates a linear trend. A second difference, that is, the difference of (3.25), can eliminate a quadratic trend, and so on. In order to define higher differences, we need a variation in notation that we will use often in our discussion of ARIMA models in Chapter 5.

Definition 3.8. *We define the* **backshift operator** *by*

$$Bx_t = x_{t-1}$$

and extend it to powers $B^2x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$*, and so on. Thus,*

$$B^k x_t = x_{t-k}. \tag{3.26}$$

The idea of an inverse operator can also be given if we require $B^{-1}B = 1$, so that

$$x_t = B^{-1}Bx_t = B^{-1}x_{t-1}.$$

That is, B^{-1} is the *forward-shift operator*. In addition, it is clear that we may rewrite (3.25) as

$$\nabla x_t = (1 - B)x_t, \tag{3.27}$$

and we may extend the notion further. For example, the second difference becomes

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2} \tag{3.28}$$

by the linearity of the operator.

Definition 3.9. *Differences of order* d *are defined as*

$$\nabla^d = (1 - B)^d, \tag{3.29}$$

where we may expand the operator $(1 - B)^d$ *algebraically to evaluate for higher integer values of* d*. When* $d = 1$*, we drop it from the notation.*

The first difference (3.25) is an example of a *linear filter* applied to eliminate a trend. Other filters, formed by averaging values near x_t, can produce adjusted series that eliminate other kinds of unwanted fluctuations, as in Chapter 6. The differencing technique is an important component of the ARIMA model discussed in Chapter 5.

Example 3.10. Differencing a Commodity

The first difference of the salmon prices series, also shown in Figure 3.6, produces different results than removing trend by detrending via regression. For example, the Kitchin business cycle we observed in the detrended series is not obvious in the differenced series (although it is still there, which can be verified using Chapter 7 techniques).

The ACF of the differenced series is also shown in Figure 3.7. In this case, the difference series exhibits a strong annual cycle that was not evident in the original or detrended data. The R code to reproduce Figure 3.6 and Figure 3.7 is as follows.

```
fit = lm(salmon~time(salmon), na.action=NULL) # the regression
par(mfrow=c(2,1))    # plot transformed data
tsplot(resid(fit), main="detrended salmon price")
tsplot(diff(salmon), main="differenced salmon price")
par(mfrow=c(2,1))    # plot their ACFs
acf1(resid(fit), 48, main="detrended salmon price")
acf1(diff(salmon), 48, main="differenced salmon price")
```

♢

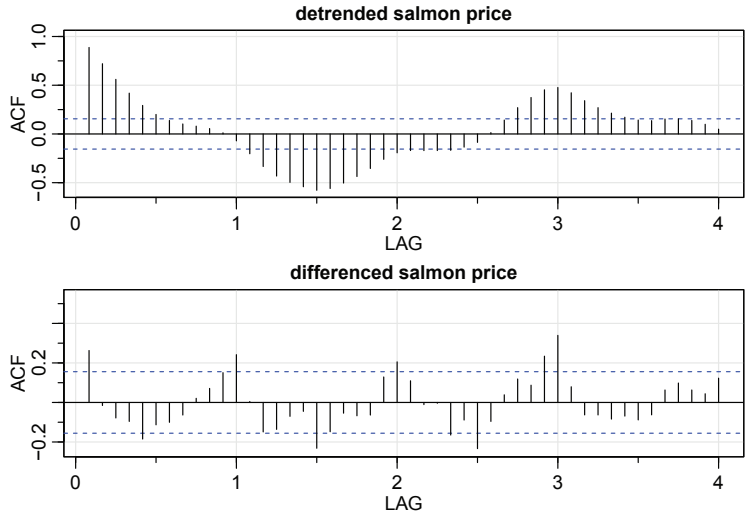


Figure 3.7 *Sample ACFs of the detrended (top) and of the differenced (bottom) salmon price series.*

Example 3.11. Differencing Global Temperature

The global temperature series shown in Figure 1.2 appears to behave more as a random walk than a trend stationary series. Hence, rather than detrend the data, it would be more appropriate to use differencing to coerce it into stationarity. The detrended data are shown in Figure 3.8 along with the corresponding sample ACF. In this case it appears that the differenced process shows minimal autocorrelation at lag 1, which may imply the global temperature series is nearly a random walk with drift.

It is interesting to note that if the series is a random walk with drift, the mean of the differenced series, which is an estimate of the drift, is about .014, or an increase of about one and a half degree centigrade per 100 years. If however, we restrict attention to the temperatures after 1980 when global temperature increase is evident (see Hansen and Lebedeff, 1987), the drift increases by more than twofold. The R code to reproduce Figure 3.8 is as follows.

```
par(mfrow=c(2,1))
tsplot(diff(gtemp_land), col=4, main="differenced global temperature")
mean(diff(gtemp_land))     # drift since 1880
 [1] 0.0143
acf1(diff(gtemp_land))
mean(window(diff(gtemp_land), start=1980)) # drift since 1980
 [1] 0.0329
```

◊

Sometimes heteroscedasticity is seen in time series data. A particularly useful transformation in this case is

$$y_t = \log x_t, \tag{3.30}$$

which tends to suppress larger fluctuations that occur over portions of the series where

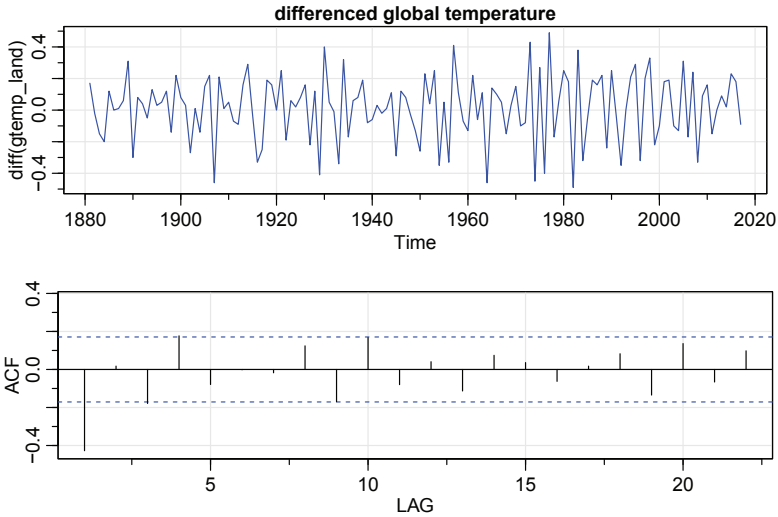


Figure 3.8 *Differenced global temperature series and its sample ACF.*

the underlying values are larger. Other possibilities are *power transformations* in the Box–Cox family of the form

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log x_t & \lambda = 0. \end{cases} \tag{3.31}$$

Methods for choosing the power λ are available (see Johnson and Wichern, 2002, §4.7) but we do not pursue them here. Often, transformations are also used to improve the approximation to normality or to improve linearity in predicting the value of one series from another.

Example 3.12. Paleoclimatic Glacial Varves

Melting glaciers deposit yearly layers of sand and silt during the spring melting seasons, which can be reconstructed yearly over a period ranging from the time deglaciation began in New England (about 12,600 years ago) to the time it ended (about 6,000 years ago). Such sedimentary deposits, called *varves*, can be used as proxies for paleoclimatic parameters, such as temperature, because, in a warm year, more sand and silt are deposited from the receding glacier. The top of Figure 3.9 shows the thicknesses of the yearly varves collected from one location in Massachusetts for 634 years, beginning 11,834 years ago. For further information, see Shumway and Verosub (1992).

Because the variation in thicknesses increases in proportion to the amount deposited, a logarithmic transformation could remove the nonstationarity observable in the variance as a function of time. Figure 3.9 shows the original and the logged transformed varves, and it is clear that this improvement has occurred. Also plotted are the corresponding normal Q-Q plots. Recall that these plots are of the quantiles

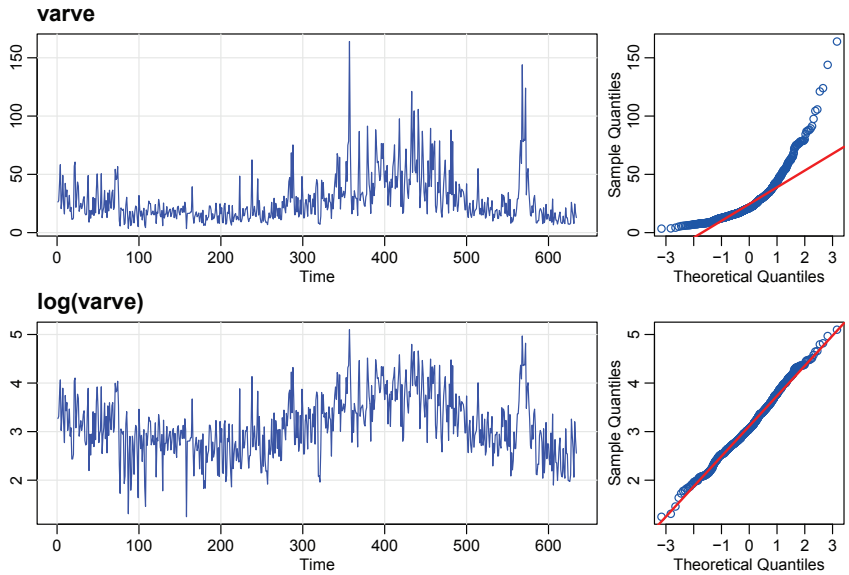


Figure 3.9 *Glacial varve thicknesses (top) from Massachusetts for* $n = 634$ *years compared with* log *transformed thicknesses (bottom). The plots on the right-side are corresponding normal Q-Q plots.*

of the data against the theoretical quantiles of the normal distribution. Normal data should fall approximately on the exhibited line of equality. In this case, we can argue that the approximation to normality is improved by the log transformation.

Figure 3.9 was generated in R as follows:

```
layout(matrix(1:4,2), widths=c(2.5,1))
par(mgp=c(1.6,.6,0), mar=c(2,2,.5,0)+.5)
tsplot(varve, main="", ylab="", col=4, margin=0)
mtext("varve", side=3, line=.5, cex=1.2, font=2, adj=0)
tsplot(log(varve), main="", ylab="", col=4, margin=0)
mtext("log(varve)", side=3, line=.5, cex=1.2, font=2, adj=0)
qqnorm(varve, main="", col=4); qqline(varve, col=2, lwd=2)
qqnorm(log(varve), main="", col=4); qqline(log(varve), col=2, lwd=2)
```

◊

Next, we consider another preliminary data processing technique that is used for the purpose of visualizing the relations between series at different lags, namely the *lagplot*. When using the ACF, we are measuring the linear relation between lagged values of a time series. The restriction of this idea to linear predictability, however, may mask possible nonlinear relationships between future values, x_{t+h}, and current values, x_t. This idea extends to two series where one may be interested in examining lagplots of y_t versus x_{t-h}.

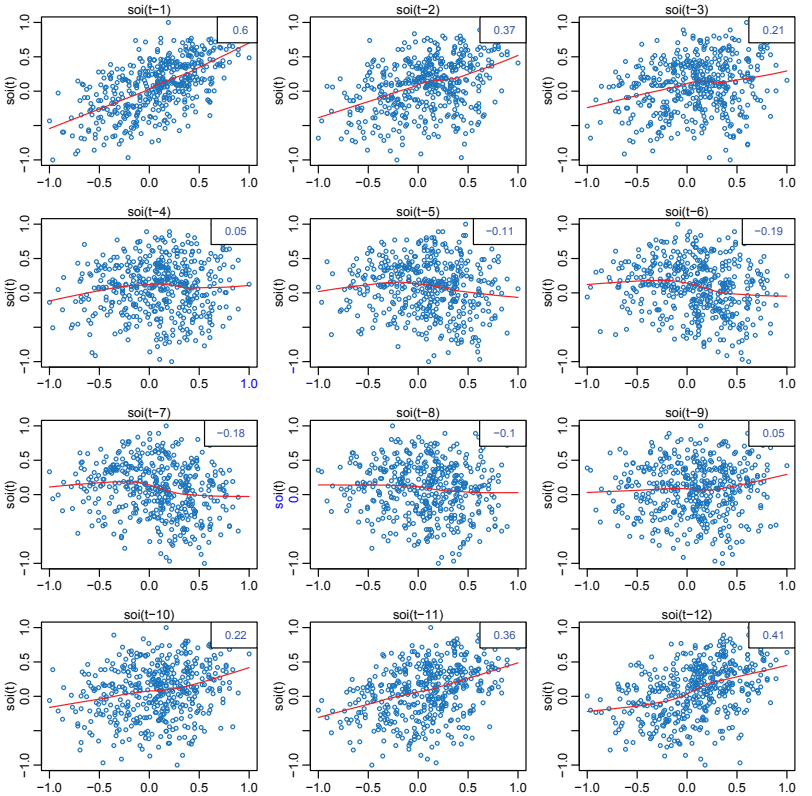


Figure 3.10 *Lagplot relating current SOI values, S_t, to past SOI values, S_{t-h}, at lags $h = 1, 2, ..., 12$. The values in the upper right corner are the sample autocorrelations and the lines are a lowess fit.*

Example 3.13. Lagplots: SOI and Recruitment

Figure 3.10 displays a lagplot of the SOI, S_t, on the vertical axis plotted against S_{t-h} on the horizontal axis. The sample autocorrelations are displayed in the upper right-hand corner and superimposed on the lagplots are locally weighted scatterplot smoothing (lowess) lines that can be used to help discover any nonlinearities. We discuss smoothing in the next section, but for now, think of lowess as a method for fitting local regression.

In Figure 3.10, we notice that the local fits are approximately linear so that the sample autocorrelations are meaningful. Also, we see strong positive linear relations at lags $h = 1, 2, 11, 12$, that is, between S_t and $S_{t-1}, S_{t-2}, S_{t-11}, S_{t-12}$, and a negative linear relation at lags $h = 6, 7$.

Similarly, we might want to look at values of one series, say Recruitment, denoted R_t plotted against another series at various lags, say the SOI, S_{t-h}, to look for possible nonlinear relations between the two series. Because, for example, we might wish to

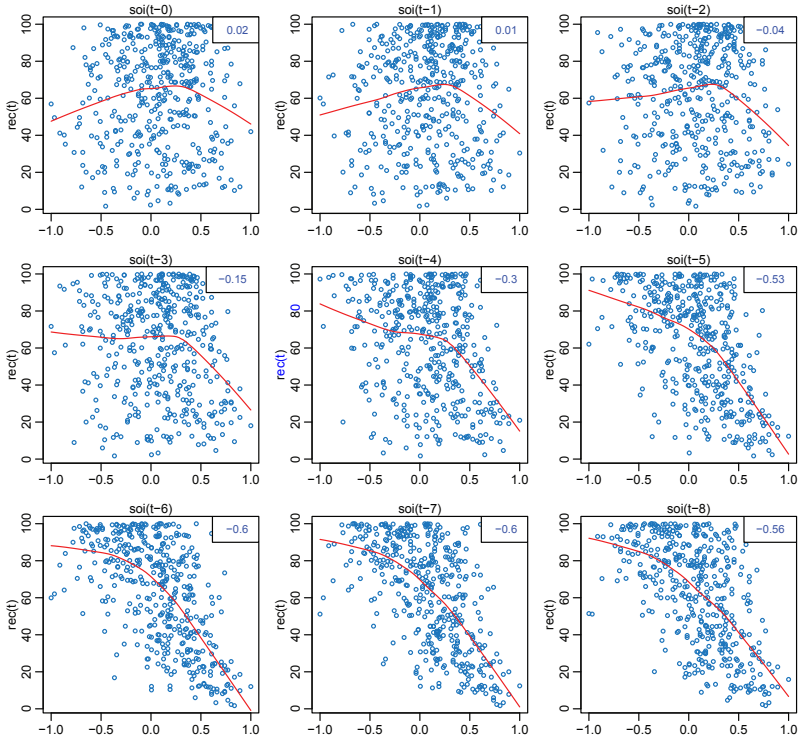


Figure 3.11 *Lagplot of the Recruitment series, R_t, on the vertical axis plotted against the SOI series, S_{t-h}, on the horizontal axis at lags $h = 0, 1, \ldots, 8$. The values in the upper right corner are the sample cross-correlations and the lines are a lowess fit.*

predict the Recruitment series, R_t, from current or past values of the SOI series, S_{t-h}, for $h = 0, 1, 2, \ldots$ it would be worthwhile to examine the scatterplot matrix. Figure 3.11 shows the lagplot of the Recruitment series R_t on the vertical axis plotted against the SOI index S_{t-h} on the horizontal axis. In addition, the figure exhibits the sample cross-correlations as well as lowess fits.

Figure 3.11 shows a fairly strong nonlinear relationship between Recruitment, R_t, and the SOI series at $S_{t-5}, S_{t-6}, S_{t-7}, S_{t-8}$, indicating the SOI series tends to lead the Recruitment series and the coefficients are negative, implying that increases in the SOI lead to decreases in the Recruitment. The nonlinearity observed in the lagplots (with the help of the superimposed lowess fits) indicate that the behavior between Recruitment and the SOI is different for positive values of SOI than for negative values of SOI.

The R code for this example is

```
lag1.plot(soi, 12, col="dodgerblue3")      #  Figure 3.10
lag2.plot(soi, rec, 8, col="dodgerblue3")  #  Figure 3.11
```

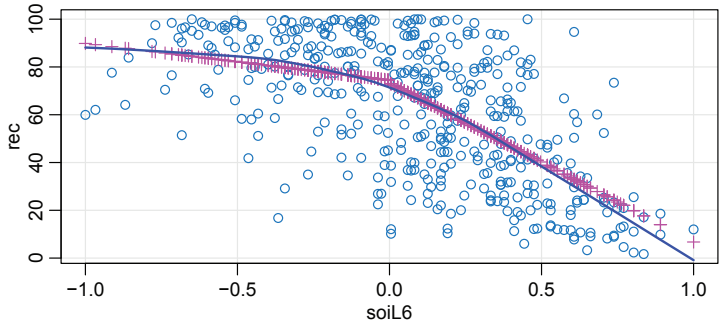


Figure 3.12 *Display for Example 3.14: Plot of Recruitment (R_t) vs. SOI lagged 6 months (S_{t-6}) with the fitted values of the regression as points (+) and a lowess fit (—).*

Example 3.14. Regression with Lagged Variables (cont.)
In Example 3.6 we regressed Recruitment on lagged SOI,

$$R_t = \beta_0 + \beta_1 S_{t-6} + w_t.$$

However, in Example 3.13, we saw that the relationship is nonlinear and different when SOI is positive or negative. In this case, we may consider adding a dummy variable to account for this change. In particular, 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 D_t is a dummy variable that is 0 if $S_t < 0$ and 1 otherwise. This means that

$$R_t = \begin{cases} \beta_0 + \beta_1 S_{t-6} + w_t & \text{if } S_{t-6} < 0, \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) S_{t-6} + w_t & \text{if } S_{t-6} \geq 0. \end{cases}$$

The result of the fit is given in the R code below. We have loaded `zoo` to ease the pain of working with lagged variables in R. Figure 3.12 shows R_t vs S_{t-6} with the fitted values of the regression and a lowess fit superimposed. The piecewise regression fit is similar to the lowess fit, but we note that the residuals are not white noise. This is followed up in Problem 5.16.

```
library(zoo)   # zoo allows easy use of the variable names
dummy = ifelse(soi<0, 0, 1)
fish = as.zoo(ts.intersect(rec, soiL6=lag(soi,-6), dL6=lag(dummy,-6)))
summary(fit <- lm(rec~ soiL6*dL6, data=fish, na.action=NULL))
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)   74.479      2.865  25.998  < 2e-16
 soiL6        -15.358      7.401  -2.075   0.0386
 dL6           -1.139      3.711  -0.307   0.7590
 soiL6:dL6    -51.244      9.523  -5.381  1.2e-07
```

```
 ---
 Residual standard error: 21.84 on 443 degrees of freedom
 F-statistic: 99.43 on 3 and 443 DF,  p-value: < 2.2e-16
plot(fish$soiL6, fish$rec, panel.first=Grid(), col="dodgerblue3")
points(fish$soiL6, fitted(fit), pch=3, col=6)
lines(lowess(fish$soiL6, fish$rec), col=4, lwd=2)
tsplot(resid(fit))     # not shown, but looks like Figure 3.5
acf1(resid(fit))       # and obviously not noise
```

◊

As a final exploratory tool, we discuss assessing periodic behavior in time series data using regression analysis; this material may be thought of as an introduction to *spectral analysis*, which we discuss in detail in Chapter 6. In Example 1.11, we briefly discussed the problem of identifying cyclic or periodic signals in time series. A number of the time series we have seen so far exhibit periodic behavior. For example, the data from the pollution study example shown in Figure 3.2 exhibit strong yearly cycles. Also, the Johnson & Johnson data shown in Figure 1.1 make one cycle every year (four quarters) on top of an increasing trend and the speech data in Figure 1.2 is highly repetitive. The monthly SOI and Recruitment series in Figure 1.7 show strong yearly cycles, but hidden in the series are clues to the El Niño cycle.

Example 3.15. Using Regression to Discover a Signal in Noise
In Example 1.11, we generated $n = 500$ observations from the model

$$x_t = A\cos(2\pi\omega t + \phi) + w_t, \tag{3.32}$$

where $\omega = 1/50$, $A = 2$, $\phi = .6\pi$, and $\sigma_w = 5$; the data are shown on the bottom panel of Figure 1.11. At this point we assume the frequency of oscillation $\omega = 1/50$ is known, but A and ϕ are unknown parameters. In this case the parameters appear in (3.32) in a nonlinear way, so we use a trigonometric identity (see Section C.5) and write

$$A\cos(2\pi\omega t + \phi) = \beta_1\cos(2\pi\omega t) + \beta_2\sin(2\pi\omega t),$$

where $\beta_1 = A\cos(\phi)$ and $\beta_2 = -A\sin(\phi)$.

Now the model (3.32) can be written in the usual linear regression form given by (no intercept term is needed here)

$$x_t = \beta_1\cos(2\pi t/50) + \beta_2\sin(2\pi t/50) + w_t. \tag{3.33}$$

Using linear regression, we find $\hat{\beta}_1 = -.74_{(.33)}$, $\hat{\beta}_2 = -1.99_{(.33)}$ with $\hat{\sigma}_w = 5.18$; the values in parentheses are the standard errors. We note the actual values of the coefficients for this example are $\beta_1 = 2\cos(.6\pi) = -.62$, and $\beta_2 = -2\sin(.6\pi) = -1.90$. It is clear that we are able to detect the signal in the noise using regression, even though the signal-to-noise ratio is small. The top of Figure 3.13 shows the data generated by (3.32); it is hard to discern the signal and the data look like noise. However, the bottom of the figure shows the same data, but with the fitted line superimposed. It is now easy to see the signal through the noise.

To reproduce the analysis and Figure 3.13 in R, use the following:

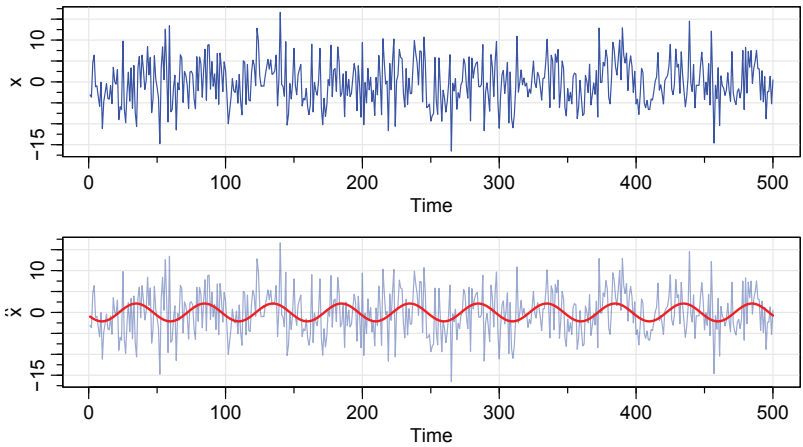


Figure 3.13 *Data generated by* (3.32) *[top] and the fitted line superimposed on the data [bottom].*

```
set.seed(90210)                 # so you can reproduce these results
x  = 2*cos(2*pi*1:500/50 + .6*pi) + rnorm(500,0,5)
z1 = cos(2*pi*1:500/50)
z2 = sin(2*pi*1:500/50)
summary(fit <- lm(x~ 0 + z1 + z2))  # zero to exclude the intercept
  Coefficients:
     Estimate Std. Error t value Pr(>|t|)
  z1  -0.7442     0.3274  -2.273   0.0235
  z2  -1.9949     0.3274  -6.093 2.23e-09
  Residual standard error: 5.177 on 498 degrees of freedom
par(mfrow=c(2,1))
tsplot(x, col=4)
tsplot(x, ylab=expression(hat(x)), col=rgb(0,0,1,.5))
lines(fitted(fit), col=2, lwd=2)
```

◊

3.3 Smoothing Time Series

In Example 1.8, we introduced the concept of smoothing a time series using a moving average. This method is useful for discovering certain traits in a time series, such as long-term trend and seasonal components (see Section 6.3 for details). In particular, if x_t represents the observations, then

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}, \tag{3.34}$$

where $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$ is a symmetric moving average.

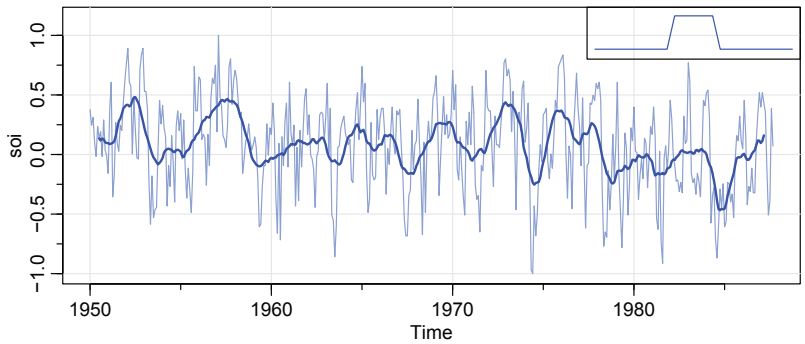


Figure 3.14 *The SOI series smoothed using* (3.34) *with* $k = 6$ *(and half-weights at the ends). The insert shows the shape of the moving average ("boxcar") kernel [not drawn to scale] described in* (3.36)*.*

Example 3.16. Moving Average Smoother

For example, Figure 3.14 shows the monthly SOI series discussed in Example 1.4 smoothed using (3.34) with $k = 6$ and weights $a_0 = a_{\pm1} = \cdots = a_{\pm5} = 1/12$, and $a_{\pm6} = 1/24$. This particular method removes (filters out) the obvious annual temperature cycle and helps emphasize the El Niño cycle. The reason half-weights are used at the ends is so the same month does not get included in the average twice. For example, if we center on a July ($j = 0$), then January ($j = -6$) of that year and January ($j = 6$) of the next year will be included in the smoother. Consequently, each January gets a half-weight, and so on.

To reproduce Figure 3.14 in R:

```
w = c(.5, rep(1,11), .5)/12
soif = filter(soi, sides=2, filter=w)
tsplot(soi, col=rgb(.5, .6, .85, .9), ylim=c(-1, 1.15))
lines(soif, lwd=2, col=4)
# insert
par(fig = c(.65, 1, .75, 1), new = TRUE)
w1 = c(rep(0,20), w, rep(0,20))
plot(w1, type="l", ylim = c(-.02,.1), xaxt="n", yaxt="n", ann=FALSE)
```

◊

Although the moving average smoother does a good job in highlighting the El Niño effect, it might be considered too choppy. We can obtain a smoother fit using the normal distribution for the weights, instead of boxcar-type weights of (3.34).

Example 3.17. Kernel Smoothing

Kernel smoothing is a moving average smoother that uses a weight function, or kernel, to average the observations. Figure 3.15 shows kernel smoothing of the SOI series, where m_t is now

$$m_t = \sum_{i=1}^{n} w_i(t) x_{t_i}\,, \tag{3.35}$$

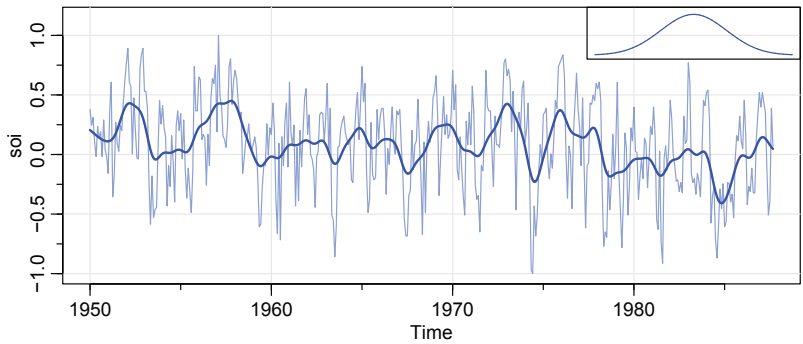


Figure 3.15 *Kernel smoother of the SOI. The insert shows the shape of the normal kernel [not drawn to scale].*

where

$$w_i(t) = K\left(\tfrac{t-t_i}{b}\right) \Big/ \sum_{k=1}^{n} K\left(\tfrac{t-t_k}{b}\right) \tag{3.36}$$

are the weights and $K(\cdot)$ is a kernel function. In this example, and typically, the normal kernel, $K(z) = \exp(-z^2/2)$, is used.

To implement this in R, we use the `ksmooth` function where a bandwidth can be chosen. Think of b as standard deviation, and the bigger the bandwidth, the smoother the result. In our case, we are smoothing over time, which is of the form $t/12$ for `soi`. In Figure 3.15, we used the value of $b = 1$ to correspond to approximately smoothing over about a year The R code for this example is

```
tsplot(soi, col=rgb(0.5, 0.6, 0.85, .9), ylim=c(-1, 1.15))
lines(ksmooth(time(soi), soi, "normal", bandwidth=1), lwd=2, col=4)
# insert
par(fig = c(.65, 1, .75, 1), new = TRUE)
curve(dnorm(x), -3, 3, xaxt="n", yaxt="n", ann=FALSE, col=4)
```

We note that if the unit of time for SOI were months, then an equivalent smoother would use a bandwidth of 12:

```
SOI = ts(soi, freq=1)
tsplot(SOI)   # the time scale matters (not shown)
lines(ksmooth(time(SOI), SOI, "normal", bandwidth=12), lwd=2, col=4)
```

◊

Example 3.18. Lowess

Another approach to smoothing is based on k-nearest neighbor regression, wherein, for $k < n$, one uses only the data $\{x_{t-k/2}, \ldots, x_t, \ldots, x_{t+k/2}\}$ to predict x_t via regression, and then sets $m_t = \hat{x}_t$.

Lowess is a method of smoothing that is rather complex, but the basic idea is close to nearest neighbor regression. Figure 3.16 shows smoothing of SOI using the R function `lowess` (see Cleveland, 1979). First, a certain proportion of nearest neighbors to x_t are included in a weighting scheme; values closer to x_t in time get more weight. Then, a robust weighted regression is used to predict x_t and obtain

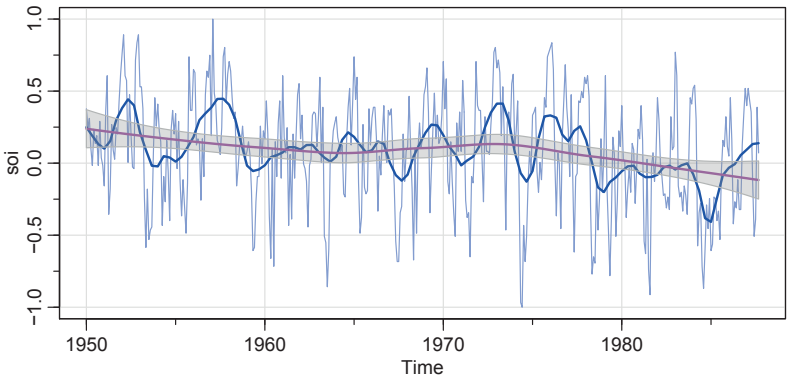


Figure 3.16 *Locally weighted scatterplot smoothers of the SOI series. The El Niño cycle is estimated using lowess and the trend with confidence intervals is estimated using loess.*

the smoothed values m_t. The larger the fraction of nearest neighbors included, the smoother the fit will be. In Figure 3.16, one smoother uses 5% of the data to obtain an estimate of the El Niño cycle of the data. In addition, a (negative) trend in SOI would indicate the long-term warming of the Pacific Ocean. To investigate this, we used the R function `loess` with the default smoother span of `f=2/3` of the data. The script `loess` is similar to `lowess`. A major difference for us is that the former strips the time series attributes whereas the latter does not, but the `loess` script allows the calculation of confidence intervals. Figure 3.16 can be reproduced in R as follows. We have commented out the trend estimate using `lowess`.

```
tsplot(soi, col=rgb(0.5, 0.6, 0.85, .9))
lines(lowess(soi, f=.05), lwd=2, col=4)     # El Niño cycle
# lines(lowess(soi), lty=2, lwd=2, col=2)  # trend (with default span)
##-- trend with CIs using loess --##
lo = predict(loess(soi~ time(soi)), se=TRUE)
trnd = ts(lo$fit, start=1950, freq=12)      # put back ts attributes
lines(trnd, col=6, lwd=2)
 L = trnd - qt(0.975, lo$df)*lo$se
 U = trnd + qt(0.975, lo$df)*lo$se
 xx = c(time(soi), rev(time(soi)))
 yy = c(L, rev(U))
polygon(xx, yy, border=8, col=gray(.6, alpha=.4) )
```

◊

Example 3.19. Smoothing One Series as a Function of Another

Smoothing techniques can also be applied to smoothing a time series as a function of another time series. In Example 3.5, we discovered a nonlinear relationship between mortality and temperature. Figure 3.17 shows a scatterplot of mortality, M_t, and temperature, T_t, along with M_t smoothed as a function of T_t using lowess. Note that

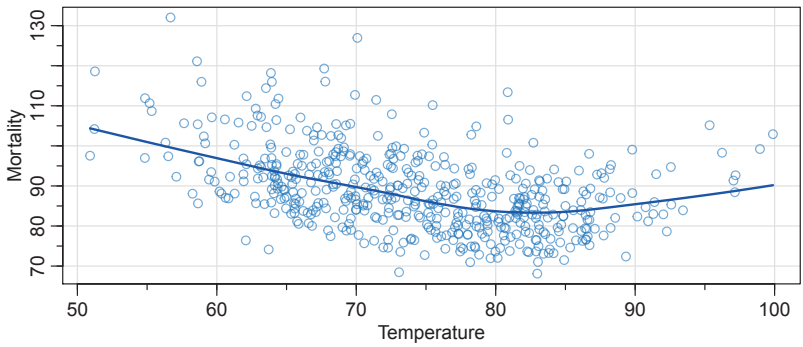


Figure 3.17 *Smooth of mortality as a function of temperature using lowess.*

mortality increases at extreme temperatures. The minimum mortality rate seems to occur at approximately 83° F. Figure 3.17 can be reproduced in R as follows.

```
plot(tempr, cmort, xlab="Temperature", ylab="Mortality",
           col="dodgerblue3", panel.first=Grid())
lines(lowess(tempr,cmort), col=4, lwd=2)
```

◊

Example 3.20. Classical Structural Modeling
A classical approach to time series analysis is to decompose data into components labeled trend (T_t), seasonal (S_t), irregular or noise (N_t). If we let x_t denote the data, we can then sometimes write

$$x_t = T_t + S_t + N_t\,.$$

Of course, not all time series data fit into such a paradigm and the decomposition may not be unique. Sometimes an additional cyclic component, say C_t, such as a business cycle is added to the model.

Figure 3.18 shows the result of the decomposition using loess on the quarterly occupancy rate of Hawaiian hotels from 2002 to 2016. R provides a few scripts to fit the decomposition. The script `decompose` uses moving averages as in Example 3.16. Another script, `stl`, uses loess to obtain each component and is similar to the approach used in Example 3.18. To use `stl`, the seasonal smoothing method must be specified. That is, specify either the character string `"periodic"` or the span of the loess window for seasonal extraction. The span should be odd and at least 7 (there is no default). By using a seasonal window, we are allowing $S_t \approx S_{t-4}$ rather than $S_t = S_{t-4}$, which is forced by specifying a periodic seasonal component.

Note that in Figure 3.18, the seasonal component is very regular showing a 2% to 4% gain in the first and third quarters, while showing a 2% to 4% loss in the second and fourth quarters. The trend component is perhaps more like a business cycle than what may be considered a trend. As previously implied, the components are not well defined and the decomposition is not unique; one person's trend may be another person's business cycle. The basic R code for this example is:

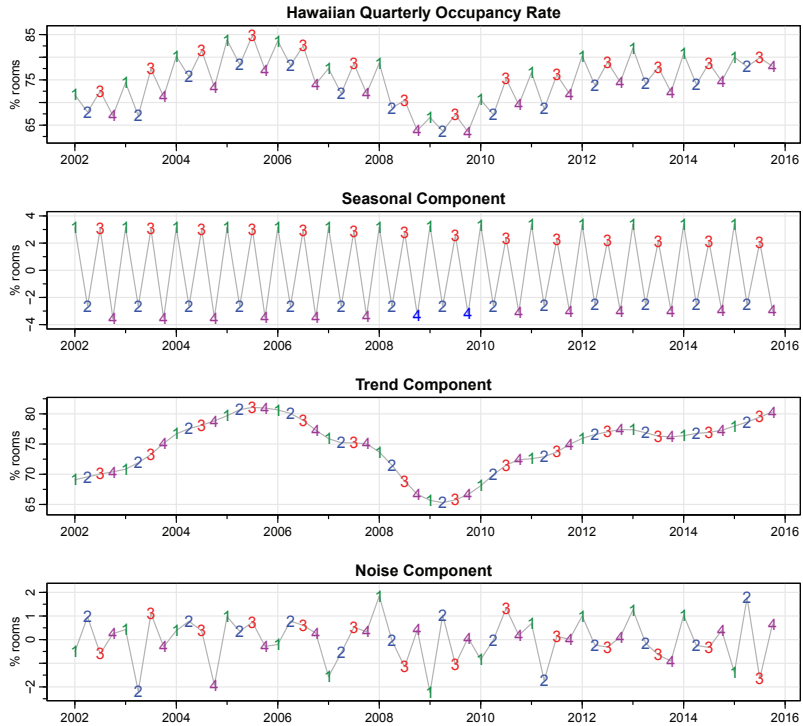


Figure 3.18 *Structural model of the Hawaiian quarterly occupancy rate.*

```
x = window(hor, start=2002)
plot(decompose(x))             # not shown
plot(stl(x, s.window="per")) # seasons are periodic - not shown
plot(stl(x, s.window=15))
```

However, a figure similar to Figure 3.18 can be generated as follows:

```
culer = c("cyan4", 4, 2, 6)
par(mfrow = c(4,1), cex.main=1)
x = window(hor, start=2002)
out = stl(x, s.window=15)$time.series
tsplot(x, main="Hawaiian Occupancy Rate", ylab="% rooms", col=gray(.7))
 text(x, labels=1:4, col=culer, cex=1.25)
tsplot(out[,1], main="Seasonal", ylab="% rooms",col=gray(.7))
 text(out[,1], labels=1:4, col=culer, cex=1.25)
tsplot(out[,2], main="Trend", ylab="% rooms", col=gray(.7))
 text(out[,2], labels=1:4, col=culer, cex=1.25)
tsplot(out[,3], main="Noise", ylab="% rooms", col=gray(.7))
 text(out[,3], labels=1:4, col=culer, cex=1.25)
```

STAT 626: Outline of Lecture 14
ARMA Models (Chapter 4)

1. **Autoregressive Models** (§4.1)

2. **Estimation** (§4.3)

   Yule-Walker Equations.

3. **Correlation Functions (§4.2)**

   ACF and PACF

4. **Forecasting/Prediction (§4.4)**

Review of Stationarity, Overview of TS Models (Chapter 4)

1. **Autoregressive Models of order p or AR(p) Models:**

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t, \quad \phi_p \neq 0.$$

QUESTION: Is a time series $\{x_t\}$ defined via an AR(p) model always stationary? If so, what is its autocovariance function?
To get a feel for the answer consider the AR(1):

$$x_t = \phi x_{t-1} + w_t,$$

what happens when $\phi = 1$?

See Examples 1.9, 1.10, 2.20 and Problem 2.4 for more details on AR models.

2. **Linear Processes:** $x_t = \mu + \sum_{j=-\infty}^{+\infty} \psi_j w_{t-j}$ is stationary with the *autocovariance function*

$$\gamma(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j.$$

3. **MA(q) Models**: $x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q}, \quad \theta_q \neq 0$, is stationary, its autocovariance is zero at lags greater than q.

4. **The Backshift Operator B:** $Bx_t = x_{t-1}$.

5. **MA(q) and B:**

$$x_t = w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q} = (1 + \theta_1 B + \ldots + \theta_q B^q) w_t = \theta(B) w_t.$$

6. **AR(p) and B:**

$$x_t - \phi_1 x_{t-1} - \ldots - \phi_p x_{t-p} = w_t, \quad (1 - \phi_1 B - \ldots - \phi_p B^p) x_t = \phi(B) x_t = w_t.$$

7. **The ROOTS of the polynomial equation**

$$\phi(B) = 0,$$

hold the key to the question of stationarity of the solutions of AR models.

Chapter 4

ARMA Models

4.1 Autoregressive Moving Average Models

Linear regression models are often unsatisfactory for explaining all of the interesting dynamics of a time series. Instead, the introduction of correlation through lagged relationships leads to autoregressive (AR) and moving average (MA) models. These models are often combined to form autoregressive moving average (ARMA) models.

Autoregressive models are an obvious extension of linear regression models. An *autoregressive model* of order p, abbreviated AR(p), is of the form

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{4.1}$$

where x_t is stationary and w_t is white noise. We note that (4.1) is similar to the regression model of Section 3.1, and hence the term auto (or self) regression. Some technical difficulties develop from applying that model because the regressors, $x_{t-1}, \ldots, x_{t-p}$, are random components, whereas in regression, the regressors are assumed to be fixed. For example, we will see that restrictions must be put on the AR parameters, as opposed to linear regression where there are no parameter restrictions.

Example 4.1. The AR(1) Model and Causality
Consider the first-order, zero-mean AR(1) model,

$$x_t = \phi x_{t-1} + w_t.$$

Because x_t must be stationary, we can rule out the case $\phi = 1$ because this would make x_t a random walk, which we know is not stationary. Similarly, we can rule out $\phi = -1$. In other words, the models

$$x_t = x_{t-1} + w_t\,, \quad \text{and} \quad x_t = -x_{t-1} + w_t\,,$$

are *not* AR models because they are not stationary.

As we saw in Example 2.20, if x_t is stationary, then

$$\text{var}(x_t) = \phi^2 \text{var}(x_{t-1}) + \text{var}(w_t)\,,$$

which, because $\text{var}(x_{t-1}) = \text{var}(x_t)$, implies

$$\text{var}(x_t) = \gamma(0) = \sigma_w^2 \frac{1}{(1-\phi^2)}\,.$$

Thus, we must have $|\phi| < 1$ for the process to have a positive (finite) variance. Similarly, in Example 2.20, we showed that ϕ is the correlation between x_t and x_{t-1}.

Provided that $|\phi| < 1$ we can represent an AR(1) model as a linear process given by

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \tag{4.2}$$

Representation (4.2) is called the *causal solution* of the model (see Section D.2 for details). The term causal refers to the fact that x_t does not depend on the future. In fact, by simple substitution,

$$\underbrace{\sum_{j=0}^{\infty} \phi^j w_{t-j}}_{x_t} = \phi \underbrace{\left(\sum_{k=0}^{\infty} \phi^k w_{t-1-k}\right)}_{x_{t-1}} + w_t\,.$$

As a check, the right-hand side is $w_t + \phi w_{t-1}\,[k=0] + \phi^2 w_{t-2}\,[k=1] + \cdots$. Using (4.2), it is easy to see that the AR(1) process is stationary with mean

$$E(x_t) = \sum_{j=0}^{\infty} \phi^j E(w_{t-j}) = 0,$$

and autocovariance function ($h \geq 0$),

$$\begin{aligned}
\gamma(h) &= \operatorname{cov}(x_{t+h}, x_t) = \operatorname{cov}\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j}, \sum_{k=0}^{\infty} \phi^k w_{t-k}\right) \\
&= \operatorname{cov}\left[w_{t+h} + \cdots + \phi^h w_t + \phi^{h+1} w_{t-1} + \cdots,\ \phi^0 w_t + \phi w_{t-1} + \cdots\right] \\
&= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{h+j}\phi^j = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2 \phi^h}{1-\phi^2}\,.
\end{aligned} \tag{4.3}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only exhibit the autocovariance function for $h \geq 0$. From (4.3), the ACF of an AR(1) is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h, \quad h \geq 0. \tag{4.4}$$

In addition, from the causal form (4.2) we see that, as required in Example 2.20, x_{t-1} and w_t are uncorrelated because $x_{t-1} = \sum_{j=0}^{\infty} \phi^j w_{t-1-j}$ is a linear filter of past shocks, $w_{t-1}, w_{t-2}, \ldots$, which are uncorrelated with w_t, the present shock. Also, the causal form of the model allows us to easily see that if we replace x_t by $x_t - \mu$, then

$$x_t = \mu + \sum_{j=0}^{\infty} \phi^j w_{t-j}\,,$$

so that the mean function is now $E(x_t) = \mu$. ◊

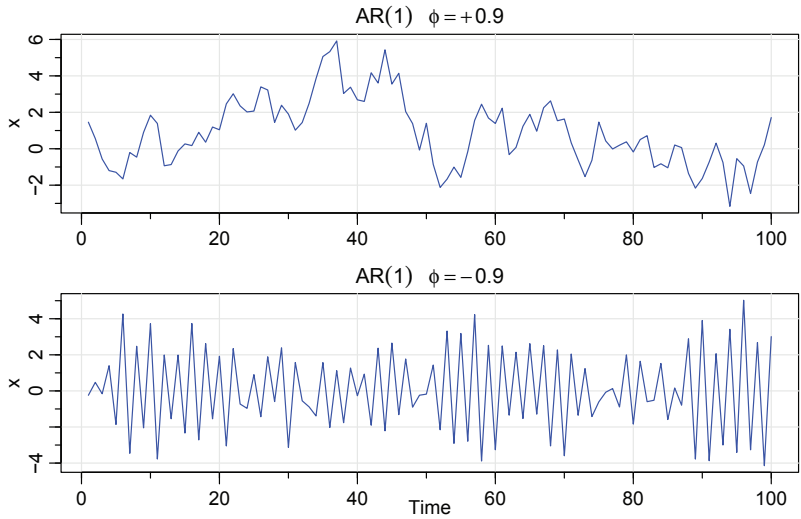


Figure 4.1 *Simulated AR(1) models: $\phi = .9$ (top); $\phi = -.9$ (bottom).*

Example 4.2. The Sample Path of an AR(1) Process

Figure 4.1 shows a time plot of two AR(1) processes, one with $\phi = .9$ and one with $\phi = -.9$; in both cases, $\sigma_w^2 = 1$. In the first case, $\rho(h) = .9^h$, for $h \geq 0$, so observations close together in time are positively correlated. Thus, observations at contiguous time points will tend to be close in value to each other; this fact shows up in the top of Figure 4.1 as a very smooth sample path for x_t. Now, contrast this with the case in which $\phi = -.9$, so that $\rho(h) = (-.9)^h$, for $h \geq 0$. This result means that observations at contiguous time points are negatively correlated but observations two time points apart are positively correlated. This fact shows up in the bottom of Figure 4.1, where, for example, if an observation, x_t, is positive, the next observation, x_{t+1}, is typically negative, and the next observation, x_{t+2}, is typically positive. Thus, in this case, the sample path is very choppy. The following R code can be used to obtain a figure similar to Figure 4.1:

```
par(mfrow=c(2,1))
tsplot(arima.sim(list(order=c(1,0,0), ar=.9), n=100), ylab="x", col=4,
           main=expression(AR(1)~~~phi==+.9))
tsplot(arima.sim(list(order=c(1,0,0),ar=-.9), n=100), ylab="x", col=4,
           main=expression(AR(1)~~~phi==-.9))
```

◊

Example 4.3. AR(p) and Causality

In Example 4.1, we saw that an AR(1) has as a causal representation; for example, the AR(1) model $x_t = .9x_{t-1} + w_t$ can also be written as $x_t = \sum_{j=0}^{\infty} .9^j w_{t-j}$. In the general case, it is more difficult to go from one version to another. It is, however, possible to use the R command `ARMAtoMA` to print some of the coefficients.

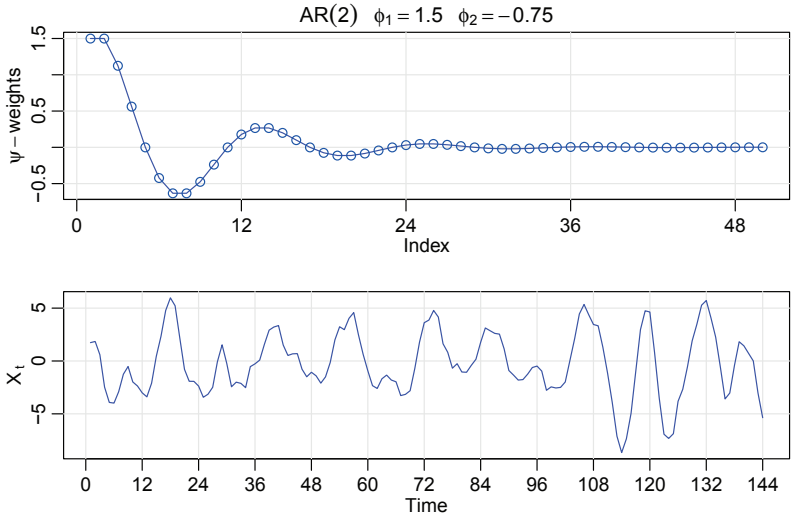


Figure 4.2 *ψ-weights and simulated data of an AR(2),* $x_t = 1.5x_{t-1} - .75x_{t-2} + w_t$.

For example, the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t\,,$$

can be written in its *causal* form, $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where $\psi_0 = 1$ and

$$\psi_j = 2(\tfrac{\sqrt{3}}{2})^j \cos(\tfrac{2\pi(j-2)}{12})\,, \qquad j = 1, 2, \ldots\,.$$

The ψ-weights were solved for using difference equation theory (see Shumway and Stoffer, 2017, §3.2). Notice that the coefficients are cyclic with a period of 12 (like monthly data), but they decrease exponentially fast to zero (because $\sqrt{3}/2 < 1$) indicating a short dependence on the past. Figure 4.2 shows a plot of the ψ_j for $j = 1, \ldots, 50$, as well as simulated data from the model. Both show the cyclic-type behavior of this particular model. It is evident that the linear process form of the model gives more insight into the model than the regression form of the model. Finally, we note that an AR(p) is also an MA(∞).

The following R code was used for this example.

```
psi = ARMAtoMA(ar = c(1.5, -.75), ma = 0, 50)
par(mfrow=c(2,1), mar=c(2,2.5,1,0)+.5, mgp=c(1.5,.6,0), cex.main=1.1)
plot(psi, xaxp=c(0,144,12), type="n", col=4,
           ylab=expression(psi-weights),
           main=expression(AR(2)~~~phi[1]==1.5~~~phi[2]==-.75))
abline(v=seq(0,48,by=12), h=seq(-.5,1.5,.5), col=gray(.9))
lines(psi, type="o", col=4)
set.seed(8675309)
simulation = arima.sim(list(order=c(2,0,0), ar=c(1.5,-.75)), n=144)
```

```
plot(simulation, xaxp=c(0,144,12), type="n", ylab=expression(X[~t]))
abline(v=seq(0,144,by=12), h=c(-5,0,5), col=gray(.9))
lines(simulation, col=4)
```

◊

We now formally define the concept of causality. The importance of this condition is to make sure that a time series model is not future-dependent. This allows us to be able to predict future values of a time series based on only the present and the past.

Definition 4.4. *A time series x_t is said to be* **causal** *if it can be written as*

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j w_{t-j}$$

for constants ψ_j satisfying $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

Remark. As stated in Property 2.21, any stationary (non-deterministic) time series has a causal representation.

As an alternative to autoregression, think of w_t as a "shock" to the process at time t. One can imagine that what happens today might be related to shocks from a few previous days. In this case, we have the moving average model of order q, abbreviated as MA(q). The *moving average model* of order q, is defined by[1]

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \tag{4.5}$$

where w_t is white noise. Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \ldots, \theta_q$. In addition, the MA($q$) is already in the form of Definition 4.4 with $\psi_j = \theta_j$ and $\theta_j = 0$ for $j > q$.

Example 4.5. The MA(1) Process

Consider the MA(1) model $x_t = w_t + \theta w_{t-1}$. Then, $E(x_t) = 0$, and if we replace x_t by $x_t - \mu$, then $E(x_t) = \mu$. The autocovariance function is

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_w^2 & h = 0, \\ \theta\sigma_w^2 & |h| = 1, \\ 0 & |h| > 1, \end{cases}$$

and the ACF is

$$\rho(h) = \begin{cases} \frac{\theta}{(1+\theta^2)} & |h| = 1, \\ 0 & |h| > 1. \end{cases}$$

Note $|\rho(1)| \leq 1/2$ for all values of θ (Problem 4.1). Also, x_t is correlated with x_{t-1}, but not with $x_{t-2}, x_{t-3}, \ldots$. Contrast this with the case of the AR(1) model in which the correlation between x_t and x_{t-k} is never zero. When $\theta = .9$, for example,

[1] Some texts and software packages write the MA model with negative coefficients; that is, $x_t = w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2} - \cdots - \theta_q w_{t-q}$.

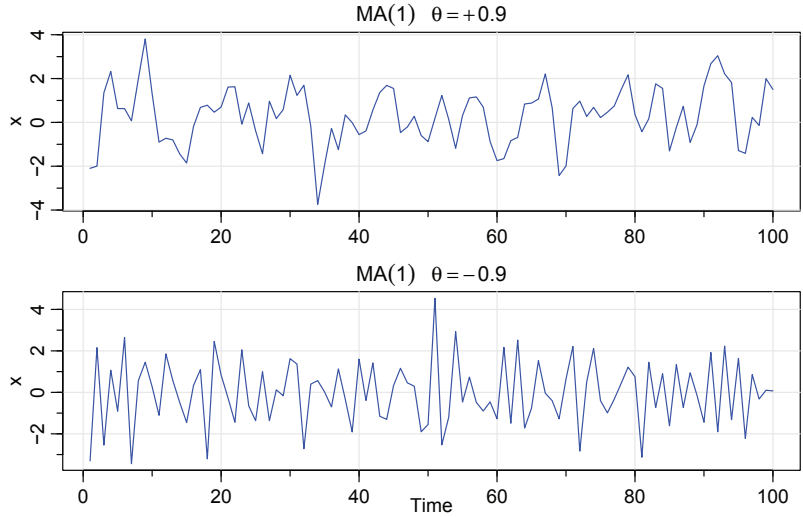


Figure 4.3 *Simulated MA(1) models: $\theta = .9$ (top); $\theta = -.9$ (bottom).*

x_t and x_{t-1} are positively correlated, and $\rho(1) = .497$. When $\theta = -.9$, x_t and x_{t-1} are negatively correlated, $\rho(1) = -.497$. Figure 4.3 shows a time plot of these two processes with $\sigma_w^2 = 1$. The series for which $\theta = .9$ is smoother than the series for which $\theta = -.9$. A figure similar to Figure 4.3 can be created in R as follows:

```
par(mfrow = c(2,1))
tsplot(arima.sim(list(order=c(0,0,1), ma=.9),  n=100), col=4,
           ylab="x", main=expression(MA(1)~~~theta==+.5))
tsplot(arima.sim(list(order=c(0,0,1), ma=-.9), n=100), col=4,
           ylab="x", main=expression(MA(1)~~~theta==-.5))
```

◊

Example 4.6. Non-uniqueness of MA Models and Invertibility
Using Example 4.5, we note that for an MA(1) model, the pair $\sigma_w^2 = 1$ and $\theta = 5$ yield the same autocovariance function as the pair $\sigma_w^2 = 25$ and $\theta = 1/5$, namely,

$$\gamma(h) = \begin{cases} 26 & h = 0, \\ 5 & |h| = 1, \\ 0 & |h| > 1. \end{cases}$$

Thus, the MA(1) processes

$$x_t = w_t + \tfrac{1}{5} w_{t-1}, \quad w_t \sim \text{iid } \mathrm{N}(0, 25)$$

and

$$y_t = v_t + 5 v_{t-1}, \quad v_t \sim \text{iid } \mathrm{N}(0, 1)$$

are stochastically the same. We can only observe the time series, x_t or y_t, and not the noise, w_t or v_t, so we cannot distinguish between the models. Hence, we will have to

choose only one of them. For convenience, by mimicking causality for AR models, we will choose the model with an infinite AR representation. Such a process is called an *invertible* process.

To discover which model is the invertible model, we can reverse the roles of x_t and w_t (because we are mimicking the AR case) and write the MA(1) model as

$$w_t = -\theta w_{t-1} + x_t\,.$$

As in (4.2), if $|\theta| < 1$, then $w_t = \sum_{j=0}^{\infty}(-\theta)^j x_{t-j}$, which is the desired infinite representation of the model. Hence, given a choice, we will choose the model with $\sigma_w^2 = 25$ and $\theta = 1/5$ because it is invertible. ◊

Henceforth, for uniqueness, we require that a moving average have an *invertible* representation:

Definition 4.7. *A time series x_t is said to be* **invertible** *if it can be written as*

$$w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}\,.$$

for constants π_j satisfying $\sum_{j=0}^{\infty}\pi_j^2 < \infty$.

Remark. Aside from the uniqueness problem, invertibility is important because it gives a representation of a present shock, w_t, in terms of the present and past data. Consequently, the current shock to the system does not depend on future data. Also, note that an MA(q) is an AR(∞).

We now proceed with the general development of mixed *autoregressive moving average* (ARMA) models for stationary time series.

Definition 4.8. *A time series $\{x_t;\ t = 0, \pm1, \pm2, \ldots\}$ is* **ARMA(p, q)** *if*

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \tag{4.6}$$

with $\phi_p \neq 0$, $\theta_q \neq 0$, $\sigma_w^2 > 0$, and the model is causal and invertible. Henceforth, unless stated otherwise, w_t is a Gaussian white noise series with mean zero and variance σ_w^2. If $E(x_t) = \mu$, then $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$.

The ARMA model may be seen as a regression of the present outcome (x_t) on the past outcomes ($x_{t-1}, \ldots, x_{t-p}$), with correlated errors. That is,

$$x_t = \beta_0 + \beta_1 x_{t-1} + \cdots + \beta_p x_{t-p} + \epsilon_t,$$

where $\epsilon_t = w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}$, although we call the regression parameters ϕ instead of β. As opposed to ordinary regression, the ϕ parameters are restricted to certain values in order to obtain causality and the θ parameters are restricted to certain values to obtain invertibility.

When $q = 0$, the model is called an autoregressive model of order p, AR(p), and when $p = 0$, the model is called a moving average model of order q, MA(q). Before

proceeding, we establish some notation based on the backshift operator defined in Definition 3.8, $B^k x_t = x_{t-k}$. Using the backshift operator, we can write the AR(p) model as

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

Thus, it is convenient to define the **autoregressive operator** as

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p. \tag{4.7}$$

so that the AR model is $\phi(B)x_t = w_t$. As in the AR(p) case, the MA(q) model may be written as

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

so we define the **moving average operator** as

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{4.8}$$

and write an MA(q) model as $x_t = \theta(B)w_t$. Consequently, an ARMA(p, q) model can be written as concisely as

$$\phi(B)(x_t - \mu) = \theta(B) w_t \,, \tag{4.9}$$

where the orders of $\phi(B)$ and $\theta(B)$ are understood to be p and q, respectively.

In addition to restricted values of the ϕs and θs, there are complications where the autoregressive side of the model can cancel the moving average side of the model. This is called over-parameterization or parameter redundancy. That is, given an ARMA(p, q) model, we can unnecessarily complicate the model by multiplying both sides by another operator, say

$$\eta(B)\phi(B)(x_t - \mu) = \eta(B)\theta(B) w_t \,,$$

without changing the dynamics. Consider the following example.

Example 4.9. Parameter Redundancy
Consider a white noise process $x_t = w_t$. Now multiply both sides of the equation by $(1 - .9B)$ to get

$$x_t - .9x_{t-1} = w_t - .9w_{t-1} \,,$$

or

$$x_t = .9x_{t-1} - .9w_{t-1} + w_t, \tag{4.10}$$

which looks like an ARMA$(1,1)$ model. Of course, x_t is still white noise; nothing has changed in this regard [i.e., $x_t = w_t$ is the solution to (4.10)], but we have hidden the fact that x_t is white noise because of the *parameter redundancy* or over-parameterization. ♢

Example 4.9 points out the need to be careful when fitting ARMA models to data. Unfortunately, *it is easy to fit an overly complex ARMA model to data*. For example, if a process is truly white noise, it is possible to fit a significant ARMA(k, k) model to the data. Consider the following example.

Example 4.10. Parameter Redundancy and Estimation
Although we have not discussed estimation yet, we present the following demonstration of the problem. We generated 150 iid normals with $\mu = 5$ and $\sigma = 1$, and then fit an ARMA$(1,1)$ to the data. Note that $\hat{\phi} = -.96$ and $\hat{\theta} = .95$, and both are significant. Below is the R code (note that the estimate called "intercept" is really the estimate of the mean).

```
set.seed(8675309)          # Jenny, I got your number
x = rnorm(150, mean=5)     # generate iid N(5,1)s
arima(x, order=c(1,0,1))   # estimation
  Coefficients:
          ar1      ma1  intercept <= misnomer
       -0.96    0.95       5.05
  s.e.  0.17    0.17       0.07
```

Of course the data are independent, but the estimation implies a seemingly different result that the data are highly dependent. ♢

Henceforth, we will require an ARMA model to be reduced to its simplest form. A simple way to discover if this problem exists with a model is to write the model with the AR part on the left and the MA part on the right, and then compare each side.

Example 4.11. Checking for Parameter Redundancy
In the previous example, it was easy to see that the left-hand and right-hand sides are nearly the same. For more complicated models, we can use R to compare each side. For example, consider the model

$$x_t = .3x_{t-1} + .4x_{t-2} + w_t + .5w_{t-1}\,,$$

which looks like an ARMA$(2,1)$. Now write the model as

$$(1 - .3B - .4B^2)x_t = (1 + .5B)w_t\,,$$

or

$$(1 + .5B)(1 - .8B)x_t = (1 + .5B)w_t\,.$$

We can cancel the $(1 + .5B)$ on each side, so the model is really an AR(1),

$$x_t = .8x_{t-1} + w_t\,.$$

These situations can be checked easily in R by looking at the roots of the polynomials in B corresponding to each side. If the roots are close, then there may be parameter redundancy:

```
AR = c(1, -.3, -.4)  # original AR coefs on the left
polyroot(AR)
 [1]  1.25-0i -2.00+0i
MA = c(1, .5)        # original MA coefs on the right
polyroot(MA)
 [1] -2+0i
```

This indicates there is one common factor (with root -2) and hence the model is over-parameterized and can be reduced. ◊

Example 4.12. Causal and Invertible ARMA
It might be useful at times to write an ARMA model in its causal or invertible forms. For example, consider the model

$$x_t = .8x_{t-1} + w_t - .5w_{t-1}\,.$$

Using R, we can list some of the causal and invertible coefficients of our ARMA(1, 1) model as follows:

```
round( ARMAtoMA(ar=.8, ma=-.5, 10), 2)   # first 10 ψ-weights
  [1] 0.30 0.24 0.19 0.15 0.12 0.10 0.08 0.06 0.05 0.04
round( ARMAtoAR(ar=.8, ma=-.5, 10), 2)  # first 10 π-weights
  [1] -0.30 -0.15 -0.08 -0.04 -0.02 -0.01  0.00  0.00  0.00  0.00
```

Thus, the causal form looks like,

$$x_t = w_t + .3w_{t-1} + .24w_{t-2} + .19w_{t-3} + \cdots + .05w_{t-9} + .04w_{t-10} + \cdots\,,$$

whereas the invertible form looks like,

$$w_t = x_t - .3x_{t-1} - .15x_{t-2} - .08x_{t-3} - .04x_{t-4} - .02x_{t-5} - .01x_{t-6} + \cdots\,.$$

If a model is not causal or invertible, the scripts will work, but the coefficients will not converge to zero. For a random walk, $x_t = x_{t-1} + w_t$, or $x_t = \sum_{j=1}^{t} w_j$, for example:

```
ARMAtoMA(ar=1, ma=0, 20)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

◊

4.2 Correlation Functions

*Autocorrelation Function (*ACF)

Example 4.13. ACF of an MA(q)
Write the model as $x_t = \sum_{j=0}^{q} \theta_j w_{t-j}$ with $\theta_0 = 1$ for ease. Because x_t is a finite linear combination of white noise terms, the process is stationary with autocovariance function

$$\gamma(h) = \operatorname{cov}(x_{t+h}, x_t) = \operatorname{cov}\Big(\sum_{j=0}^{q} \theta_j w_{t+h-j},\ \sum_{k=0}^{q} \theta_k w_{t-k}\Big)$$

$$= \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \le h \le q \\ 0 & h > q\,, \end{cases} \tag{4.11}$$

which is similar to the calculation in (2.16). The cutting off of $\gamma(h)$ after q lags is the signature of the MA(q) model. Dividing (4.11) by $\gamma(0)$ yields the ACF of an MA(q):

$$\rho(h) = \begin{cases} \dfrac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1+\theta_1^2+\cdots+\theta_q^2} & 1 \le h \le q \\ 0 & h > q. \end{cases} \tag{4.12}$$

In addition, we note that $\rho(q) \neq 0$ because $\theta_q \neq 0$. ♢

Example 4.14. ACF of an AR(p) and ARMA(p,q)
For an AR(p) or ARMA(p,q) model, write the model in its causal MA(∞) form,

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}. \tag{4.13}$$

It follows immediately that the autocovariance function of x_t can be written as

$$\gamma(h) = \operatorname{cov}(x_{t+h}, x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_{j+h}\psi_j, \quad h \ge 0, \tag{4.14}$$

as was calculated in (2.16). The ACF is given by

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\sum_{j=0}^{\infty} \psi_{j+h}\psi_j}{\sum_{j=0}^{\infty} \psi_j^2}, \quad h \ge 0. \tag{4.15}$$

Unlike the MA(q), the ACF of an AR(p) or an ARMA(p,q) does not cut off at any lag, so using the ACF to help identify the order of an AR or ARMA is difficult. ♢

Result (4.15) is not appealing in that it provides little information about the appearance of the ACF of various models. We can, however, look at what happens for some specific models.

Example 4.15. ACF of an AR(2)
Figure 4.2 shows $n = 144$ observations from the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

with $\sigma_w^2 = 1$. We examined this model in Example 4.3 where we noted that the process exhibits pseudo-cyclic behavior at the rate of one cycle every 12 time points. Because the ψ-weights are cyclic, the ACF of the model will also be cyclic with a period of 12. The R code to calculate and display the ACF for this model as shown on the left side of Figure 4.4 is:

```
ACF = ARMAacf(ar=c(1.5,-.75), ma=0, 50)
plot(ACF, type="h", xlab="lag", panel.first=Grid())
abline(h=0)
```

♢

The general behavior of the ACF of an AR(p) or an ARMA(p,q) is controlled by the AR part because the MA part has only finite influence.

Example 4.16. The ACF of an ARMA(1,1)
Consider the ARMA(1,1) process $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$. Using the theory of difference equations, we can show that the ACF is given by

$$\rho(h) = \frac{(1+\theta\phi)(\phi+\theta)}{\phi(1+2\theta\phi+\theta^2)}\,\phi^h, \quad h \geq 1. \tag{4.16}$$

Notice that the general pattern of $\rho(h)$ in (4.16) is not different from that of an AR(1) given in (4.4), $\rho(h) = \phi^h$. Hence, it is unlikely that we will be able to tell the difference between an ARMA(1,1) and an AR(1) based solely on an ACF estimated from a sample. This consideration will lead us to the partial autocorrelation function. ◊

*Partial Autocorrelation Function (*PACF)

In (4.12), we saw that for MA(q) models, the ACF will be zero for lags greater than q. Moreover, because $\theta_q \neq 0$, the ACF will not be zero at lag q. Thus, the ACF provides a considerable amount of information about the order of the dependence when the process is a moving average process.

If the process, however, is ARMA or AR, the ACF alone tells us little about the orders of dependence. Hence, it is worthwhile pursuing a function that will behave like the ACF of MA models, but for AR models, namely, the *partial autocorrelation function (PACF)*.

Recall that if X, Y, and Z are random variables, then the partial correlation between X and Y given Z is obtained by regressing X on Z to obtain the predictor $\hat{X}$, regressing Y on Z to obtain $\hat{Y}$, and then calculating

$$\rho_{XY|Z} = \text{corr}\{X - \hat{X},\, Y - \hat{Y}\}.$$

The idea is that $\rho_{XY|Z}$ measures the correlation between X and Y with the linear effect of Z removed (or partialled out). If the variables are multivariate normal, then this definition coincides with $\rho_{XY|Z} = \text{corr}(X, Y \mid Z)$.

To motivate the idea of partial *auto*correlation, consider a causal AR(1) model, $x_t = \phi x_{t-1} + w_t$. Then,

$$\begin{aligned}\gamma_x(2) &= \text{cov}(x_t, x_{t-2}) = \text{cov}(\phi x_{t-1} + w_t, x_{t-2}) \\ &= \text{cov}(\phi x_{t-1}, x_{t-2}) = \phi\gamma_x(1).\end{aligned}$$

Note that $\text{cov}(w_t, x_{t-2}) = 0$ from causality because x_{t-2} involves $\{w_{t-2}, w_{t-3}, \ldots\}$, which are all uncorrelated with w_t. The correlation between x_t and x_{t-2} is not zero as it would be for an MA(1) because x_t is dependent on x_{t-2} through x_{t-1}. Suppose we break this chain of dependence by removing (or partialling out) the effect of x_{t-1}. That is, we consider the correlation between $x_t - \phi x_{t-1}$ and $x_{t-2} - \phi x_{t-1}$, because it is the correlation between x_t and x_{t-2} with the linear dependence of each on x_{t-1} removed. In this way, we have broken the dependence chain between x_t and x_{t-2},

$$\text{cov}(x_t - \phi x_{t-1}, x_{t-2} - \phi x_{t-1}) = \text{cov}(w_t, x_{t-2} - \phi x_{t-1}) = 0.$$

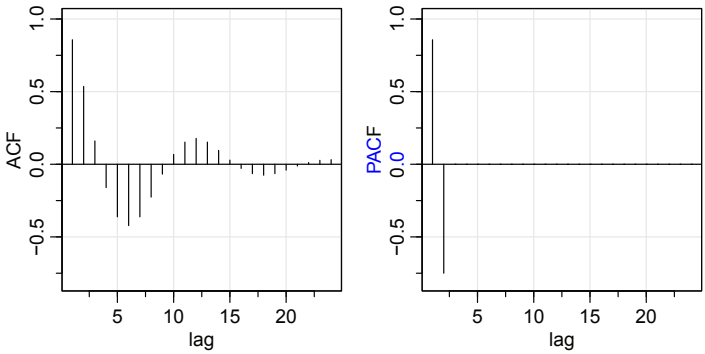


Figure 4.4 *The ACF and PACF of an AR(2) model with* $\phi_1 = 1.5$ *and* $\phi_2 = -.75$.

Hence, the tool we need is partial autocorrelation, which is the correlation between x_s and x_t with the linear effect of everything "in the middle" removed.

Definition 4.17. *The* **partial autocorrelation function (PACF)** *of a stationary process,* x_t*, denoted* ϕ_{hh}*, for* $h = 1, 2, \ldots,$ *is*

$$\phi_{11} = \text{corr}(x_1, x_0) = \rho(1) \tag{4.17}$$

and

$$\phi_{hh} = \text{corr}(x_h - \hat{x}_h,\ x_0 - \hat{x}_0), \quad h \geq 2, \tag{4.18}$$

where $\hat{x}_h$ *is the regression of* x_h *on* $\{x_1, x_2, \ldots, x_{h-1}\}$ *and* $\hat{x}_0$ *is the regression of* x_0 *on* $\{x_1, x_2, \ldots, x_{h-1}\}$.

Thus, due to the stationarity, the PACF, ϕ_{hh}, is the correlation between x_{t+h} and x_t with the linear dependence of everything between them, namely $\{x_{t+1}, \ldots, x_{t+h-1}\}$, on each, removed.

It is not necessary to actually run regressions to compute the PACF because the values can be computed recursively based on what is known as the Durbin–Levinson algorithm due to Levinson (1947) and Durbin (1960).

Example 4.18. The PACF of an AR(p)

The PACF of an AR(p) model will be zero for all lags larger than p and the PACF at lag p will not be zero because it can be shown that $\phi_{pp} = \phi_p$ (the last parameter in the model).

In Example 4.15 we looked at the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t\,.$$

In this case, $\phi_{11} = \rho(1) = \phi_1/(1-\phi_2) = 1.5/1.75 \approx .86$, $\phi_{22} = \phi_2 = -.75$, and $\phi_{hh} = 0$ for $h > 2$. Figure 4.4 shows the ACF and the PACF of this AR(2) model. To reproduce Figure 4.4 in R, use the following commands:

Table 4.1 *Behavior of the ACF and PACF for ARMA Models*

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

```
ACF = ARMAacf(ar=c(1.5,-.75), ma=0, 24)[-1]
PACF = ARMAacf(ar=c(1.5,-.75), ma=0, 24, pacf=TRUE)
par(mfrow=1:2)
tsplot(ACF, type="h", xlab="lag", ylim=c(-.8,1))
abline(h=0)
tsplot(PACF, type="h", xlab="lag", ylim=c(-.8,1))
abline(h=0)
```

◊

We also have the following large sample result for the PACF, which may be compared to the similar result for the ACF given in Property 2.28.

Property 4.19 (Large Sample Distribution of the PACF). *If a time series is an AR(p) and the sample size n is large, then for $h > p$, the $\hat{\phi}_{hh}$ are approximately independent normal with mean 0 and standard deviation $1/\sqrt{n}$. This result also holds for $p = 0$, wherein the process is white noise.*

Example 4.20. The PACF of an MA(q)
An MA(q) is invertible, so it has an AR(∞) representation,

$$x_t = -\sum_{j=1}^{\infty} \pi_j x_{t-j} + w_t\,.$$

Moreover, no finite representation exists. From this result, it should be apparent that the PACF will never cut off, as in the case of an AR(p). For an MA(1), $x_t = w_t + \theta w_{t-1}$, with $|\theta| < 1$, it can be shown that

$$\phi_{hh} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}, \quad h \geq 1.$$

◊

The PACF for MA models behaves much like the ACF for AR models. Also, the PACF for AR models behaves much like the ACF for MA models. Because an invertible ARMA model has an infinite AR representation, the PACF will not cut off. We may summarize these results in Table 4.1.

Example 4.21. Preliminary Analysis of the Recruitment Series
We consider the problem of modeling the Recruitment series shown in Figure 1.5. There are 453 months of observed recruitment ranging over the years 1950–1987. The ACF and the PACF given in Figure 4.5 are consistent with the behavior of

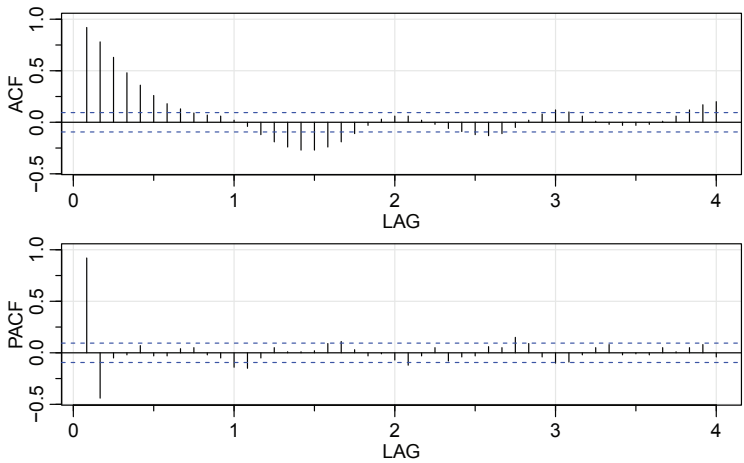


Figure 4.5 *ACF and PACF of the Recruitment series. Note that the lag axes are in terms of season (12 months in this case).*

an AR(2). The ACF has cycles corresponding roughly to a 12-month period, and the PACF has large values for $h = 1, 2$ and then is essentially zero for higher-order lags. Based on Table 4.1, these results suggest that a second-order ($p = 2$) autoregressive model might provide a good fit. Although we will discuss estimation in detail in Section 4.3, we ran a regression (OLS) using the data triplets $\{(x; z_1, z_2) : (x_3; x_2, x_1), (x_4; x_3, x_2), \ldots, (x_{453}; x_{452}, x_{451})\}$ to fit the model

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

for $t = 3, 4, \ldots, 453$. The values of the estimates were $\hat{\phi}_0 = 6.74_{(1.11)}$, $\hat{\phi}_1 = 1.35_{(.04)}, \hat{\phi}_2 = -.46_{(.04)}$, and $\hat{\sigma}_w^2 = 89.72$, where the estimated standard errors are in parentheses.

The following R code can be used for this analysis. We use the script `acf2` from `astsa` to print and plot the ACF and PACF.

```
acf2(rec, 48)      # will produce values and a graphic
(regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE))
   Coefficients:
         1        2
   1.3541  -0.4632
 Intercept: 6.737 (1.111)
 sigma^2 estimated as  89.72
regr$asy.se.coef  # standard errors of the estimates
 $ar
   [1] 0.04178901 0.04187942
```

We could have used `lm()` to do the regression, however using `ar.ols()` is much simpler for pure AR models. Also, the term `intercept` is used correctly here. ◊

4.3 Estimation

Throughout this section, we assume we have n observations, $x_1, \ldots, x_n$, from an ARMA(p, q) process in which, initially, the order parameters, p and q, are known. Our goal is to estimate the parameters, $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$, and σ_w^2.

We begin with *method of moments* estimators. The idea behind these estimators is that of equating population moments, $E(x_t^k)$, to sample moments, $\frac{1}{n}\sum_{t=1}^n x_t^k$, for $k = 1, 2, \ldots$, and then solving for the parameters in terms of the sample moments. We immediately see that if $E(x_t) = \mu$, the method of moments estimator of μ is the sample average, $\bar{x}$ ($k = 1$). Thus, while discussing method of moments, we will assume $\mu = 0$. Although the method of moments can produce good estimators, they can sometimes lead to suboptimal estimators. We first consider the case in which the method leads to optimal (efficient) estimators, that is, AR(p) models,

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t\,.$$

If we multiply each side of the AR equation by x_{t-h} for $h = 0, 1, \ldots, p$ and take expectation, we obtain the following result.

Definition 4.22. *The* **Yule–Walker equations** *are given by*

$$\rho(h) = \phi_1\rho(h-1) + \cdots + \phi_p\rho(h-p), \quad h = 1, 2, \ldots, p, \tag{4.19}$$

$$\sigma_w^2 = \gamma(0)\,[1 - \phi_1\rho(1) - \cdots - \phi_p\rho(p)]. \tag{4.20}$$

The estimators obtained by replacing $\gamma(0)$ with its estimate, $\hat{\gamma}(0)$ and $\rho(h)$ with its estimate, $\hat{\rho}(h)$, are called the *Yule–Walker estimators*. For AR(p) models, if the sample size is large, the Yule–Walker estimators are approximately normally distributed, and $\hat{\sigma}_w^2$ is close to the true value of σ_w^2. In addition, the estimates are close to the OLS estimates discussed in Example 4.21.

Example 4.23. Yule–Walker Estimation for an AR(1)
For an AR(1), $(x_t - \mu) = \phi(x_{t-1} - \mu) + w_t$, the mean estimate is $\hat{\mu} = \bar{x}$, and (4.19) is

$$\rho(1) = \phi\rho(0) = \phi\,,$$

so

$$\hat{\phi} = \hat{\rho}(1) = \frac{\sum_{t=1}^{n-1}(x_{t+1} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}\,,$$

as expected. The estimate of the error variance is then

$$\hat{\sigma}_w^2 = \hat{\gamma}(0)\,[1 - \hat{\phi}^2]\,;$$

recall $\gamma(0) = \sigma_w^2/(1 - \phi^2)$ from (4.3). ♢

Example 4.24. Yule–Walker Estimation of the Recruitment Series
In Example 4.21 we fit an AR(2) model to the Recruitment series using regression. Below are the results of fitting the same model using Yule–Walker estimation, which are close to the regression values in Example 4.21.

```
rec.yw = ar.yw(rec, order=2)
rec.yw$x.mean     # mean estimate
 [1] 62.26278
rec.yw$ar         # phi parameter estimates
 [1]  1.3315874  -0.4445447
sqrt(diag(rec.yw$asy.var.coef))  # their standard errors
 [1] 0.04222637 0.04222637
rec.yw$var.pred  # error variance estimate
 [1] 94.79912
```

◊

In the case of AR(p) models, the Yule–Walker estimators are optimal estimators, but this is not true for MA(q) or ARMA(p,q) models. AR(p) models are basically linear models, and the Yule–Walker estimators are essentially least squares estimators. MA or ARMA models are nonlinear models, so this technique does not give optimal estimators.

Example 4.25. Method of Moments Estimation for an MA(1)
Consider the MA(1) model, $x_t = w_t + \theta w_{t-1}$, where $|\theta| < 1$. The model can then be written as

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

which is nonlinear in θ. The first two population autocovariances are $\gamma(0) = \sigma_w^2(1+\theta^2)$ and $\gamma(1) = \sigma_w^2\theta$, so the estimate of θ is found by solving

$$\hat{\rho}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\hat{\theta}}{1+\hat{\theta}^2}.$$

Two solutions exist, so we would pick the invertible one. If $|\hat{\rho}(1)| \leq \frac{1}{2}$, the solutions are real, otherwise, a real solution does not exist. Even though $|\rho(1)| < \frac{1}{2}$ for an invertible MA(1), it may happen that $|\hat{\rho}(1)| \geq \frac{1}{2}$ because it is an estimator. For example, the following simulation in R produces a value of $\hat{\rho}(1) = .51$ when the true value is $\rho(1) = .9/(1+.9^2) = .497$.

```
set.seed(2)
ma1 = arima.sim(list(order = c(0,0,1), ma = 0.9), n = 50)
acf1(ma1, plot=FALSE)[1]
 [1] 0.51
```

◊

The preferred method of estimation is maximum likelihood estimation (MLE), which determines the values of the parameters that are most *likely* to have produced the observations. MLE for an AR(1) is discussed in detail in Section D.1. For normal models, this is the same as weighted least squares. For ease, we first discuss conditional least squares.

Conditional Least Squares

Recall from Chapter 3, in simple linear regression, $x_t = \beta_0 + \beta_1 z_t + w_t$, we minimize

$$S(\beta) = \sum_{t=1}^{n} w_t^2(\beta) = \sum_{t=1}^{n} (x_t - [\beta_0 + \beta_1 z_t])^2$$

with respect to the βs. This is a simple problem because we have all the data pairs, (z_t, x_t) for $t = 1, \ldots, n$. For ARMA models, we do not have this luxury.

Consider a simple AR(1) model, $x_t = \phi x_{t-1} + w_t$. In this case, the error sum of squares is

$$S(\phi) = \sum_{t=1}^{n} w_t^2(\phi) = \sum_{t=1}^{n} (x_t - \phi x_{t-1})^2 .$$

We have a problem because we didn't observe x_0. Let's make life easier by forgetting the problem and dropping the first term. That is, let's perform least squares using the (conditional) sum of squares,

$$S_c(\phi) = \sum_{t=2}^{n} w_t^2(\phi) = \sum_{t=2}^{n} (x_t - \phi x_{t-1})^2$$

because that's easy (it's just OLS) and if n is large, it shouldn't matter much. We know from regression that the solution is

$$\hat{\phi} = \frac{\sum_{t=2}^{n} x_t x_{t-1}}{\sum_{t=2}^{n} x_{t-1}^2} ,$$

which is nearly the Yule–Walker estimate in Example 4.23 (replace x_t by $x_t - \bar{x}$ if the mean is not zero).

Now we focus on conditional least squares for ARMA(p, q) models via *Gauss–Newton*. Write the model parameters as $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$, and for the ease of discussion, we will put $\mu = 0$. Write the ARMA model in terms of the errors

$$w_t(\beta) = x_t - \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{k=1}^{q} \theta_k w_{t-k}(\beta), \tag{4.21}$$

emphasizing the dependence of the errors on the parameters (recall that $w_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}$ by invertibility, and the π_j are complicated functions of β).

Again we have the problem that we don't observe the x_t for $t \leq 0$, nor the errors w_t. For *conditional least squares*, we condition on $x_1, \ldots, x_p$ (if $p > 0$) and set $w_p = w_{p-1} = w_{p-2} = \cdots = w_{p+1-q} = 0$ (if $q > 0$), in which case, given β, we may evaluate (4.21) for $t = p + 1, \ldots, n$. For example, for an ARMA(1, 1),

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

we would start at $p+1=2$ and set $w_1=0$ so that

$$\begin{aligned} w_2 &= x_2 - \phi x_1 - \theta w_1 = x_2 - \phi x_1 \\ w_3 &= x_3 - \phi x_2 - \theta w_2 \\ &\vdots \\ w_n &= x_n - \phi x_{n-1} - \theta w_{n-1} \end{aligned}$$

Given data, we can evaluate these errors at any values of the parameters; e.g., $\phi = \theta = .5$. Using this conditioning argument, the conditional error sum of squares is

$$S_c(\beta) = \sum_{t=p+1}^{n} w_t^2(\beta). \tag{4.22}$$

Minimizing $S_c(\beta)$ with respect to β yields the conditional least squares estimates. We could use a brute force method where we evaluate $S_c(\beta)$ over a grid of possible values for the parameters and choose the values with the smallest error sum of squares, but this method becomes prohibitive if there are many parameters.

If $q=0$, the problem is linear regression as we saw in the case of the AR(1). If $q>0$, the problem becomes nonlinear regression and we will rely on numerical optimization. Gauss–Newton is an iterative method for solving the problem of minimizing (4.22). We demonstrate the method for an MA(1).

Example 4.26. Gauss–Newton for an MA(1)
Consider an MA(1) process, $x_t = w_t + \theta w_{t-1}$. Write the errors as

$$w_t(\theta) = x_t - \theta w_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{4.23}$$

where we condition on $w_0(\theta) = 0$. Our goal is to find the value of θ that minimizes $S_c(\theta) = \sum_{t=1}^{n} w_t^2(\theta)$, which is a nonlinear function of θ.

Let $\theta_{(0)}$ be an initial estimate of θ, for example the method of moments estimate. Now we use a first-order Taylor approximation of $w_t(\theta)$ at $\theta_{(0)}$ to get

$$S_c(\theta) = \sum_{t=1}^{n} w_t^2(\theta) \approx \sum_{t=1}^{n} \big[w_t(\theta_{(0)}) - (\theta - \theta_{(0)}) z_t(\theta_{(0)})\big]^2, \tag{4.24}$$

where

$$z_t(\theta_{(0)}) = -\frac{\partial w_t(\theta)}{\partial \theta}\bigg|_{\theta=\theta_{(0)}},$$

(writing the derivative in the negative simplifies the algebra at the end). It turns out that the derivatives have a simple form that makes them easy to evaluate. Taking derivatives in (4.23),

$$\frac{\partial w_t(\theta)}{\partial \theta} = -w_{t-1}(\theta) - \theta \frac{\partial w_{t-1}(\theta)}{\partial \theta}, \quad t = 1, \ldots, n, \tag{4.25}$$

where we set $\partial w_0(\theta)/\partial\theta = 0$. We can also write (4.25) as

$$z_t(\theta) = w_{t-1}(\theta) - \theta z_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{4.26}$$

where $z_0(\theta) = 0$. This implies that the derivative sequence is an AR process, which we may easily compute recursively given a value of θ.

We will write the right side of (4.24) as

$$Q(\theta) = \sum_{t=1}^{n} \Big[\underbrace{w_t(\theta_{(0)})}_{y_t} - \underbrace{(\theta - \theta_{(0)})}_{\beta} \underbrace{z_t(\theta_{(0)})}_{z_t} \Big]^2 \tag{4.27}$$

and this is the quantity that we will minimize. The problem is now simple linear regression ("$y_t = \beta z_t + \epsilon_t$"), so that

$$\widehat{(\theta - \theta_{(0)})} = \textstyle\sum_{t=1}^{n} z_t(\theta_{(0)}) w_t(\theta_{(0)}) \;/\; \sum_{t=1}^{n} z_t^2(\theta_{(0)}),$$

or

$$\hat{\theta} = \theta_{(0)} + \textstyle\sum_{t=1}^{n} z_t(\theta_{(0)}) w_t(\theta_{(0)}) \;/\; \sum_{t=1}^{n} z_t^2(\theta_{(0)}).$$

Consequently, the Gauss–Newton procedure in this case is, on iteration $j+1$, set

$$\theta_{(j+1)} = \theta_{(j)} + \frac{\sum_{t=1}^{n} z_t(\theta_{(j)}) w_t(\theta_{(j)})}{\sum_{t=1}^{n} z_t^2(\theta_{(j)})}, \quad j = 0, 1, 2, \ldots, \tag{4.28}$$

where the values in (4.28) are calculated recursively using (4.23) and (4.26). The calculations are stopped when $|\theta_{(j+1)} - \theta_{(j)}|$, or $|Q(\theta_{(j+1)}) - Q(\theta_{(j)})|$, are smaller than some preset amount. ♢

Example 4.27. Fitting the Glacial Varve Series

Consider the glacial varve series (say x_t) analyzed in Example 3.12 and in Problem 3.6, where it was argued that a first-order moving average model might fit the logarithmically transformed and differenced varve series, say,

$$\nabla \log(x_t) = \log(x_t) - \log(x_{t-1}).$$

The transformed series and the sample ACF and PACF are shown in Figure 4.6 and based on Table 4.1, confirm the tendency of $\nabla \log(x_t)$ to behave as a first-order moving average. The code to display the output of Figure 4.6 is:

```
tsplot(diff(log(varve)), col=4, ylab=expression(nabla~log~X[~t]),
          main="Transformed Glacial Varves")
acf2(diff(log(varve)))
```

We see $\hat{\rho}(1) = -.4$ and using method of moments for our initial estimate:

$$\theta_{(0)} = \frac{1 - \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)} = -.5$$

based on Example 4.25 and the quadratic formula. The R code to run the Gauss–Newton and the results are:

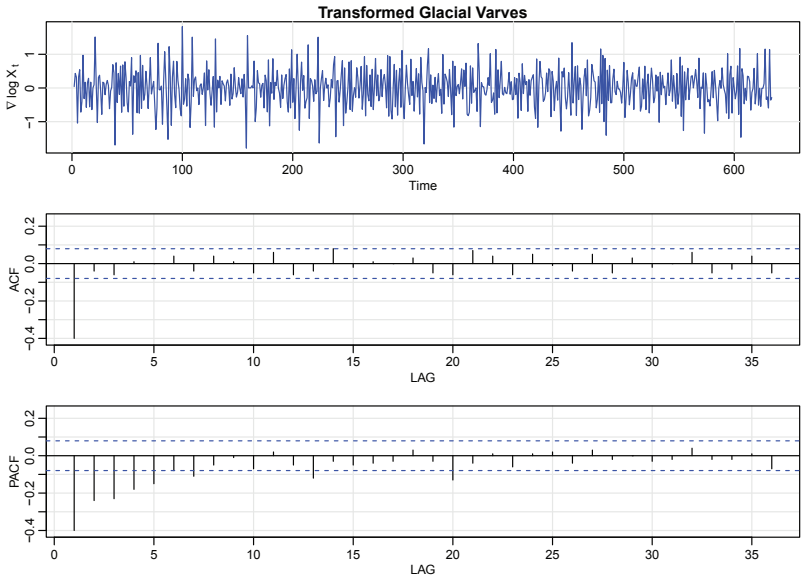


Figure 4.6 *Transformed glacial varves and corresponding sample ACF and PACF.*

```
x = diff(log(varve))                          # data
r = acf1(x, 1, plot=FALSE)                    # acf(1)
c(0) -> w -> z -> Sc -> Sz -> Szw -> para # initialize
num  = length(x)                              # = 633
## Estimation
para[1] = (1-sqrt(1-4*(r^2)))/(2*r)           # MME
niter   = 12
for (j in 1:niter){
  for  (i in 2:num){ w[i] = x[i]   - para[j]*w[i-1]
                     z[i] = w[i-1] - para[j]*z[i-1]
  }
  Sc[j]     = sum(w^2)
  Sz[j]     = sum(z^2)
  Szw[j]    = sum(z*w)
  para[j+1] = para[j] + Szw[j]/Sz[j]
}
## Results
cbind(iteration=1:niter-1, thetahat=para[1:niter], Sc, Sz)
iteration    thetahat        Sc        Sz
        0  -0.5000000  158.4258  172.1110
        1  -0.6704205  150.6786  236.8917
        2  -0.7340825  149.2539  301.6214
        3  -0.7566814  149.0291  337.3468
        4  -0.7656857  148.9893  354.4164
        5  -0.7695230  148.9817  362.2777
```

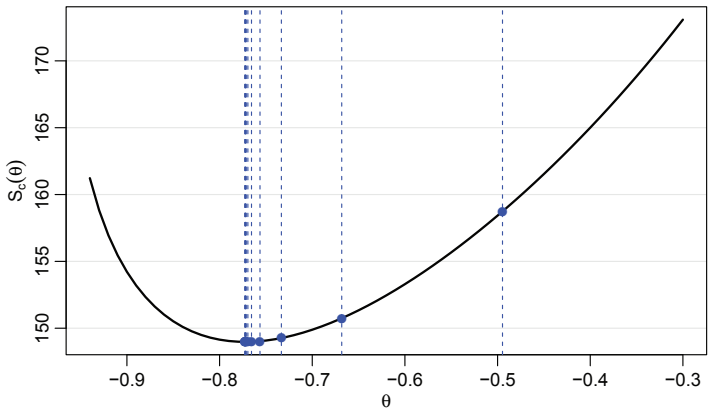


Figure 4.7 *Conditional sum of squares versus values of the moving average parameter for the glacial varve example, Example 4.27. Vertical lines indicate the values of the parameter obtained via Gauss–Newton.*

```
 6  -0.7712091  148.9802  365.8518
 7  -0.7719602  148.9799  367.4683
 8  -0.7722968  148.9799  368.1978
 9  -0.7724482  148.9799  368.5266
10  -0.7725162  148.9799  368.6748
11  -0.7725469  148.9799  368.7416
```

The estimate is

$$\hat{\theta} = \theta_{(11)} = -.773\,,$$

which results in the conditional sum of squares at convergence being

$$S_c(-.773) = 148.98\,.$$

The final estimate of the error variance is

$$\hat{\sigma}_w^2 = \frac{148.98}{632} = .236$$

with 632 degrees of freedom. The value of the sum of the squared derivatives at convergence is $\sum_{t=1}^{n} z_t^2(\theta_{(11)}) = 368.74$ and consequently, the estimated standard error of $\hat{\theta}$ is

$$\mathrm{SE}(\hat{\theta}) = \sqrt{.236/368.74} = .025$$

using the standard regression results as an approximation. This leads to a t-value of $-.773/.025 = -30.92$ with 632 degrees of freedom.

Figure 4.7 displays the conditional sum of squares, $S_c(\theta)$ as a function of θ, as well as indicating the values of each step of the Gauss–Newton algorithm. Note that the Gauss–Newton procedure takes large steps toward the minimum initially, and then takes very small steps as it gets close to the minimizing value.

```
## Plot conditional SS
c(0) -> w -> cSS
th = -seq(.3, .94, .01)
for (p in 1:length(th)){
  for (i in 2:num){ w[i] = x[i] - th[p]*w[i-1]
  }
  cSS[p] = sum(w^2)
}
tsplot(th, cSS, ylab=expression(S[c](theta)), xlab=expression(theta))
abline(v=para[1:12], lty=2, col=4)    #  add previous results to plot
points(para[1:12], Sc[1:12], pch=16, col=4)
```

◊

Unconditional Least Squares and MLE

Estimation of the parameters in an ARMA model is more like weighted least squares than ordinary least squares. Consider the normal regression model

$$x_t = \beta_0 + \beta_1 z_t + \epsilon_t\,,$$

where now, the errors have possibly different variances,

$$\epsilon_t \sim N(0, \sigma^2 h_t)\,.$$

In this case, we use weighted least squares to minimize

$$S(\beta) = \sum_{t=1}^{n} \frac{\epsilon_t^2(\beta)}{h_t} = \sum_{t=1}^{n} \frac{1}{h_t}\Big(x_t - [\beta_0 + \beta_1 z_t]\Big)^2$$

with respect to the βs. This problem is more difficult because the weights, $1/h_t$, are often unknown (the case $h_t = 1$ is ordinary least squares). For ARMA models, however, we do know the structure of these variances.

For ease, we'll concentrate on the full AR(1) model,

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t \tag{4.29}$$

where $|\phi| < 1$ and $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$. Given data $x_1, x_2, \ldots, x_n$, we cannot regress x_1 on x_0 because it is not observed. However, we know from Example 4.1 that

$$x_1 = \mu + \epsilon_1 \qquad \epsilon_1 \sim N(0, \sigma_w^2/(1-\phi^2))\,.$$

In this case, we have $h_1 = 1/(1-\phi^2)$. For $t = 2, \ldots, n$, the model is ordinary linear regression with w_t as the regression error, so that $h_t = 1$ for $t \geq 2$. Thus, the unconditional sum of squares is now

$$S(\mu, \phi) = (1-\phi^2)(x_1 - \mu)^2 + \sum_{t=2}^{n} [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2\,. \tag{4.30}$$

In conditional least squares, we conditioned away the nasty part involving x_1 to make the problem easier. For unconditional least squares, we need to use numerical optimization even for the simple AR(1) case.

This problem generalizes in an obvious way to AR(p) models and in a not so obvious way to ARMA models. For us, unconditional least squares is equivalent to maximum likelihood estimation (MLE). MLE involves finding the "most likely" parameters given the data and is discussed further in Section D.1. In the general case of causal and invertible ARMA(p, q) models, maximum likelihood estimation, least squares estimation (conditional and unconditional), and Yule–Walker estimation in the case of AR models, all lead to optimal estimators for large sample sizes.

Example 4.28. Transformed Glacial Varves (cont)
In Example 4.27, we used Gauss–Newton to fit an MA(1) model to the transformed glacial varve series via conditional least squares. To use unconditional least squares (equivalently MLE), we can use the script `sarima` from `astsa` as follows. The script requires specification of the AR order (p), the MA order (q), and the order of differencing (d). In this case, we are already differencing the data, so we set $d = 0$; we will discuss this further in the next chapter. In addition, the transformed data appear to have a zero mean function so we do not fit a mean to the data. This is accomplished by specifying `no.constant=TRUE` in the call.

```
sarima(diff(log(varve)), p=0, d=0, q=1, no.constant=TRUE)
# partial output
 initial  value -0.551778
 iter   2 value -0.671626
 iter   3 value -0.705973
 iter   4 value -0.707314
 iter   5 value -0.722372
 iter   6 value -0.722738   # conditional SS
 iter   7 value -0.723187
 iter   8 value -0.723194
 iter   9 value -0.723195
 final  value -0.723195
 converged
 initial  value -0.722700
 iter   2 value -0.722702   # unconditional SS (MLE)
 iter   3 value -0.722702
 final  value -0.722702
 converged
 ---
 Coefficients:
          ma1
      -0.7705
 s.e.   0.0341
 sigma^2 estimated as 0.2353:  log likelihood = -440.72,  aic = 885.44
```

The script starts by using the data to pick initial values of the estimates that are

within the causal and invertible region of the parameter space. Then, the script uses conditional least squares as in Example 4.27. Once that process has converged, the next step is to use the conditional estimates to find the unconditional least squares estimates (or MLEs).

The output shows only the iteration number and the value of the sum of squares. It is a good idea to look at the results of the numerical optimization to make sure it converges and that there are no warnings. If there is trouble converging or there are warnings, it usually means that the proposed model is not even close to reality.

The final estimates are $\hat{\theta} = -.7705_{(.034)}$ and $\hat{\sigma}_w^2 = .2353$. These are nearly the values obtained in Example 4.27, which were $\hat{\theta} = -.771_{(.025)}$ and $\hat{\sigma}_w^2 = .236$. $\Diamond$

Most packages use large sample theory to evaluate the estimated standard errors (standard deviation of an estimate). We give a few examples in the following proposition.

Property 4.29 (Some Specific Large Sample Distributions). *In the following, read* AN *as "approximately normal for large sample size".*

AR(1):

$$\hat{\phi}_1 \sim \text{AN}\left[\phi_1, n^{-1}(1-\phi_1^2)\right] \tag{4.31}$$

Thus, an approximate $100(1-\alpha)\%$ *confidence interval for* ϕ_1 *is*

$$\hat{\phi}_1 \pm z_{\alpha/2}\sqrt{\tfrac{1-\hat{\phi}_1^2}{n}}\,.$$

AR(2):

$$\hat{\phi}_1 \sim \text{AN}\left[\phi_1, n^{-1}(1-\phi_2^2)\right] \quad \text{and} \quad \hat{\phi}_2 \sim \text{AN}\left[\phi_2, n^{-1}(1-\phi_2^2)\right] \tag{4.32}$$

Thus, approximate $100(1-\alpha)\%$ *confidence intervals for* ϕ_1 *and* ϕ_2 *are*

$$\hat{\phi}_1 \pm z_{\alpha/2}\sqrt{\tfrac{1-\hat{\phi}_2^2}{n}} \quad \textit{and} \quad \hat{\phi}_2 \pm z_{\alpha/2}\sqrt{\tfrac{1-\hat{\phi}_2^2}{n}}\,.$$

MA(1):

$$\hat{\theta}_1 \sim \text{AN}\left[\theta_1, n^{-1}(1-\theta_1^2)\right] \tag{4.33}$$

Confidence intervals for the MA examples are similar to the AR examples.

MA(2):

$$\hat{\theta}_1 \sim \text{AN}\left[\theta_1, n^{-1}(1-\theta_2^2)\right] \quad \text{and} \quad \hat{\theta}_2 \sim \text{AN}\left[\theta, n^{-1}(1-\theta_2^2)\right] \tag{4.34}$$

Example 4.30. Overfitting Caveat

The large sample behavior of the parameter estimators gives us an additional insight into the problem of fitting ARMA models to data. For example, suppose a time series follows an AR(1) process and we decide to fit an AR(2) to the data. Do any problems occur in doing this? More generally, why not simply fit large-order

AR models to make sure that we capture the dynamics of the process? After all, if the process is truly an AR(1), the other autoregressive parameters will not be significant. The answer is that if we *overfit*, we obtain less efficient, or less precise parameter estimates. For example, if we fit an AR(1) to an AR(1) process, for large n, $\mathrm{var}(\hat{\phi}_1) \approx n^{-1}(1-\phi_1^2)$. But, if we fit an AR(2) to the AR(1) process, for large n, $\mathrm{var}(\hat{\phi}_1) \approx n^{-1}(1-\phi_2^2) = n^{-1}$ because $\phi_2 = 0$. Thus, the variance of ϕ_1 has been inflated, making the estimator less precise.

We do want to mention, however, that overfitting can be used as a diagnostic tool. For example, if we fit an AR(1) model to the data and are satisfied with that model, then adding one more parameter and fitting an AR(2) should lead to approximately the same model as in the AR(1) fit. We will discuss model diagnostics in more detail in Section 5.2. ◊

4.4 Forecasting

In forecasting, the goal is to predict future values of a time series, x_{n+m}, $m = 1, 2, \ldots$, based on the data, $x_1, \ldots, x_n$, collected to the present. Throughout this section, we will assume that the model parameters are known. When the parameters are unknown, we replace them with their estimates.

To understand how to forecast an ARMA process, it is instructive to investigate forecasting an AR(1),

$$x_t = \phi x_{t-1} + w_t\,.$$

First, consider *one-step-ahead prediction*, that is, given data $x_1, \ldots, x_n$, we wish to forecast the value of the time series at the next time point, x_{n+1}. We will call the forecast x_{n+1}^n. In general, the notation x_t^n refers to what we can expect x_t to be given the data $x_1, \ldots, x_n$.[2] Since

$$x_{n+1} = \phi x_n + w_{n+1}\,,$$

we should have

$$x_{n+1}^n = \phi x_n^n + w_{n+1}^n\,.$$

But since we know x_n (it is one of our observations), $x_n^n = x_n$, and since w_{n+1} is a future error and independent of $x_1, \ldots, x_n$, we have $w_{n+1}^n = E(w_{n+1}) = 0$. Consequently, the *one-step-ahead forecast* is

$$x_{n+1}^n = \phi x_n\,. \tag{4.35}$$

The one-step-ahead *mean squared prediction error* (MSPE) is given by

$$P_{n+1}^n = E[x_{n+1} - x_{n+1}^n]^2 = E[x_{n+1} - \phi x_n]^2 = E w_{n+1}^2 = \sigma_w^2.$$

The two-step-ahead forecast is obtained similarly. Since the model is

$$x_{n+2} = \phi x_{n+1} + w_{n+2}\,,$$

[2] Formally $x_t^n = E(x_t \mid x_1, \ldots, x_n)$ is conditional expectation, which is discussed in Section B.4.

we should have

$$x^n_{n+2} = \phi x^n_{n+1} + w^n_{n+2}\,.$$

Again, w_{n+2} is a future error, so $w^n_{n+2} = 0$. Also, we already know $x^n_{n+1} = \phi x_n$, so the forecast is

$$x^n_{n+2} = \phi x^n_{n+1} = \phi^2 x_n\,. \tag{4.36}$$

The two-step-ahead MSPE is given by

$$\begin{aligned} P^n_{n+2} &= E[x_{n+2} - x^n_{n+2}]^2 = E[\phi x_{n+1} + w_{n+2} - \phi^2 x_n]^2 \\ &= E[w_{n+2} + \phi(x_{n+1} - \phi x_n)]^2 = E[w_{n+2} + \phi w_{n+1}]^2 = \sigma_w^2(1+\phi^2). \end{aligned}$$

Generalizing these results, it is easy to see that the m-step-ahead forecast is.

$$x^n_{n+m} = \phi^m x_n\,, \tag{4.37}$$

with MSPE

$$P^n_{n+m} = E[x_{n+m} - x^n_{n+m}]^2 = \sigma_w^2(1 + \phi^2 + \cdots + \phi^{2(m-1)})\,. \tag{4.38}$$

for $m = 1, 2, \ldots$.

Note that since $|\phi| < 1$, we will have $\phi^m \to 0$ fast as $m \to \infty$. Thus the forecasts in (4.37) will soon go to zero (or the mean) and become useless. In addition, the MSPE will converge to $\sigma_w^2 \sum_{j=0}^{\infty} \phi^{2j} = \sigma_w^2/(1-\phi^2)$, which is the variance of the process x_t; recall (4.3).

Forecasting an AR(p) model is basically the same as forecasting an AR(1) provided the sample size n is larger than the order p, which it is most of the time. Since MA(q) and ARMA(p,q) are AR(∞) by invertibility, the same basic techniques can be used. Because ARMA models are invertible; i.e., $w_t = x_t + \sum_{j=1}^{\infty} \pi_j x_{t-j}$, we may write

$$x_{n+m} = -\sum_{j=1}^{\infty} \pi_j x_{n+m-j} + w_{n+m}.$$

If we had the infinite history $\{x_n, x_{n-1}, \ldots, x_1, x_0, x_{-1}, \ldots\}$, of the data available, we would predict x_{n+m} by

$$x^n_{n+m} = -\sum_{j=1}^{\infty} \pi_j x^n_{n+m-j}$$

successively for $m = 1, 2, \ldots$. In this case, $x^n_t = x_t$ for $t = n, n-1, \ldots$. We only have the actual data $\{x_n, x_{n-1}, \ldots, x_1\}$ available, but a practical solution is to truncate the forecasts as

$$x^n_{n+m} = -\sum_{j=1}^{n+m-1} \pi_j x^n_{n+m-j},$$

with $x^n_t = x_t$ for $1 \le t \le n$. For ARMA models in general, as long as n is large,

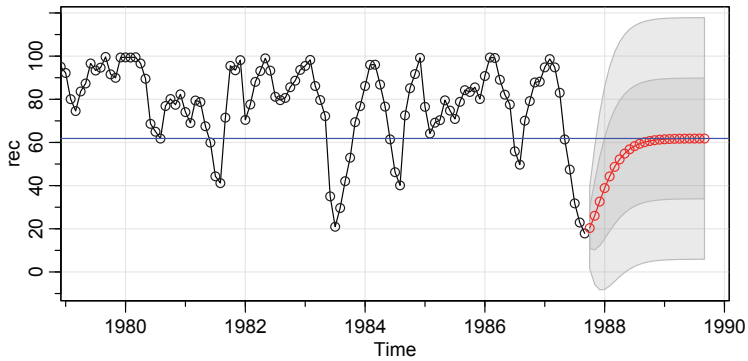


Figure 4.8 *Twenty-four-month forecasts for the Recruitment series. The actual data shown are from about January 1979 to September 1987, and then the forecasts plus and minus one and two standard error are displayed. The solid horizontal line is the estimated mean function.*

the approximation works well because the π-weights are going to zero exponentially fast. For large n, it can be shown (see Problem 4.10) that the mean squared prediction error for ARMA(p,q) models is approximately (exact if $q=0$)

$$P_{n+m}^{n} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2 . \tag{4.39}$$

We saw this result in (4.38) for the AR(1) because in that case, $\psi_j^2 = \phi^{2j}$.

Example 4.31. Forecasting the Recruitment Series
In Example 4.21 we fit an AR(2) model to the Recruitment series using OLS. Here, we use maximum likelihood estimation (MLE), which is similar to unconditional least squares for ARMA models:

```
sarima(rec, p=2, d=0, q=0)    # fit the model
       Estimate     SE  t.value p.value
 ar1     1.3512 0.0416  32.4933       0
 ar2    -0.4612 0.0417 -11.0687       0
 xmean  61.8585 4.0039  15.4494       0
```

The results are nearly the same as using OLS. Using the parameter estimates as the actual parameter values, the forecasts and root MSPEs can be calculated in a similar fashion to the introduction to this section.

Figure 4.8 shows the result of forecasting the Recruitment series over a 24-month horizon, $m = 1, 2, \ldots, 24$, obtained in R as

```
sarima.for(rec, n.ahead=24, p=2, d=0, q=0)
abline(h=61.8585, col=4)   # display estimated mean
```

Note how the forecast levels off to the mean quickly and the prediction intervals are wide and become constant. That is, because of the short memory, the forecasts settle

to the estimated mean, 61.86, and the root MSPE becomes quite large (and eventually settles at the standard deviation of all the data). ◊

Problems

4.1. For an MA(1), $x_t = w_t + \theta w_{t-1}$, show that $|\rho_x(1)| \leq 1/2$ for any number θ. For which values of θ does $\rho_x(1)$ attain its maximum and minimum?

4.2. Let $\{w_t; t = 0, 1, \dots\}$ be a white noise process with variance σ_w^2 and let $|\phi| < 1$ be a constant. Consider the process $x_0 = w_0$, and

$$x_t = \phi x_{t-1} + w_t, \quad t = 1, 2, \dots .$$

We might use this method to simulate an AR(1) process from simulated white noise.

(a) Show that $x_t = \sum_{j=0}^{t} \phi^j w_{t-j}$ for any $t = 0, 1, \dots$.

(b) Find the $E(x_t)$.

(c) Show that, for $t = 0, 1, \dots$,

$$\text{var}(x_t) = \frac{\sigma_w^2}{1-\phi^2}(1 - \phi^{2(t+1)})$$

(d) Show that, for $h \geq 0$,

$$\text{cov}(x_{t+h}, x_t) = \phi^h \text{var}(x_t)$$

(e) Is x_t stationary?

(f) Argue that, as $t \to \infty$, the process becomes stationary, so in a sense, x_t is "asymptotically stationary."

(g) Comment on how you could use these results to simulate n observations of a stationary Gaussian AR(1) model from simulated iid N(0,1) values.

(h) Now suppose $x_0 = w_0/\sqrt{1-\phi^2}$. Is this process stationary? *Hint*: Show $\text{var}(x_t)$ is constant.

4.3. Consider the following two models:

(i) $x_t = .80x_{t-1} - .15x_{t-2} + w_t - .30w_{t-1}$.

(ii) $x_t = x_{t-1} - .50x_{t-2} + w_t - w_{t-1}$.

(a) Using Example 4.10 as a guide, check the models for parameter redundancy. If a model has redundancy, find the reduced form of the model.

(b) A way to tell if an ARMA model is causal is to examine the roots of AR term $\phi(B)$ to see if there are no roots less than or equal to one in magnitude. Likewise, to determine invertibility of a model, the roots of the MA term $\theta(B)$ must not be less than or equal to one in magnitude. Use Example 4.11 as a guide to determine if the reduced (if appropriate) models (i) and (ii), are causal and/or invertible.

17

STAT 626: Outline of Lecture ~~22~~

The ARIMA (p, d, q) Model Building Process (§5.2)

1. **Plot the Data, Transform to Stationarity if Necessary,**
   Select the Differencing Order d.
2. **Model Formulation: Use the ACF and PACF to Select p, q:**
   ARIMA(p, d, q).
3. **Model Estimation: Find the MLE of the $p + q + 1$ Parameters**
4. **Model Diagnostic: Check the Residuals for Independence**
5. **If Not Happy, Go to Step 2 and Repeat the PROCESS**
6. **Choose from the Competing Models Using AIC/BIC**

7. Review of HWs on the NEGATIVE role of NONSTATIONARITY.

Example 5.6: Analysis of GNP Data

Example 5.8 Diagnostics for the Glacial Varve Series

ALL Models Are Wrong, But SOME Are Useful.

Who said the above?

STAT 626: Review of Past Lectures

1. Forecating: Begins when a good model is identified for the time series,

2. Given the time series data $x_1, \ldots, x_n$: **What are the principles for model-based forecasting ?**

$$x_t = f(\beta, \text{Past of the Series}) + w_t.$$

Example:

$$x_t = \phi x_{t-1} + w_t.$$

Principle: Replace the unknowns by the best ESTIMATES.

Example:

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

3. **Forecasting ARMA Models**

Recall that causal ARMA models can be written as One-Sided MA(∞) of a white noise:

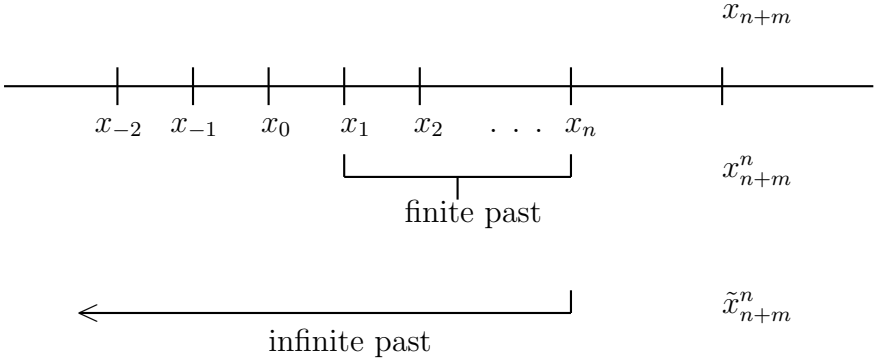
$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

and **invertible** ARMA models can be written as One-Sided AR(∞);

$$x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} + w_t.$$

Forecasting

4. A pictorial setup for forecasting the future values $x_{n+m}, m = 1, 2, \ldots$:



5. What are their forecasts, forecast error, and forecast error variances?

   Forecast error: $x_{n+m} - x^n_{n+m}$

   Error variance: $P^n_{n+m} = \mathrm{Var}(x_{n+m} - x^n_{n+m})$

6. Their 95% forecast intervals?

$$x^n_{n+m} \pm 1.96\sqrt{P^n_{n+m}}.$$

Chapter 5

ARIMA Models

5.1 Integrated Models

Adding nonstationary to ARMA models leads to the *autoregressive integrated moving average* (ARIMA) model popularized by Box and Jenkins (1970). Seasonal data, such as the data discussed in Example 1.1 and Example 1.4 lead to seasonal autoregressive integrated moving average (SARIMA) models.

In previous chapters, we saw that if x_t is a random walk, $x_t = x_{t-1} + w_t$, then by differencing x_t, we find that $\nabla x_t = w_t$ is stationary. In many situations, time series can be thought of as being composed of two components, a nonstationary trend component and a zero-mean stationary component. For example, in Section 3.1 we considered the model

$$x_t = \mu_t + y_t, \tag{5.1}$$

where $\mu_t = \beta_0 + \beta_1 t$ and y_t is stationary. Differencing such a process will lead to a stationary process:

$$\nabla x_t = x_t - x_{t-1} = \beta_1 + y_t - y_{t-1} = \beta_1 + \nabla y_t.$$

Another model that leads to first differencing is the case in which μ_t in (5.1) is stochastic and slowly varying according to a random walk. That is,

$$\mu_t = \mu_{t-1} + v_t$$

where v_t is stationary and uncorrelated with y_t. In this case,

$$\nabla x_t = v_t + \nabla y_t,$$

is stationary.

On a rare occasion, the differenced data ∇x_t may still have linear trend or random walk behavior. In this case, it may be appropriate to difference the data again, $\nabla(\nabla x_t) = \nabla^2 x_t$. For example, if μ_t in (5.1) is quadratic, $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$, then the twice differenced series $\nabla^2 x_t$ is stationary.

The *integrated* ARMA, or ARIMA, model is a broadening of the class of ARMA models to include differencing. The basic idea is that if differencing the data at some order d produces an ARMA process, then the original process is said to be ARIMA. Recall that the difference operator

Definition 5.1. *A process x_t is said to be* **ARIMA**(p,d,q) *if*

$$\nabla^d x_t = (1-B)^d x_t$$

is ARMA(p,q). In general, we will write the model as

$$\phi(B)(1-B)^d x_t = \alpha + \theta(B) w_t\,, \tag{5.2}$$

where $\alpha = \delta(1-\phi_1-\cdots-\phi_p)$ and $\delta = E(\nabla^d x_t)$.

Estimation for ARIMA models is the same as for ARMA models except that the data are differenced first. For example, if $d = 1$, we fit an ARMA model to $\nabla x_t = x_t - x_{t-1}$ instead of x_t.

Example 5.2. Fitting the Glacial Varve Series (cont.)
In Example 4.28, we fit an MA(1) to the differenced logged varve series as using the commands as follows:

```
sarima(diff(log(varve)), p=0, d=0, q=1, no.constant=TRUE)
```

Equivalently, we can fit an ARIMA$(0,1,1)$ to the logged series:

```
sarima(log(varve), p=0, d=1, q=1, no.constant=TRUE)
 Coefficients:
          ma1
       -0.7705
 s.e.   0.0341
 sigma^2 estimated as 0.2353
```

The results are identical to Example 4.28. The only difference will be when we forecast. In Example 4.28 we would get forecasts of $\nabla \log x_t$ and in this example we would get forecasts for $\log x_t$, where x_t represents the varve series. ◊

Forecasting ARIMA is also similar to the ARMA case, but needs some additional consideration. Since $y_t = \nabla^d x_t$ is ARMA, we can use Section 4.4 methods to obtain forecasts of y_t, which in turn lead to forecasts for x_t. For example, if $d = 1$, given forecasts y^n_{n+m} for $m = 1,2,\ldots$, we have $y^n_{n+m} = x^n_{n+m} - x^n_{n+m-1}$, so that

$$x^n_{n+m} = y^n_{n+m} + x^n_{n+m-1}$$

with initial condition $x^n_{n+1} = y^n_{n+1} + x_n$ (noting $x^n_n = x_n$).

It is a little more difficult to obtain the prediction errors P^n_{n+m}, but for large n, the approximation (4.39) works well. That is, the mean-squared prediction error (MSPE) can be approximated by

$$P^n_{n+m} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2, \tag{5.3}$$

where ψ_j is the coefficient of z^j in $\psi(z) = \theta(z)/\phi(z)(1-z)^d$; Section D.2 has more details on how the ψ-weights are determined.

To better understand forecasting integrated models, we examine the properties of some simple cases.

Example 5.3. Random Walk with Drift
To fix ideas, we begin by considering the random walk with drift model first presented in Example 1.10, that is,

$$x_t = \delta + x_{t-1} + w_t,$$

for $t = 1,2,\ldots$, and $x_0 = 0$. Given data $x_1,\ldots,x_n$, the one-step-ahead forecast is given by

$$x^n_{n+1} = \delta + x^n_n + w^n_{n+1} = \delta + x_n\,.$$

The two-step-ahead forecast is given by $x^n_{n+2} = \delta + x^n_{n+1} = 2\delta + x_n$, and consequently, the m-step-ahead forecast, for $m = 1,2,\ldots$, is

$$x^n_{n+m} = m\,\delta + x_n, \tag{5.4}$$

To obtain the forecast errors, it is convenient to recall equation (1.4) wherein $x_n = n\,\delta + \sum_{j=1}^n w_j$. In this case we may write

$$x_{n+m} = (n+m)\,\delta + \sum_{j=1}^{n+m} w_j = m\,\delta + x_n + \sum_{j=n+1}^{n+m} w_j\,. \tag{5.5}$$

Using the difference of (5.5) and (5.4), it follows that the m-step-ahead prediction error is given by

$$P^n_{n+m} = E(x_{n+m} - x^n_{n+m})^2 = E\Big(\sum_{j=n+1}^{n+m} w_j\Big)^2 = m\,\sigma_w^2. \tag{5.6}$$

Unlike the stationary case, as the forecast horizon grows, the prediction errors, (5.6), increase without bound and the forecasts follow a straight line with slope δ emanating from x_n.

We note that (5.3) is exact in this case because the ψ-weights for this model are all equal to one. Thus, the MSPE is $P^n_{n+m} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2 = m\sigma_w^2$.

```
ARMAtoMA(ar=1, ma=0, 20)  # ψ-weights
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

◊

Example 5.4. Forecasting an ARIMA(1,1,0)
To get a better idea of what forecasts for ARIMA models will look like, we generated 150 observations from an ARIMA(1,1,0) model,

$$\nabla x_t = .9\nabla x_{t-1} + w_t\,.$$

Alternately, the model is $x_t - x_{t-1} = .9(x_{t-1} - x_{t-2}) + w_t$, or

$$x_t = 1.9x_{t-1} - .9x_{t-2} + w_t\,.$$

Although this form looks like an AR(2), the model is not causal in x_t and therefore not an AR(2). As a check, notice that the ψ-weights do not converge to zero (and in

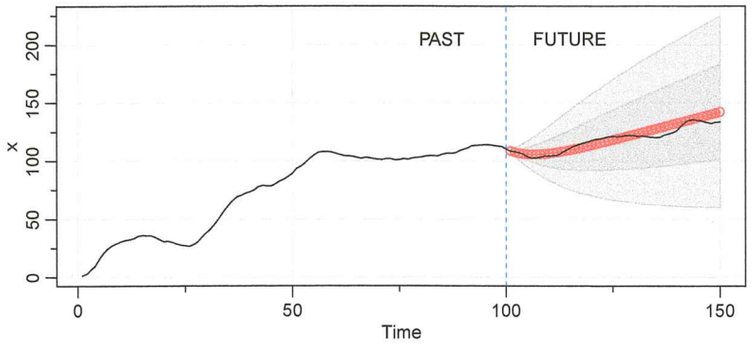


Figure 5.1 *Output for Example 5.4: Simulated ARIMA(1, 1, 0) series (solid line) with out of sample forecasts (points) and error bounds (gray area) based on the first 100 observations.*

```
round( ARMAtoMA(ar=c(1.9,-.9), ma=0, 60), 1 )
 [1]  1.9  2.7  3.4  4.1  4.7  5.2  5.7  6.1  6.5  6.9  7.2  7.5
[13]  7.7  7.9  8.1  8.3  8.5  8.6  8.8  8.9  9.0  9.1  9.2  9.3
[25]  9.4  9.4  9.5  9.5  9.6  9.6  9.7  9.7  9.7  9.7  9.8  9.8
[37]  9.8  9.8  9.9  9.9  9.9  9.9  9.9  9.9  9.9  9.9  9.9  9.9
[49]  9.9 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.0 10.0
```

We used the first 100 (of 150) generated observations to estimate a model and then predicted out-of-sample, 50 time units ahead. The results are displayed in Figure 5.1 where the solid line represents all the data, the points represent the forecasts, and the gray areas represent ± 1 and ± 2 root MSPEs. Note that, unlike the forecasts of an ARMA model from the previous chapter, the error bounds continue to increase.

The R code to generate Figure 5.1 is below. Note that `sarima.for` fits an ARIMA model and then does the forecasting out to a chosen horizon. In this case, `x` is the entire time series of 150 points, whereas `y` is only the first 100 values of `x`.

```
set.seed(1998)
x <- ts(arima.sim(list(order = c(1,1,0), ar=.9), n=150)[-1])
y <- window(x, start=1, end=100)
sarima.for(y, n.ahead = 50, p = 1, d = 1, q = 0, plot.all=TRUE)
text(85, 205, "PAST"); text(115, 205, "FUTURE")
abline(v=100, lty=2, col=4)
lines(x)
```

◊

Example 5.5. IMA(1,1) and EWMA

The ARIMA(0,1,1), or IMA(1,1) model is of interest because many economic time series can be successfully modeled this way. The model leads to a frequently used method called exponentially weighted moving average (EWMA). We will write the model as

$$x_t = x_{t-1} + w_t - \lambda w_{t-1} \tag{5.7}$$

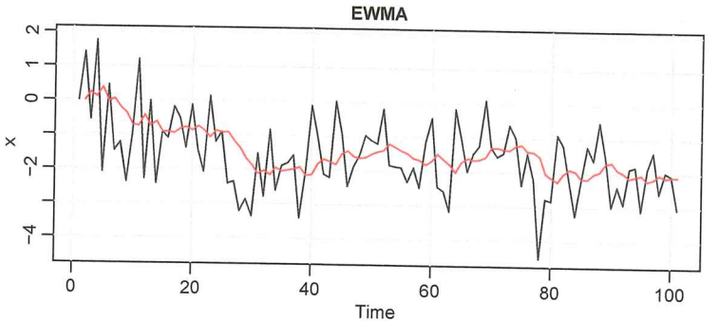


Figure 5.2 *Output for Example 5.5: Simulated data with an EWMA superimposed.*

with $|\lambda| < 1$, because this model formulation is easier to work with here, and it leads to the standard representation for EWMA.

In this case, the one-step-ahead predictor is

$$x^n_{n+1} = (1-\lambda)x_n + \lambda x^{n-1}_n . \tag{5.8}$$

That is, the predictor is a linear combination of the present value of the process, x_n, and the prediction of the present, x^{n-1}_n. Details are given in Problem 5.17. This method of forecasting is popular because it is easy to use; we need only retain the previous forecast value and the current observation to forecast the next time period. EWMA is widely used, for example in control charts (Shewhart, 1931), and economic forecasting (Winters, 1960) whether or not the underlying dynamics are IMA(1,1).

The MSPE is given by

$$P^n_{n+m} \approx \sigma^2_w[1 + (m-1)(1-\lambda)^2] . \tag{5.9}$$

In EWMA, the parameter $1-\lambda$ is often called the smoothing parameter, is denoted by α, and is restricted to be between zero and one. Larger values of λ (or smaller values of α) lead to smoother forecasts.

In the following, we show how to generate 100 observations from an IMA(1,1) model with $\alpha = 1-\lambda = .2$ and then calculate and display the fitted EWMA superimposed on the data. This can accomplished using the Holt-Winters command in R (see the help file `?HoltWinters` for details). This and related techniques are generally called *exponential smoothing*; the ideas were made popular in the late 1950s and are still used today. To reproduce Figure 5.2, use the following.

```
set.seed(666)
x = arima.sim(list(order = c(0,1,1), ma = -0.8), n = 100)
(x.ima = HoltWinters(x, beta=FALSE, gamma=FALSE))  # α below is 1−λ
  Smoothing parameter:  alpha:  0.1663072
plot(x.ima, main="EWMA")
```

◊

5.2 Building ARIMA Models

There are a few basic steps to fitting ARIMA models to time series data. These steps involve

- plotting the data,
- possibly transforming the data,
- identifying the dependence orders of the model,
- parameter estimation,
- diagnostics, and
- model choice.

First, as with any data analysis, construct a time plot of the data and inspect the graph for any anomalies. It may be of interest to transform the data and as we have seen in numerous examples, if the data behave as $x_t = (1 + r_t)x_{t-1}$, where r_t is a stable process of small percent changes, then $\nabla \log(x_t) \approx r_t$ will be stable. This general idea was used in Example 4.27, and we will use it again in Example 5.6.

After suitably transforming the data, the next step is to identify preliminary values of the autoregressive order, p, the order of differencing, d, and the moving average order, q. A time plot of the data will typically suggest whether any differencing is needed. If differencing is called for, then difference the data once, $d = 1$, and inspect the time plot of ∇x_t. If additional differencing is necessary, then try differencing again and inspect a time plot of $\nabla^2 x_t$; *it is rare for d to be bigger than 1*. Be careful not to overdifference because this may introduce dependence where none exists. For example, $x_t = w_t$ is serially uncorrelated, but $\nabla x_t = w_t - w_{t-1}$ is a non-invertible MA(1). In addition to time plots, the sample ACF can help in indicating whether differencing is needed. A slow (linear) decay in the ACF is an indication that differencing may be needed.

When preliminary values of d have been chosen (including no differencing, $d = 0$), the next step is to look at the sample ACF and PACF of $\nabla^d x_t$. Using Table 4.1 as a guide, preliminary values of p and q are chosen. Note that it cannot be the case that both the ACF and PACF cut off. Because we are dealing with estimates, it will not always be clear whether the sample ACF or PACF is tailing off or cutting off. Also, two models that are seemingly different can actually be very similar. It is a good idea to *start small* and up the orders slowly. Also, watch out for parameter redundancy and do not increase p and q at the same time. At this point, a few preliminary values of p, d, and q should be at hand, and we can start estimating the parameters and performing diagnostics and model choice.

Example 5.6. Analysis of GNP Data

In this example, we consider the analysis of quarterly U.S. GNP from 1947(1) to 2002(3), $n = 223$ observations. The data are real U.S. gross national product in billions of chained 1996 dollars and have been seasonally adjusted. Figure 5.3 shows a plot of the data, say, y_t. Because strong trend tends to obscure other effects, it is difficult to see any other variability in data except for periodic large dips in

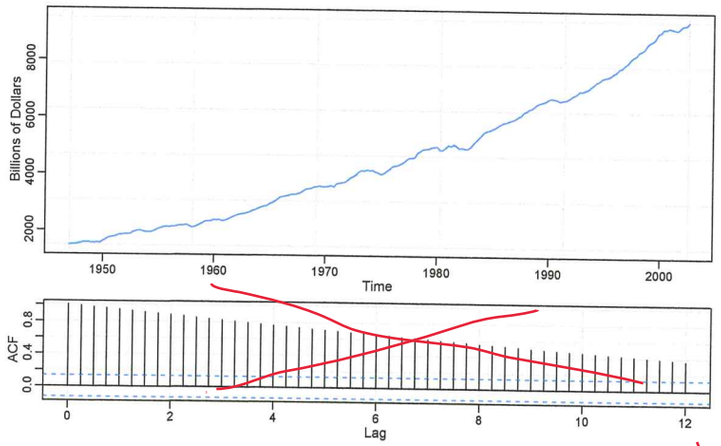


Figure 5.3 Top: *Quarterly U.S. GNP from 1947(1) to 2002(3).* Bottom: *Sample ACF of the GNP data. Lag is in terms of years.*

the economy. Typically, GNP and similar economic indicators are given in terms of growth rate (percent change) rather than in actual values. The growth rate, say $x_t = \nabla \log(y_t)$, is plotted in Figure 5.4 and it appears to be a stable process.

```
##-- Figure 5.3 --##
layout(1:2, heights=2:1)
tsplot(gnp, col=4)
acf1(gnp, main="")
##-- Figure 5.4 --##
tsplot(diff(log(gnp)), ylab="GNP Growth Rate", col=4)
abline(mean(diff(log(gnp))), col=6)
##-- Figure 5.5 --##
acf2(diff(log(gnp)), main="")
```

The sample ACF and PACF of the quarterly growth rate are plotted in Figure 5.5. Inspecting the sample ACF and PACF, we might feel that the ACF is cutting off at lag 2 and the PACF is tailing off. This would suggest the GNP growth rate follows an MA(2) process, or log GNP follows an ARIMA$(0, 1, 2)$ model.

The MA(2) fit to the growth rate, x_t, is

$$\hat{x}_t = .008_{(.001)} + .303_{(.065)}\hat{w}_{t-1} + .204_{(.064)}\hat{w}_{t-2} + \hat{w}_t, \qquad (5.10)$$

where $\hat{\sigma}_w = .0094$ is based on 219 degrees of freedom.

```
sarima(diff(log(gnp)), 0,0,2)    # MA(2) on growth rate
       Estimate     SE t.value p.value
ma1      0.3028 0.0654  4.6272  0.0000
ma2      0.2035 0.0644  3.1594  0.0018
xmean    0.0083 0.0010  8.7178  0.0000
sigma^2 estimated as 8.919e-05
```

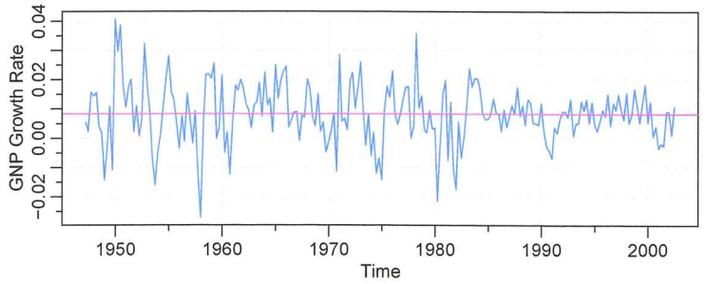


Figure 5.4 *U.S. GNP quarterly growth rate. The horizontal line displays the average growth of the process, which is close to 1%.*

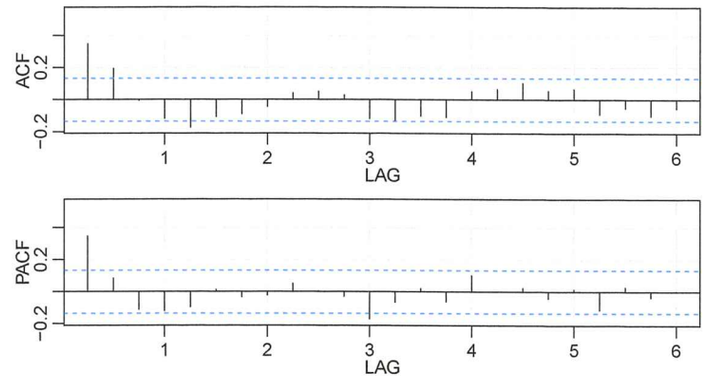


Figure 5.5 *Sample ACF and PACF of the GNP quarterly growth rate. Lag is in years.*

We note that `sarima(log(gnp), p=0, d=1, q=2)` will produce the same results.

All of the regression coefficients are significant, including the constant. *We make a special note of this because, as a default, some computer packages—including the* R *stats package—do not fit a constant in a differenced model, assuming without reason that there is no drift.* In this example, not including a constant leads to the wrong conclusions about the nature of the U.S. economy. Not including a constant assumes the average quarterly growth rate is zero, whereas the U.S. GNP average quarterly growth rate is about 1% (which can be seen easily in Figure 5.4).

Rather than focus on one model, we will also suggest that it appears that the ACF is tailing off and the PACF is cutting off at lag 1. This suggests an AR(1) model for the growth rate, or ARIMA$(1,1,0)$ for log GNP. The estimated AR(1) model is

$$\hat{x}_t = .008_{(.001)}\,(1 - .347) + .347_{(.063)}x_{t-1} + \hat{w}_t, \tag{5.11}$$

where $\hat{\sigma}_w = .0095$ on 220 degrees of freedom; note that the constant in (5.11) is $.008\,(1 - .347) = .005$.


```
sarima(diff(log(gnp)), 1,0,0)    # AR(1) on growth rate
        Estimate     SE t.value p.value
  ar1     0.3467 0.0627  5.5255       0
  xmean   0.0083 0.0010  8.5398       0
  sigma^2 estimated as 9.03e-05
```

As before, `sarima(log(gnp), p=1, d=1, q=0)` will produce the same results.

We will discuss diagnostics next, but assuming both of these models fit well, how are we to reconcile the apparent differences of the estimated models (5.10) and (5.11)? In fact, the fitted models are nearly the same. To show this, consider an AR(1) model of the form in (5.11) without a constant term; that is,

$$x_t = .35x_{t-1} + w_t,$$

and write it in its causal form, $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where we recall $\psi_j = .35^j$. Thus, $\psi_0 = 1, \psi_1 = .350, \psi_2 = .123, \psi_3 = .043, \psi_4 = .015, \psi_5 = .005, \psi_6 = .002, \psi_7 = .001, \psi_8 = 0, \psi_9 = 0, \psi_{10} = 0$, and so forth. The AR(1) model is approximately an MA(2) model,

$$x_t \approx .35w_{t-1} + .12w_{t-2} + w_t,$$

which is similar to the fitted MA(2) model in (5.10).

```
round( ARMAtoMA(ar=.35, ma=0, 10), 3)  # print psi-weights
 [1] 0.350 0.122 0.043 0.015 0.005 0.002 0.001 0.000 0.000 0.000
```

◊

The next step in model fitting is residual diagnostics. The first step involves a time plot of the *innovations* (or residuals), $x_t - \hat{x}_t^{t-1}$, or of the *standardized innovations*

$$e_t = \left(x_t - \hat{x}_t^{t-1}\right) / \sqrt{\hat{P}_t^{t-1}}, \tag{5.12}$$

where $\hat{x}_t^{t-1}$ is the one-step-ahead prediction of x_t based on the fitted model and $\hat{P}_t^{t-1}$ is the estimated one-step-ahead error variance. If the model fits well, the standardized residuals should behave as an independent normal sequence with mean zero and variance one. The time plot should be inspected for any obvious departures from this assumption. Investigation of marginal normality can be accomplished visually by inspecting a normal Q-Q plot.

We should also inspect the sample autocorrelations of the residuals, say $\hat{\rho}_e(h)$, for any patterns or large values. In addition to plotting $\hat{\rho}_e(h)$, we can perform a general test of whiteness that takes into consideration the magnitudes of $\hat{\rho}_e(h)$ as a group. The *Ljung–Box–Pierce Q-statistic* given by

$$Q = n(n+2)\sum_{h=1}^{H} \frac{\hat{\rho}_e^2(h)}{n-h} \tag{5.13}$$

can be used to perform such a test. The value H in (5.13) is chosen somewhat arbitrarily, but not too large. For large sample sizes, under the null hypothesis of model adequacy $Q \sim \chi^2_{H-p-q}$. Thus, we would reject the null hypothesis at level α if the value of Q exceeds the $(1-\alpha)$-quantile of the χ^2 distribution.

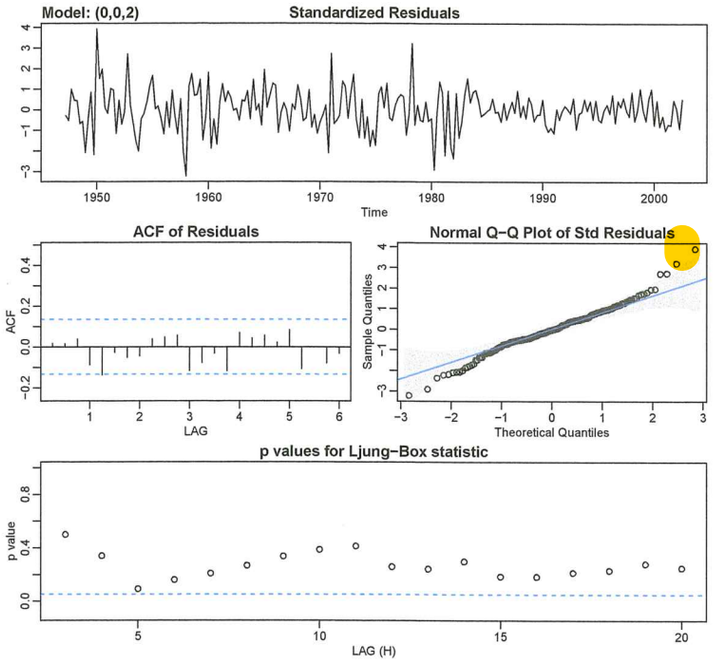


Figure 5.6 *Diagnostics of the residuals from MA(2) fit on GNP growth rate.*

Example 5.7. Diagnostics for GNP Growth Rate Example
We will focus on the MA(2) fit from Example 5.6; the analysis of the AR(1) residuals is similar. Figure 5.6 displays a plot of the standardized residuals, the ACF of the residuals, a Q-Q plot of the standardized residuals, and the p-values associated with the Q-statistic, (5.13). The residual analysis figure is generated as part of the call:

```
sarima(diff(log(gnp)), 0, 0, 2)   # MA(2) fit with diagnostics
```

You can turn off the diagnostics by adding `details=FALSE` in the `sarima` call.

Inspection of the time plot of the standardized residuals in Figure 5.6 shows no obvious patterns. Notice that there may be outliers because a few standardized residuals exceed 3 standard deviations in magnitude. However, there are no values that are exceedingly large in magnitude.

The ACF of the residuals shows no apparent departure from the model assumptions. The normal Q-Q plot of the residuals suggests that the assumption of normality is not unreasonable, however, there may be one large positive outlier.

Next, consider the Q-statistic. The graphic shows the p-values for the tests based on the lags $H = 3$ through $H = 20$ (with corresponding degrees of freedom $H - 2$). The dashed horizontal line on the bottom indicates the .05 level. The way to view this graphic is not as doing 17 highly dependent tests, but as another way to view the ACF of the residuals. In particular, the Q-statistic looks at the accumulation

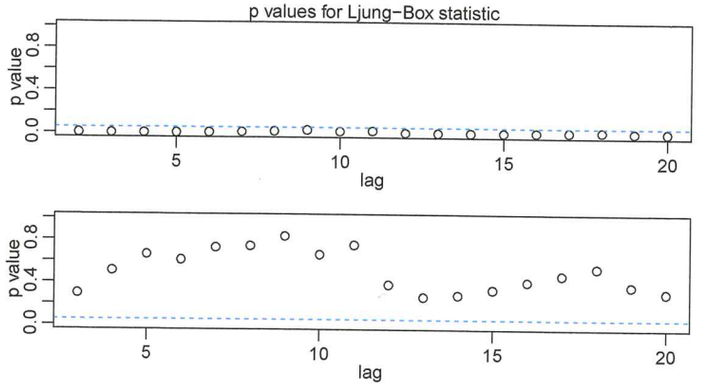


Figure 5.7 *Q-statistic p-values for the ARIMA*$(0, 1, 1)$ *fit (top) and the ARIMA*$(1, 1, 1)$ *fit (bottom) to the logged varve data.*

of autocorrelation rather than individual autocorrelations seen in the ACF. In this example all the p-values exceed .05, so we can feel comfortable not rejecting the null hypothesis that the residuals are white.

As a final check, we might consider overfitting a model to see if the results change significantly. For example, we might try the following,

```
sarima(diff(log(gnp)), 0, 0, 3)   # try an MA(2+1) fit (not shown)
sarima(diff(log(gnp)), 2, 0, 0)   # try an AR(1+1) fit (not shown)
```

and conclude that the extra parameter does not significantly change the results. ◊

Example 5.8. Diagnostics for the Glacial Varve Series
In Example 5.2, we fit an ARIMA$(0, 1, 1)$ model to the logarithms of the glacial varve data and there appears to be a small amount of autocorrelation left in the residuals and the Q-tests are all significant; see Figure 5.7.

To adjust for the small amount of autocorrelation left by the model, we added an AR parameter to the mix and fit an ARIMA$(1, 1, 1)$ to the logged varve data.

```
sarima(log(varve), 0, 1, 1, no.constant=TRUE)   # ARIMA(0,1,1)
sarima(log(varve), 1, 1, 1, no.constant=TRUE)   # ARIMA(1,1,1)
     Estimate      SE   t.value p.value
ar1   0.2330  0.0518    4.4994       0
ma1  -0.8858  0.0292  -30.3861       0
sigma^2 estimated as 0.2284
```

Hence the additional AR term is significant. The Q-statistic p-values for this model are also displayed in Figure 5.7, and it appears this model fits the data well.

As previously stated, the diagnostics are byproducts of the individual `sarima` runs. We note that we did not fit a constant in either model because there is no apparent drift in the differenced, logged varve series. This fact can be verified by noting the constant is not significant when the command `no.constant=TRUE` is removed in the code. ◊

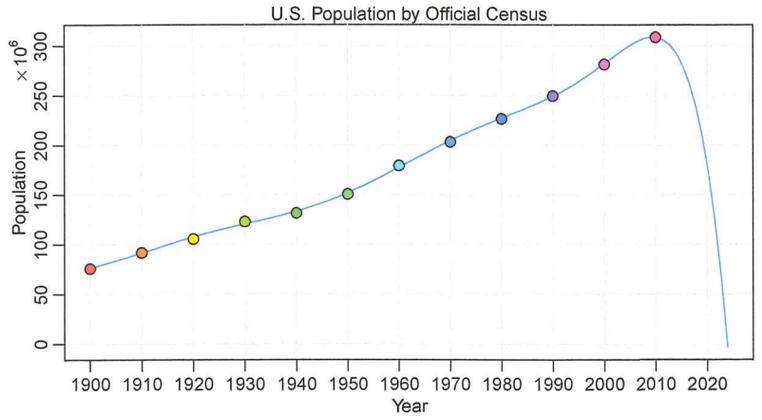


Figure 5.8 *A near perfect fit and a terrible forecast.*

In Example 5.6, we have two competing models, an AR(1) and an MA(2) on the GNP growth rate, that each appear to fit the data well. In addition, we might also consider that an AR(2) or an MA(3) might do better for forecasting. Perhaps combining both models, that is, fitting an ARMA(1, 2) to the GNP growth rate, would be the best. As previously mentioned, we have to be concerned with *overfitting* the model; it is not always the case that more is better. Overfitting leads to less-precise estimators, and adding more parameters may fit the data better but may also lead to bad forecasts. This result is illustrated in the following example.

Example 5.9. A Near Perfect Fit and a Terrible Forecast

Figure 5.8 shows the U.S. population by official census, every ten years from 1900 to 2010, as points. If we use these observations to predict the future population, we can fit a high degree polynomial so that the fit will be nearly perfect. There are twelve observations, so we could use an eight-degree polynomial to get a near perfect fit. The model in this case is

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_8 t^8 + w_t.$$

The fitted line is also plotted in Figure 5.8 and it nearly passes through all the observations ($R^2 = 99.97\%$). The model predicts that the population of the United States will cross zero before 2025! This may or may not be true.

The R code to reproduce these results is as follows. We note that the data are not in `astsa` and there is a different R data set called `uspop`.

```
uspop = c(75.995, 91.972, 105.711, 123.203, 131.669,150.697,
          179.323, 203.212, 226.505, 249.633, 281.422, 308.745)
uspop = ts(uspop, start=1900, freq=.1)
t = time(uspop) - 1955
reg = lm( uspop~ t+I(t^2)+I(t^3)+I(t^4)+I(t^5)+I(t^6)+I(t^7)+I(t^8) )
  Multiple R-squared:  0.9997
```



```
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

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)
```

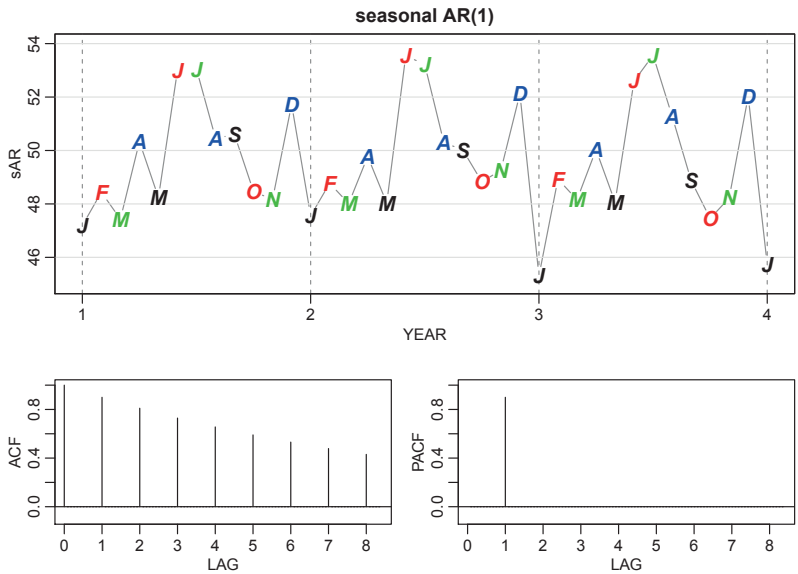


Figure 5.9 *Data generated from an SAR(1)$_{12}$ model, and the true ACF and PACF of the model* $(x_t - 50) = .9(x_{t-12} - 50) + w_t$. *LAG is in terms of seasons.*

```
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

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

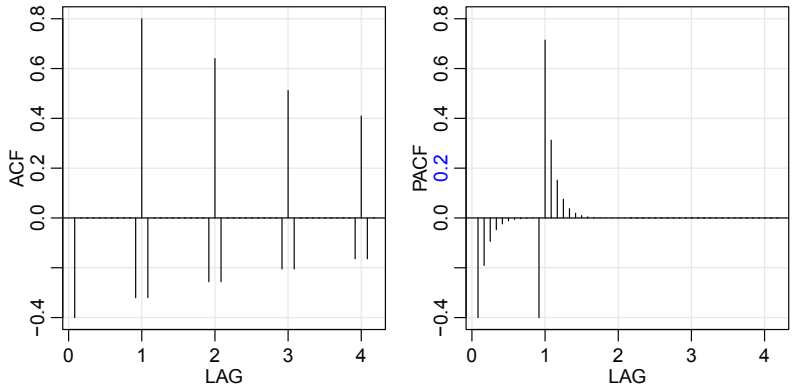


Figure 5.10 *ACF and PACF of the mixed seasonal ARMA model* $x_t = .8x_{t-12} + w_t - .5w_{t-1}$.

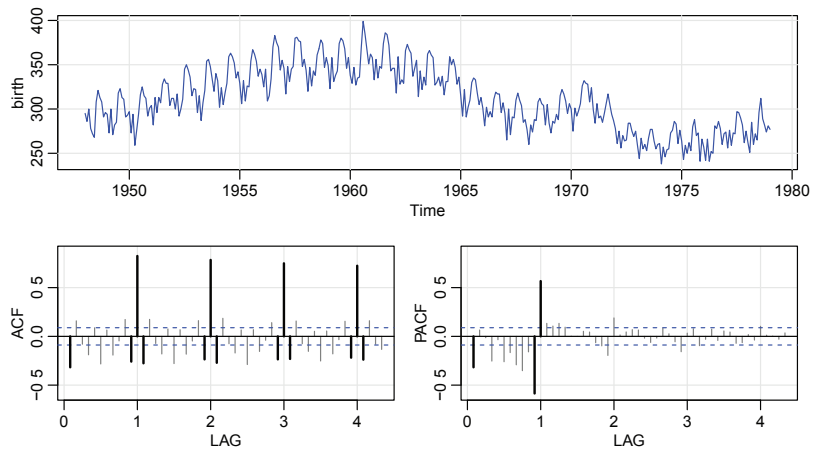


Figure 5.11 *Monthly live births in thousands for the United States during the "baby boom," 1948–1979. Sample ACF and PACF of the data with certain lags highlighted.*

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

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.

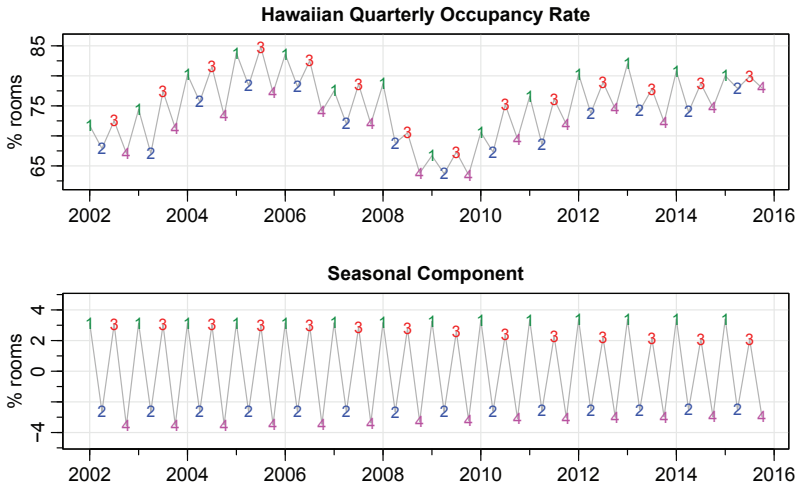


Figure 5.12 *Seasonal persistence: The quarterly occupancy rate of Hawaiian hotels and the extracted seasonal component, say $S_t \approx S_{t-4}$, where t is in quarters.*

```
##-- 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

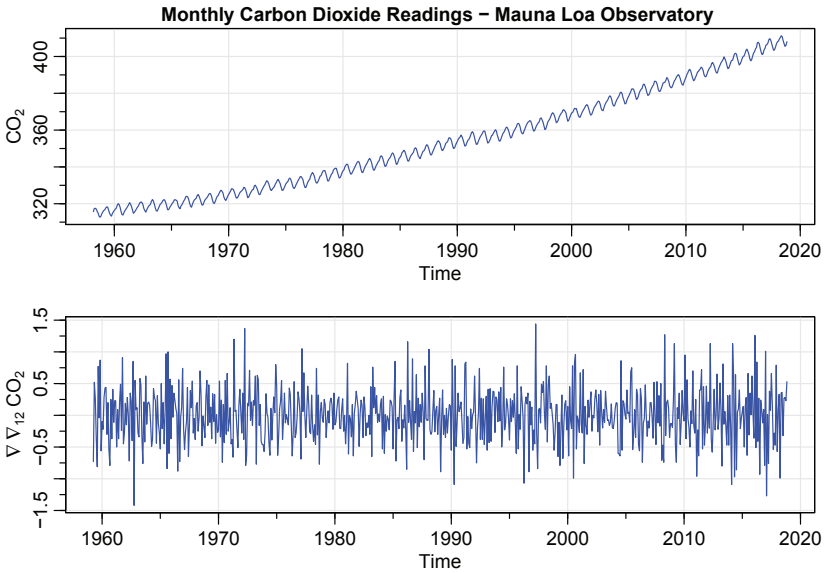


Figure 5.13 *Monthly CO_2 levels (ppm) taken at the Mauna Loa, Hawaii observatory (top) and the data differenced to remove trend and seasonal persistence (bottom).*

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.

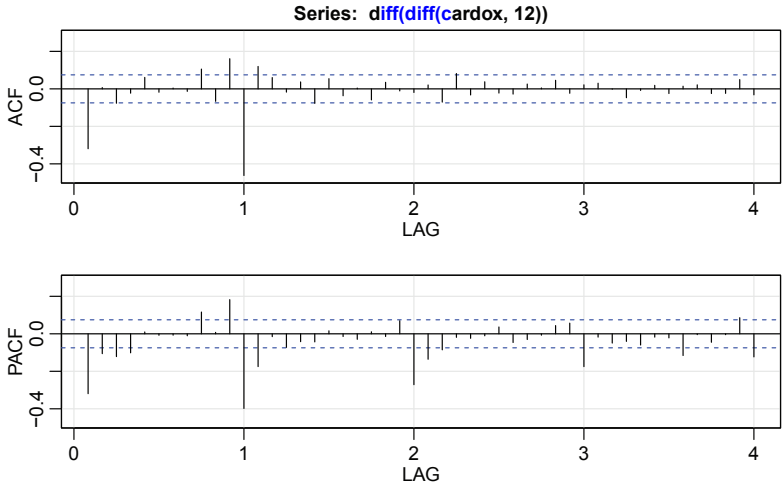


Figure 5.14 *Sample ACF and PACF of the differenced CO_2 data.*

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.

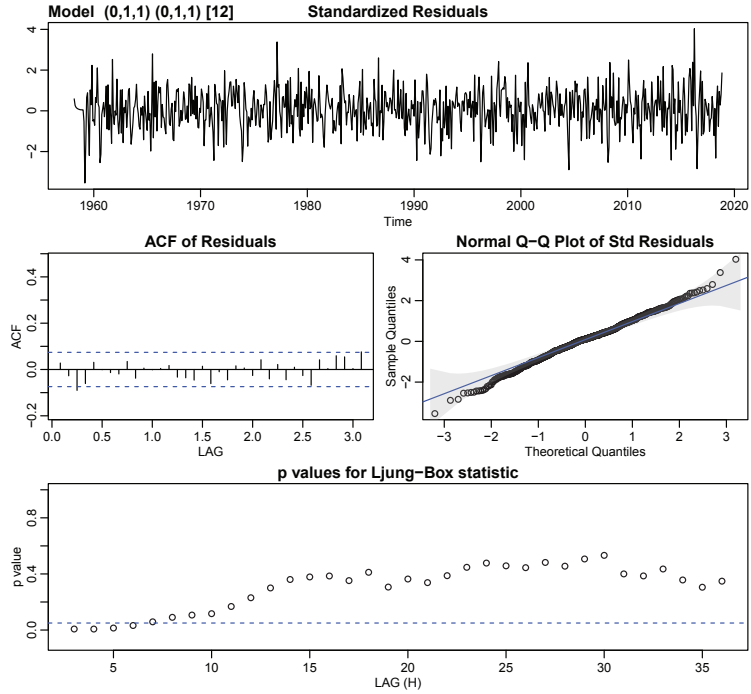


Figure 5.15 *Residual analysis for the ARIMA*$(0,1,1) \times (0,1,1)_{12}$ *fit to the* CO_2 *data set.*

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.

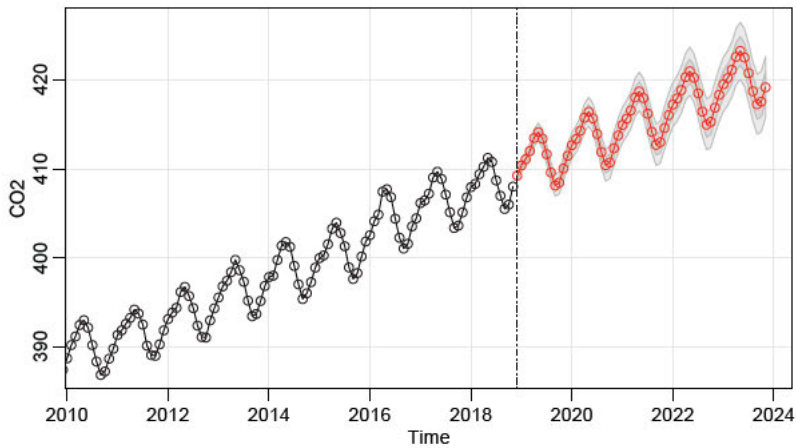


Figure 5.16 *Five-year-ahead forecasts using the ARIMA*$(1,1,1) \times (0,1,1)_{12}$ *model on the Mauna Loa carbon dioxide readings.*

```
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.

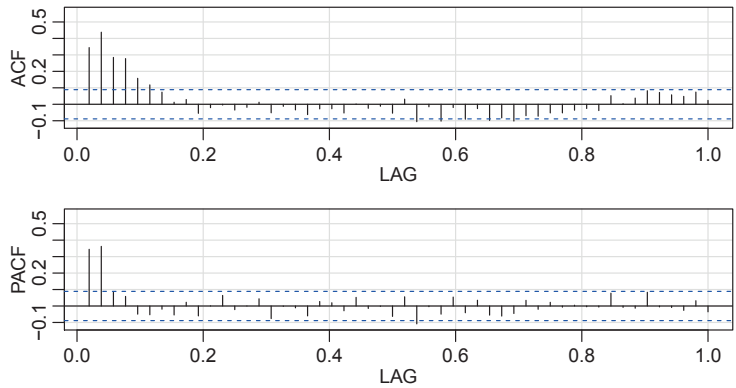


Figure 5.17 *Sample ACF and PACF of the mortality residuals indicating an AR(2) process.*

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

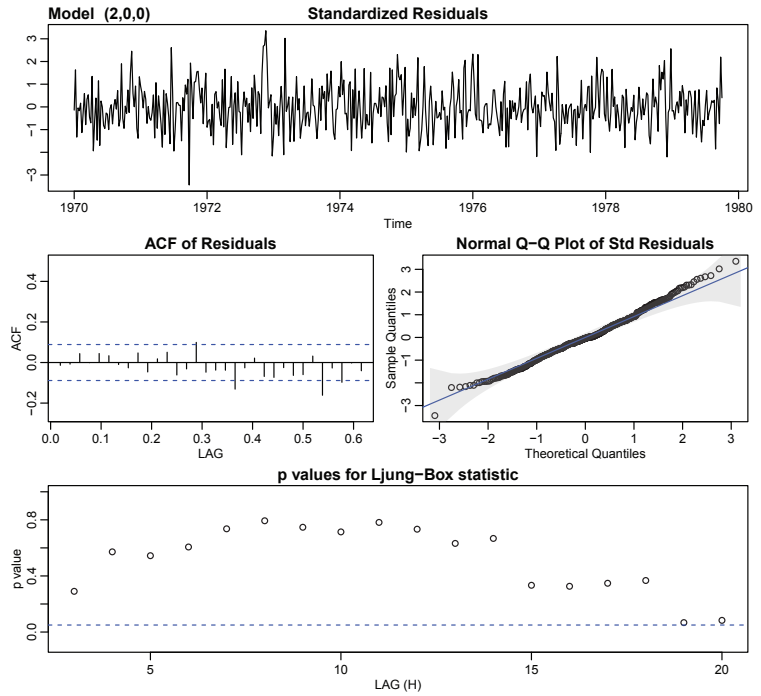


Figure 5.18 *Diagnostics for the regression of mortality on temperature and particulate pollution with autocorrelated errors, Example 5.16.*

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

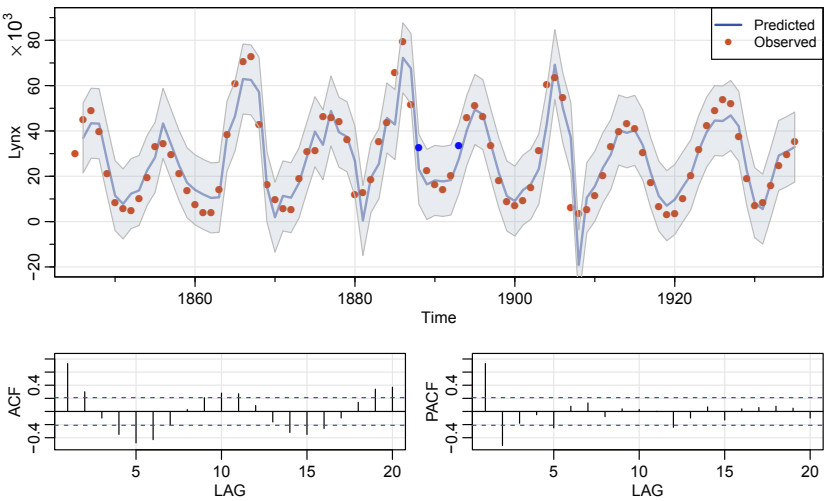


Figure 5.19 *Top: Observed lynx population size (points) and one-year-ahead prediction (line) with ±2 root MSPE (ribbon). Bottom: ACF and PACF of the residuals from* (5.23).

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).

STAT 626: Outline Lecture 219

ARCH-GARCH Models (§8.1)

1. **WN ⇒ ARMA, ARIMA, SARIMA, ARCH/GARCH, ...**

2. **Taking Care of Time-Varying Variances:** σ_t^2

3. **Time Series Decomposition:** $x_t = \mu_t + \sigma_t \varepsilon_t$, $\quad$ **Var**$(\sigma_t \varepsilon_t) = \sigma_t^2$.

4. **How to Model Time-Varying Variances?**
   Recall that Squared Residuals r_t^2 are Reasonable "Estimates" of σ_t^2:

$$r_t^2 \approx \sigma_t^2.$$

5. **Often r_t^2's appear more correlated than r_t's (Granger, 1970's).**

6. **AutoRegressive Conditionally Heteroscedastic (ARCH) Models:**(Engle, 1982)

$$r_t = \sigma_t \varepsilon_t,$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2.$$

   AR Models for Squared Residuals r_t^2.

   This point of view is helpful in using the ACF and PACF of the series r_t^2 to identify the orders of the ARCH(p) models.

7. **Generalized ARCH (GARCH) Models**

   ARMA Models for Squared Residuals r_t^2.

Unit-Root Test and Random Walks

8. **Random Walk vs AR(1):** $x_t = \phi x_{t-1} + w_t$,

$$H_0 : \phi = 1 \quad \text{vs} \quad H_1 : |\phi| < 1.$$

9. **Unit-Root Tests: DF, ADF, PP**.

10. **Why Unit-Root Test is Important in Economics and Finance?**

L. Bachelier Dissertation (1900).

Random Walk Hypothesis,

Efficient Market Hypothesis:

The weak form: All information about market prices is already reflected in the current stock price.

The strong form: All publicly available information about a company is already reflected in its stock price.

I. A Random Walk Down Wall Street, by Burton G. Malkiel

II. A Non-Random Walk Down Wall Street, by Andrew W. Lo & A. Craig MacKinlay

Books By Peter Bernstein:

III. Capital Ideas: The Improbable Origins of Modern Wall Street, (Free Press), 1991.

IV. Against the Gods: The Remarkable Story of Risk, (John Wiley & Son), 1996, Story of (Random Walk) Brownian Motion and how it Entered the World of Finance.

Chapter 8

Additional Topics *

In this chapter, we present special topics in the time domain. The sections may be read in any order. Each topic depends on a basic knowledge of ARMA models, forecasting and estimation, which is the material covered in Chapter 4 and Chapter 5.

8.1 GARCH Models

Various problems such as option pricing in finance have motivated the study of the *volatility*, or variability, of a time series. ARMA models were used to model the conditional mean (μ_t) of a process when the conditional variance (σ_t^2) was constant. For example, in the AR(1) model $x_t = \phi_0 + \phi_1 x_{t-1} + w_t$ we have

$$\mu_t = E(x_t \mid x_{t-1}, x_{t-2}, \dots) = \phi_0 + \phi_1 x_{t-1}$$
$$\sigma_t^2 = \text{var}(x_t \mid x_{t-1}, x_{t-2}, \dots) = \text{var}(w_t) = \sigma_w^2 .$$

In many problems, however, the assumption of a constant conditional variance is violated. Models such as the *autoregressive conditionally heteroscedastic* or ARCH model, first introduced by Engle (1982), were developed to model changes in volatility. These models were later extended to generalized ARCH, or GARCH models by Bollerslev (1986).

In these problems, we are concerned with modeling the return or growth rate of a series. Recall if x_t is the value of an asset at time t, then the return or relative gain, r_t, of the asset at time t is

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}} \approx \nabla \log(x_t) . \tag{8.1}$$

Either value, $\nabla \log(x_t)$ or $(x_t - x_{t-1})/x_{t-1}$, will be called the *return* and will be denoted by r_t.[1]

Typically, for financial series, the return r_t, has a constant conditional mean (typically $\mu_t = 0$ for assets), but does not have a constant conditional variance, and highly volatile periods tend to be clustered together. In addition, the autocorrelation

[1] Although it is a misnomer, $\nabla \log x_t$ is often called the *log-return*; but the returns are NOT being logged.

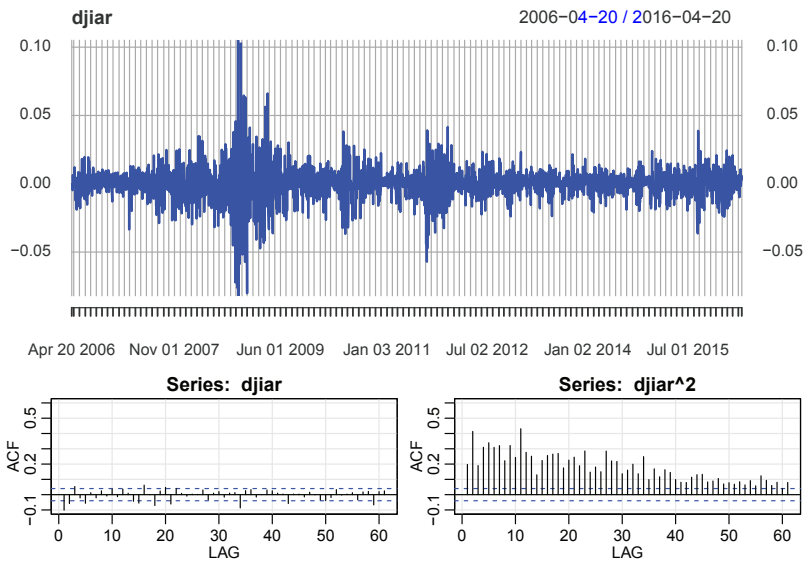


Figure 8.1 *DJIA daily closing returns and the sample ACF of the returns and of the squared returns.*

structure of r_t is that of white noise, while the returns are dependent. This can often be seen by looking at the sample ACF of the squared-returns (or some power transformation of the returns). For example, Figure 8.1 shows the daily returns of the Dow Jones Industrial Average (DJIA) that we saw in Chapter 1. In this case, as is typical, the return r_t is fairly constant (with $\mu_t = 0$) and nearly white noise, but there are short-term bursts of high volatility and the squared returns are autocorrelated.

The simplest ARCH model, the ARCH(1), models the returns as

$$r_t = \sigma_t \epsilon_t \tag{8.2}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2, \tag{8.3}$$

where ϵ_t is standard Gaussian white noise, $\epsilon_t \sim \text{iid N}(0,1)$. The normal assumption may be relaxed; we will discuss this later. As with ARMA models, we must impose some constraints on the model parameters to obtain desirable properties. An obvious constraint is that $\alpha_0, \alpha_1 \geq 0$ because σ_t^2 is a variance.

It is possible to write the ARCH(1) model as a non-Gaussian AR(1) model in the square of the returns r_t^2. First, rewrite (8.2)–(8.3) as

$$\begin{aligned} r_t^2 &= \sigma_t^2 \epsilon_t^2 \\ \alpha_0 + \alpha_1 r_{t-1}^2 &= \sigma_t^2, \end{aligned}$$

by squaring (8.2). Now subtract the two equations to obtain

$$r_t^2 - (\alpha_0 + \alpha_1 r_{t-1}^2) = \sigma_t^2 \epsilon_t^2 - \sigma_t^2 \,,$$

and rearrange it as

$$r_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + v_t, \tag{8.4}$$

where $v_t = \sigma_t^2(\epsilon_t^2 - 1)$. Because ϵ_t^2 is the square of a $\mathrm{N}(0,1)$ random variable, $\epsilon_t^2 - 1$ is a shifted (to have mean-zero), χ_1^2 random variable. In this case, v_t is non-normal white noise (see Section D.3 for details).

Thus, if $0 \le \alpha_1 < 1$, r_t^2 is a non-normal AR(1). This means that the ACF of the squared process is

$$\rho_{r^2}(h) = a_1^h \quad \text{for } h \ge 0\,.$$

In addition, it is shown in Section D.3 that, unconditionally, r_t is white noise with mean 0 and variance

$$\mathrm{var}(r_t) = \frac{\alpha_0}{1-\alpha_1}\,,$$

but conditionally,

$$r_t \mid r_{t-1} \sim \mathrm{N}(0, \alpha_0 + \alpha_1 r_{t-1}^2). \tag{8.5}$$

Hence, the model characterizes what we see in Figure 8.1:

- The returns are white noise.
- The conditional variance of a return depends on the previous return.
- The squared returns are autocorrelated.

Estimation of the parameters α_0 and α_1 of the ARCH(1) model is typically accomplished by conditional MLE based on the normal density specified in (8.5). This leads to weighted conditional least squares, which finds the values of α_0 and α_1 that minimize

$$S(\alpha_0, \alpha_1) = \tfrac{1}{2}\sum_{t=2}^{n} \ln(\alpha_0 + \alpha_1 r_{t-1}^2) + \tfrac{1}{2}\sum_{t=2}^{n}\left(\frac{r_t^2}{\alpha_0 + \alpha_1 r_{t-1}^2}\right), \tag{8.6}$$

using numerical methods, as described in Section 4.3.

The ARCH(1) model can be extended to the general ARCH(p) model in an obvious way. That is, (8.2), $r_t = \sigma_t\epsilon_t$, is retained, but (8.3) is extended to

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \cdots + \alpha_p r_{t-p}^2. \tag{8.7}$$

Estimation for ARCH(p) also follows in an obvious way from the discussion of estimation for ARCH(1) models.

It is also possible to combine a regression or an ARMA model for the conditional mean, say

$$r_t = \mu_t + \sigma_t\epsilon_t\,, \tag{8.8}$$

where, for example, a simple AR-ARCH model would have

$$\mu_t = \phi_0 + \phi_1 r_{t-1}.$$

Of course the model can be generalized to have various types of behavior for μ_t.

To fit ARMA-ARCH models, simply follow these two steps:

1. First, look at the P/ACF of the *returns*, r_t, and identify an ARMA structure, if any. There is typically either no autocorrelation or very small autocorrelation and often a low order AR or MA will suffice if needed. Estimate μ_t in order to center the returns if necessary.
2. Look at the P/ACF of the *centered squared returns*, $(r_t - \hat{\mu}_t)^2$, and decide on an ARCH model. If the P/ACF indicate an AR structure (i.e., ACF tails off, PACF cuts off), then fit an ARCH. If the P/ACF indicate an ARMA structure (i.e., both tail off), use the approach discussed after the next example.

Example 8.1. Analysis of U.S. GNP

In Example 5.6, we fit an AR(1) model to the U.S. GNP series and we concluded that the residuals appeared to behave like a white noise process. Hence, we would propose that $\mu_t = \phi_0 + \phi_1 r_{t-1}$ where r_t is the quarterly growth rate in U.S. GNP.

It has been suggested that the GNP series has ARCH errors, and in this example, we will investigate this claim. If the GNP noise term is ARCH, the squares of the residuals from the fit should behave like a non-Gaussian AR(1) process, as pointed out in (8.4). Figure 8.2 shows the ACF and PACF of the squared residuals and it appears that there may be some dependence, albeit small, left in the residuals. The figure was generated in R as follows.

```
res = resid( sarima(diff(log(gnp)), 1,0,0, details=FALSE)$fit )
acf2(res^2, 20)
```

We used the R package `fGarch` to fit an AR(1)-ARCH(1) model to the U.S. GNP returns with the following results. A partial output is shown; we note that `garch(1,0)` specifies an ARCH(1) in the code below (details later).

```
library(fGarch)
gnpr = diff(log(gnp))
summary( garchFit(~arma(1,0) + garch(1,0), data = gnpr) )
          Estimate   Std.Error  t.value  Pr(>|t|) <- 2-sided !!!
   mu        0.005       0.001    5.867     0.000
   ar1       0.367       0.075    4.878     0.000
   omega     0.000       0.000    8.135     0.000 <- these parameters
   alpha1    0.194       0.096    2.035     0.042 <- can't be negative

   Standardised Residuals Tests:  Statistic p-Value
    Jarque-Bera Test   R    Chi^2     9.118   0.010
    Shapiro-Wilk Test  R    W         0.984   0.014
    Ljung-Box Test     R    Q(20)    23.414   0.269
    Ljung-Box Test     R^2  Q(20)    37.743   0.010
```

Note that the given p-values are two-sided, so they should be halved when considering the ARCH parameters. In this example, we obtain $\hat{\phi}_0 = .005$ (called `mu` in the output) and $\hat{\phi}_1 = .367$ (called `ar1`) for the AR(1) parameter estimates; in Example 5.6 the values were .005 and .347, respectively. The ARCH(1) parameter

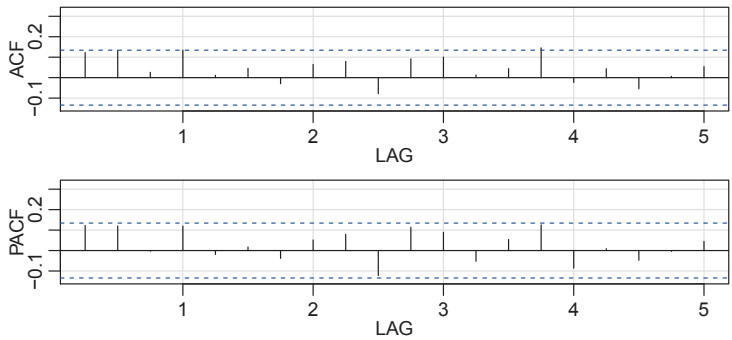


Figure 8.2 *ACF and PACF of the squares of the residuals from the AR(1) fit on U.S. GNP.*

estimates are $\hat{\alpha}_0 = 0$ (called `omega`) for the constant and $\hat{\alpha}_1 = .194$, which is significant with a p-value of about .02. There are a number of tests that are performed on the residuals [R] or the squared residuals [R^2]. For example, the Jarque–Bera statistic tests the residuals of the fit for normality based on the observed skewness and kurtosis, and it appears that the residuals have some non-normal skewness and kurtosis. The Shapiro–Wilk statistic tests the residuals of the fit for normality based on the empirical order statistics. The other tests, primarily based on the Q-statistic, are used on the residuals and their squares. ♢

The analysis of Example 8.1 had a few problems. First, it appears that the residuals are not normal (which was the assumption for the ϵ_t, and there may be some autocorrelation left in the squared residuals; see Problem 8.2). To address this kind of problem, the ARCH model was extended to generalized ARCH or GARCH. For example, a GARCH$(1,1)$ model retains (8.8), $r_t = \mu_t + \sigma_t\epsilon_t$, but extends (8.3) as follows:

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1\sigma_{t-1}^2. \tag{8.9}$$

Under the condition that $\alpha_1 + \beta_1 < 1$, using similar manipulations as in (8.4), the GARCH$(1,1)$ model, (8.2) and (8.9), admits a non-Gaussian ARMA$(1,1)$ model for the squared process

$$r_t^2 = \alpha_0 + (\alpha_1 + \beta_1)r_{t-1}^2 + v_t - \beta_1 v_{t-1}, \tag{8.10}$$

where we have set $\mu_t = 0$ for ease, and where v_t is as defined in (8.4). Representation (8.10) follows by writing (8.2) as

$$\begin{aligned} r_t^2 - \sigma_t^2 &= \sigma_t^2(\epsilon_t^2 - 1) \\ \beta_1(r_{t-1}^2 - \sigma_{t-1}^2) &= \beta_1\sigma_{t-1}^2(\epsilon_{t-1}^2 - 1), \end{aligned}$$

subtracting the second equation from the first, and using the fact that, from (8.9), $\sigma_t^2 - \beta_1\sigma_{t-1}^2 = \alpha_0 + \alpha_1 r_{t-1}^2$, on the left-hand side of the result. The GARCH(p,q)

model retains (8.8) and extends (8.9) to

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j r_{t-j}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2. \tag{8.11}$$

Estimation of the model parameters is similar to the estimation of ARCH parameters. We explore these concepts in the following example.

Example 8.2. GARCH Analysis of the DJIA Returns

As previously mentioned, the daily returns of the DJIA shown in Figure 8.1 exhibit classic GARCH features. In addition, there is some low level autocorrelation in the series itself, and to include this behavior, we used the R `fGarch` package to fit an AR(1)-GARCH(1,1) model to the series using t-errors (rather than normal):

```
library(xts)
djiar = diff(log(djia$Close))[-1]
acf2(djiar)     #  exhibits some autocorrelation - see Figure 8.1
u = resid( sarima(djiar, 1,0,0, details=FALSE)$fit )
acf2(u^2)       #  oozes autocorrelation - see Figure 8.1
library(fGarch)
summary(djia.g <- garchFit(~arma(1,0)+garch(1,1), data=djiar,
           cond.dist="std"))
            Estimate   Std.Error  t.value    Pr(>|t|)
   mu      8.585e-04   1.470e-04    5.842   5.16e-09
   ar1    -5.531e-02   2.023e-02   -2.735   0.006239
   omega   1.610e-06   4.459e-07    3.611   0.000305
   alpha1  1.244e-01   1.660e-02    7.497   6.55e-14
   beta1   8.700e-01   1.526e-02   57.022   < 2e-16
   shape   5.979e+00   7.917e-01    7.552   4.31e-14
   ---
   Standardised Residuals Tests:
                                 Statistic    p-Value
    Ljung-Box Test     R   Q(10)  16.81507  0.0785575
    Ljung-Box Test     R^2 Q(10)  15.39137  0.1184312
plot(djia.g, which=3) # similar to Figure 8.3
```

The `shape` parameter is the degrees of freedom for the t error distribution, which is estimated to be about 6. Also notice that $\hat{\alpha}_1 + \hat{\beta}_1$ is close to 1; this is often the case. To explore the GARCH predictions of volatility, we calculated and plotted part of the data surrounding the financial crises of 2008 along with the one-step-ahead predictions of the corresponding volatility, σ_t^2 as a solid line in Figure 8.3. ♢

Another model that we mention briefly is the *asymmetric power ARCH* model. The model retains (8.2), $r_t = \sigma_t \epsilon_t$, but the conditional variance is modeled as

$$\sigma_t^\delta = \alpha_0 + \sum_{j=1}^{p} \alpha_j (|r_{t-j}| - \gamma_j r_{t-j})^\delta + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^\delta . \tag{8.12}$$

Note that the model is GARCH when $\delta = 2$ and $\gamma_j = 0$, for $j \in \{1, \ldots, p\}$.

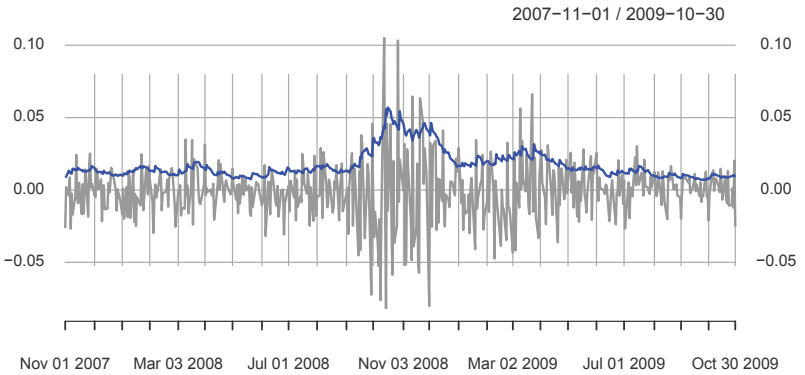


Figure 8.3 *GARCH one-step-ahead predictions of the DJIA volatility, σ_t, superimposed on part of the data including the financial crisis of 2008.*

The parameters γ_j ($|\gamma_j| \leq 1$) are the *leverage* parameters, which are a measure of asymmetry, and $\delta > 0$ is the parameter for the power term. A positive [negative] value of γ_j's means that past negative [positive] shocks have a deeper impact on current conditional volatility than past positive [negative] shocks. This model couples the flexibility of a varying exponent with the asymmetry coefficient to take the *leverage effect* into account. Further, to guarantee that $\sigma_t > 0$, we assume that $\alpha_0 > 0$, $\alpha_j \geq 0$ with at least one $\alpha_j > 0$, and $\beta_j \geq 0$.

We continue the analysis of the DJIA returns in the following example.

Example 8.3. APARCH Analysis of the DJIA Returns

The R package `fGarch` was used to fit an AR-APARCH model to the DJIA returns discussed in Example 8.2. As in the previous example, we include an AR(1) in the model to account for the conditional mean. In this case, we may think of the model as $r_t = \mu_t + y_t$ where μ_t is an AR(1), and y_t is APARCH noise with conditional variance modeled as (8.12) with t-errors. A partial output of the analysis is given below. We do not include displays, but we show how to obtain them. The predicted volatility is, of course, different than the values shown in Figure 8.3, but appear similar when graphed.

```
lapply( c("xts", "fGarch"), library, char=TRUE) # load 2 packages
djiar = diff(log(djia$Close))[-1]
summary(djia.ap <- garchFit(~arma(1,0)+aparch(1,1), data=djiar,
          cond.dist="std"))
plot(djia.ap)   # to see all plot options (none shown)
           Estimate  Std. Error  t value   Pr(>|t|)
  mu      5.234e-04   1.525e-04    3.432  0.000598
  ar1    -4.818e-02   1.934e-02   -2.491  0.012727
  omega   1.798e-04   3.443e-05    5.222  1.77e-07
  alpha1  9.809e-02   1.030e-02    9.525   < 2e-16
  gamma1  1.000e+00   1.045e-02   95.731   < 2e-16
```

```
beta1   8.945e-01   1.049e-02   85.280   < 2e-16
delta   1.070e+00   1.350e-01    7.928  2.22e-15
shape   7.286e+00   1.123e+00    6.489  8.61e-11
---
Standardised Residuals Tests:
                               Statistic p-Value
 Ljung-Box Test     R    Q(10)  15.71403  0.108116
 Ljung-Box Test     R^2  Q(10)  16.87473  0.077182
```

◊

In most applications, the distribution of the noise, ϵ_t in (8.2), is rarely normal. The R package `fGarch` allows for various distributions to be fit to the data; see the help file for information. Some drawbacks of GARCH and related models are as follows. (i) The GARCH model assumes positive and negative returns have the same effect because volatility depends on squared returns; the asymmetric models help alleviate this problem. (ii) These models are often restrictive because of the tight constraints on the model parameters. (iii) The likelihood is flat unless n is very large. (iv) The models tend to overpredict volatility because they respond slowly to large isolated returns.

Various extensions to the original model have been proposed to overcome some of the shortcomings we have just mentioned. For example, we have already discussed the fact that `fGarch` allows for asymmetric return dynamics. In the case of persistence in volatility, the integrated GARCH (IGARCH) model may be used. Recall (8.10) where we showed the GARCH$(1,1)$ model can be written as

$$r_t^2 = \alpha_0 + (\alpha_1 + \beta_1) r_{t-1}^2 + v_t - \beta_1 v_{t-1}$$

and r_t^2 is stationary if $\alpha_1 + \beta_1 < 1$. The IGARCH model sets $\alpha_1 + \beta_1 = 1$, in which case the IGARCH$(1,1)$ model is

$$r_t = \sigma_t \epsilon_t \quad \text{and} \quad \sigma_t^2 = \alpha_0 + (1-\beta_1) r_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

There are many different extensions to the basic ARCH model that were developed to handle the various situations noticed in practice. Interested readers might find the general discussions in Bollerslev et al. (1994) and Shephard (1996) worthwhile reading. Two excellent texts on financial time series analysis are Chan (2002) and Tsay (2005).

8.2 Unit Root Testing

The use of the first difference $\nabla x_t = (1-B)x_t$ can sometimes be too severe a modification in the sense that an integrated model might represent an overdifferencing of the original process. For example, in Example 5.8 we fit an ARIMA(1,1,1) model to the logged varve series. The idea of differencing the series was first made in Example 4.27 because the series appeared to take long 100+ year walks in positive and negative directions.

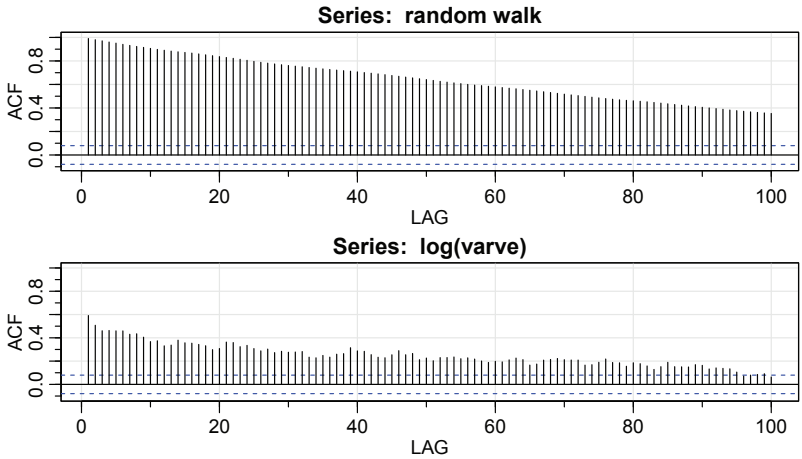


Figure 8.4 *Sample ACFs a random walk and of the log transformed varve series.*

Figure 8.4 compares the sample ACF of a generated random walk with that of the logged varve series. Although in both cases the sample correlations decrease linearly and remain significant for many lags, the sample ACF of the random walk has much larger values. (Recall that there is no ACF in terms of lag only for a random walk. But that doesn't stop us from computing one.)

```
layout(1:2)
acf1(cumsum(rnorm(634)), 100, main="Series:  random walk")
acf1(log(varve), 100, ylim=c(-.1,1))
```

Consider the normal AR(1) process,

$$x_t = \phi x_{t-1} + w_t. \tag{8.13}$$

A unit root test provides a way to test whether (8.13) is a random walk (the null case) as opposed to a causal process (the alternative). That is, it provides a procedure for testing

$$H_0 : \phi = 1 \quad \text{versus} \quad H_1 : |\phi| < 1.$$

To see if the null hypothesis is reasonable, an obvious test statistic would be to consider $(\widehat{\phi} - 1)$, appropriately normalized, in the hope to develop a t-test, where $\widehat{\phi}$ is one of the optimal estimators discussed in Section 4.3. Note that the distribution in Property 4.29 does not work in this case; if it did, under the null hypothesis, $\widehat{\phi} \sim N(1, 0)$, which is nonsense. The theory of Section 4.3 does not work in the null case because the process is not stationary under the null hypothesis.

However, the test statistic

$$T = n(\widehat{\phi} - 1)$$

can be used, and it is known as the unit root or Dickey–Fuller (DF) statistic, although the actual DF test statistic is normalized a little differently. In this case, the distribution

of the test statistic does not have a closed form and quantiles of the distribution must be computed by numerical approximation or by simulation. The R package `tseries` provides this test along with more general tests that we mention briefly.

Toward a more general model, we note that the DF test was established by noting that if $x_t = \phi x_{t-1} + w_t$, then

$$\nabla x_t = (\phi - 1)x_{t-1} + w_t = \gamma x_{t-1} + w_t\,,$$

and one could test $H_0\colon \gamma = 0$ by regressing ∇x_t on x_{t-1} and obtaining the regression coefficient estimate $\widehat{\gamma}$. Then, the statistic $n\widehat{\gamma}$ was formed and its large sample distribution derived.

The test was extended to accommodate AR(p) models, $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t$, in a similar way. For example, write an AR(2) model

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t\,,$$

as

$$x_t = (\phi_1 + \phi_2)x_{t-1} - \phi_2(x_{t-1} - x_{t-2}) + w_t\,,$$

and subtract x_{t-1} from both sides. This yields

$$\nabla x_t = \gamma x_{t-1} + \phi_2 \nabla x_{t-1} + w_t, \tag{8.14}$$

where $\gamma = \phi_1 + \phi_2 - 1$. To test the hypothesis that the process has a unit root at 1 (i.e., the AR polynoimial $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 = 0$ when $z = 1$), we can test $H_0\colon \gamma = 0$ by estimating γ in the regression of ∇x_t on x_{t-1} and ∇x_{t-1} and forming a test statistic. For AR(p) model, one regresses ∇x_t on x_{t-1} and $\nabla x_{t-1} \ldots, \nabla x_{t-p+1}$, in a similar fashion to the AR(2) case.

This test leads to the so-called augmented Dickey–Fuller test (ADF). While the calculations for obtaining the large sample null distribution change, the basic ideas and machinery remain the same as in the simple case. The choice of p is crucial, and we will discuss some suggestions in the example. For ARMA(p, q) models, the ADF test can be used by assuming p is large enough to capture the essential correlation structure; recall ARMA(p, q) models are AR(∞) models. An alternative is the Phillips–Perron (PP) test, which differs from the ADF tests mainly in how it deals with serial correlation and heteroscedasticity in the errors.

Example 8.4. Testing Unit Roots in the Glacial Varve Series

In this example we use the R package `tseries` to test the null hypothesis that the log of the glacial varve series has a unit root, versus the alternate hypothesis that the process is stationary. We test the null hypothesis using the available DF, ADF, and PP tests; note that in each case, the general regression equation incorporates a constant and a linear trend. In the ADF test, the default number of AR components included in the model is $k \approx (n-1)^{\frac{1}{3}}$, which has theoretical justification on how k should grow compared to the sample size n. For the PP test, the default value is $k \approx .04n^{\frac{1}{4}}$.

```
library(tseries)
adf.test(log(varve), k=0)                # DF test
  Dickey-Fuller = -12.8572, Lag order = 0, p-value < 0.01
   alternative hypothesis: stationary
adf.test(log(varve))                     # ADF test
  Dickey-Fuller = -3.5166, Lag order = 8, p-value = 0.04071
   alternative hypothesis: stationary
pp.test(log(varve))                      # PP test
   Dickey-Fuller Z(alpha) = -304.5376,
    Truncation lag parameter = 6, p-value < 0.01
    alternative hypothesis: stationary
```

In each test, we reject the null hypothesis that the logged varve series has a unit root. The conclusion of these tests supports the conclusion of Example 8.5 in Section 8.3, where it is postulated that the logged varve series is long memory. Fitting a long memory model to these data would be the natural progression of model fitting once the unit root test hypothesis is rejected. ◊

8.3 Long Memory and Fractional Differencing

The conventional ARMA(p,q) process is often referred to as a short-memory process because the coefficients in the representation

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

are dominated by exponential decay where $\sum_{j=0}^{\infty} |\psi_j| < \infty$ (e.g., recall Example 4.3). This result implies the ACF of the short memory process $\rho(h) \to 0$ exponentially fast as $h \to \infty$. When the sample ACF of a time series decays slowly, the advice given in Chapter 6 has been to difference the series until it seems stationary. Following this advice with the glacial varve series first presented in Example 4.27 leads to the first difference of the logarithms of the data, say $x_t =$`log(varve)`, being represented as a first-order moving average. In Example 5.8, further analysis of the residuals leads to fitting an ARIMA$(1,1,1)$ model, where the estimates of the parameters (and the standard errors) were $\widehat{\phi} = .23_{(.05)}$, $\widehat{\theta} = -.89_{(.03)}$, and $\widehat{\sigma}_w^2 = .23$:

$$\nabla \hat{x}_t = .23 \nabla \hat{x}_{t-1} + \hat{w}_t - .89 \hat{w}_{t-1}.$$

What the fitted model is saying is that the series itself, x_t, is not stationary and has random walk behavior, and the only way to make it stationary is to difference it. In terms of the actual logged varve series, the fitted model is

$$\hat{x}_t = (1 + .23)\hat{x}_{t-1} - .23\hat{x}_{t-2} + \hat{w}_t - .89\hat{w}_{t-1}$$

and there is no causal representation for the data because the ψ-weights are not square summable (in fact, they do not even go to zero):

```
round(ARMAtoMA(ar=c(1.23,-.23), ma=c(1,-.89), 20), 3)
 [1] 2.230 1.623 1.483 1.451 1.444 1.442 1.442 1.442 1.442 1.442
[11] 1.442 1.442 1.442 1.442 1.442 1.442 1.442 1.442 1.442 1.442
```

But the use of the first difference $\nabla x_t = (1 - B)x_t$ can be too severe of a transformation. For example, if x_t is a causal AR(1), say

$$x_t = .9x_{t-1} + w_t\,,$$

then shifting back one unit of time,

$$x_{t-1} = .9x_{t-2} + w_{t-1}\,.$$

Now subtract the two to get,

$$x_t - x_{t-1} = .9(x_{t-1} - x_{t-2}) + w_t - w_{t-1}\,,$$

or

$$\nabla x_t = .9\nabla x_{t-1} + w_t - w_{t-1}\,.$$

This means that ∇x_t is a problematic ARMA$(1,1)$ because the moving average part is non-invertible. Thus, by overdifferencing in this example, we have gone from x_t being a simple causal AR(1) to x_t being a non-invertible ARIMA$(1,1,1)$. This is precisely why we gave several warnings about the overuse of differencing in Chapter 4 and Chapter 5.

Long memory time series were considered in Hosking (1981) and Granger and Joyeux (1980) as intermediate compromises between the short memory ARMA type models and the fully integrated nonstationary processes in the Box–Jenkins class. The easiest way to generate a long memory series is to think of using the difference operator $(1 - B)^d$ for fractional values of d, say, $0 < d < .5$, so a basic long memory series gets generated as

$$(1 - B)^d x_t = w_t, \tag{8.15}$$

where w_t still denotes white noise with variance σ_w^2. The fractionally differenced series (8.15), for $|d| < .5$, is often called *fractional noise* (except when d is zero). Now, d becomes a parameter to be estimated along with σ_w^2. Differencing the original process, as in the Box–Jenkins approach, may be thought of as simply assigning a value of $d = 1$. This idea has been extended to the class of fractionally integrated ARMA, or ARFIMA models, where $-.5 < d < .5$; when d is negative, the term antipersistent is used. Long memory processes occur in hydrology (see Hurst, 1951, McLeod and Hipel, 1978) and in environmental series, such as the varve data we have previously analyzed, to mention a few examples. Long memory time series data tend to exhibit sample autocorrelations that are not necessarily large (as in the case of $d = 1$), but persist for a long time. Figure 8.4 shows the sample ACF, to lag 100, of the log-transformed varve series, which exhibits classic long memory behavior.

To investigate its properties, we can use the binomial expansion[2] ($d > -1$) to write

$$w_t = (1-B)^d x_t = \sum_{j=0}^{\infty} \pi_j B^j x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} \tag{8.16}$$

where

$$\pi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \tag{8.17}$$

with $\Gamma(x+1) = x\Gamma(x)$ being the gamma function. Similarly $(d < 1)$, we can write

$$x_t = (1-B)^{-d} w_t = \sum_{j=0}^{\infty} \psi_j B^j w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \tag{8.18}$$

where

$$\psi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}. \tag{8.19}$$

When $|d| < .5$, the processes (8.16) and (8.18) are well-defined stationary processes (see Brockwell and Davis, 2013, for details). In the case of fractional differencing, however, the coefficients satisfy $\sum \pi_j^2 < \infty$ and $\sum \psi_j^2 < \infty$ as opposed to the absolute summability of the coefficients in ARMA processes.

Using the representation (8.18)–(8.19), and after some nontrivial manipulations, it can be shown that the ACF of x_t is

$$\rho(h) = \frac{\Gamma(h+d)\Gamma(1-d)}{\Gamma(h-d+1)\Gamma(d)} \sim h^{2d-1} \tag{8.20}$$

for large h. From this we see that for $0 < d < .5$

$$\sum_{h=-\infty}^{\infty} |\rho(h)| = \infty$$

and hence the term *long memory*.

In order to examine a series such as the varve series for a possible long memory pattern, it is convenient to look at ways of estimating d. Using (8.17) it is easy to derive the recursions

$$\pi_{j+1}(d) = \frac{(j-d)\pi_j(d)}{(j+1)}, \tag{8.21}$$

for $j = 0, 1, \ldots$, with $\pi_0(d) = 1$. In the normal case, we may estimate d by minimizing the sum of squared errors

$$Q(d) = \sum w_t^2(d)\,.$$

The usual Gauss–Newton method, described in Section 4.3, leads to the expansion

$$w_t(d) \approx w_t(d_0) + w_t'(d_0)(d - d_0),$$

[2]The binomial expansion in this case is the Taylor series about $z = 0$ for functions of the form $(1-z)^d$

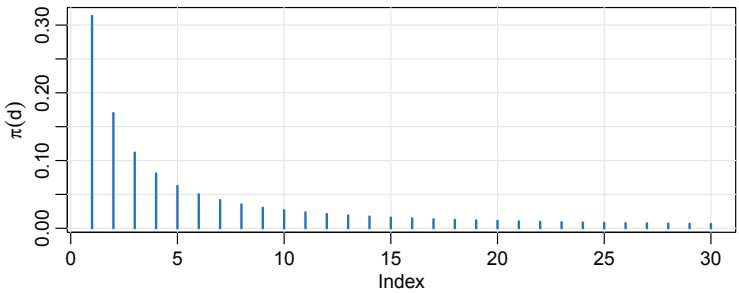


Figure 8.5 *Coefficients* $\pi_j(.373)$, $j = 1, 2, \ldots, 30$ *in the representation* (8.21).

where

$$w_t'(d_0) = \left.\frac{\partial w_t}{\partial d}\right|_{d=d_0}$$

and d_0 is an initial estimate (guess) at to the value of d. Setting up the usual regression leads to

$$d = d_0 - \frac{\sum_t w_t'(d_0) w_t(d_0)}{\sum_t w_t'(d_0)^2}. \tag{8.22}$$

The derivatives are computed recursively by differentiating (8.21) successively with respect to d: $\pi_{j+1}'(d) = [(j-d)\pi_j'(d) - \pi_j(d)]/(j+1)$, where $\pi_0'(d) = 0$. The errors are computed from an approximation to (8.16), namely,

$$w_t(d) = \sum_{j=0}^{t} \pi_j(d) x_{t-j}. \tag{8.23}$$

It is advisable to omit a number of initial terms from the computation and start the sum, (8.22), at some fairly large value of t to have a reasonable approximation.

Example 8.5. Long Memory Fitting of the Glacial Varve Series

We consider analyzing the glacial varve series discussed in Example 3.12 and Example 4.27. Figure 3.9 shows the original and log-transformed series (which we denote by x_t). In Example 5.8, we noted that x_t could be modeled as an ARIMA$(1,1,1)$ process. We fit the fractionally differenced model, (8.15), to the mean-adjusted series, $x_t - \bar{x}$. Applying the Gauss–Newton iterative procedure previously described leads to a final value of $d = .373$, which implies the set of coefficients $\pi_j(.373)$, as given in Figure 8.5 with $\pi_0(.373) = 1$.

```
d = 0.3727893
p = c(1)
for (k in 1:30){
  p[k+1] = (k-d)*p[k]/(k+1)
}
tsplot(1:30, p[-1], ylab=expression(pi(d)), lwd=2, xlab="Index",
           type="h", col="dodgerblue3")
```

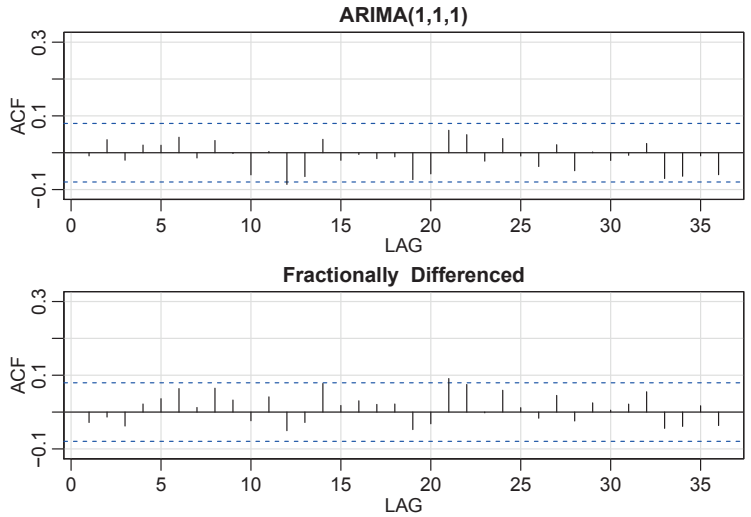


Figure 8.6 *ACF of residuals from the ARIMA*$(1,1,1)$ *fit to* x_t*, the logged varve series (top) and of the residuals from the long memory model fit,* $(1-B)^d x_t = w_t$*, with* $d = .373$ *(bottom).*

We can compare roughly the performance of the fractional difference operator with the ARIMA model by examining the autocorrelation functions of the two residual series as shown in Figure 8.6. The ACFs of the two residual series are roughly comparable with the white noise model.

To perform this analysis in R, use the `arfima` package. Note that after the analysis, when the innovations (residuals) are pulled out of the results, they are in the form of a list and thus the need for double brackets (`[[ ]]`) below:

```
library(arfima)
summary(varve.fd <- arfima(log(varve), order = c(0,0,0)))
  Mode 1 Coefficients:
              Estimate Std. Error Th. Std. Err. z-value   Pr(>|z|)
 d.f         0.3727893  0.0273459     0.0309661 13.6324 < 2.22e-16
 Fitted mean 3.0814142  0.2646507            NA 11.6433 < 2.22e-16
 ---
 sigma^2 estimated as 0.229718;
 Log-likelihood = 466.028; AIC = -926.056; BIC = 969.944
# innovations (aka residuals)
innov = resid(varve.fd)[[1]]  # resid() produces a `list`
tsplot(innov)       # not shown
par(mfrow=2:1, cex.main=1)
acf1(resid(sarima(log(varve),1,1,1, details=FALSE)$fit),
           main="ARIMA(1,1,1)")
acf1(innov, main="Fractionally Differenced")
```

◊

Forecasting long memory processes is similar to forecasting ARIMA models.

That is, (8.16) and (8.21) can be used to obtain the truncated forecasts

$$x^n_{n+m} = -\sum_{j=1}^{n+m-1} \pi_j(\widehat{d})\, x^n_{n+m-j}, \tag{8.24}$$

for $m = 1, 2, \ldots$. Error bounds can be approximated by using

$$P^n_{n+m} = \widehat{\sigma}^2_w \sum_{j=0}^{m-1} \psi^2_j(\widehat{d}) \tag{8.25}$$

where, as in (8.21),

$$\psi_j(\widehat{d}) = \frac{(j+\widehat{d})\psi_j(\widehat{d})}{(j+1)}, \tag{8.26}$$

with $\psi_0(\widehat{d}) = 1$.

No obvious short memory ARMA-type component can be seen in the ACF of the residuals from the fractionally differenced varve series shown in Figure 8.6. It is natural, however, that cases will exist in which substantial short memory-type components will also be present in data that exhibits long memory. Hence, it is natural to define the general ARFIMA(p, d, q), $-.5 < d < .5$ process as

$$\phi(B)\nabla^d(x_t - \mu) = \theta(B)w_t, \tag{8.27}$$

where $\phi(B)$ and $\theta(B)$ are as given in Chapter 4. Writing the model in the form

$$\phi(B)\pi_d(B)(x_t - \mu) = \theta(B)w_t \tag{8.28}$$

makes it clear how we go about estimating the parameters for the more general model. Forecasting for the ARFIMA(p, d, q) series can be easily done, noting that we may equate coefficients in

$$\phi(z)\psi(z) = (1-z)^{-d}\theta(z) \tag{8.29}$$

and

$$\theta(z)\pi(z) = (1-z)^{d}\phi(z) \tag{8.30}$$

to obtain the representations

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j w_{t-j}$$

and

$$w_t = \sum_{j=0}^{\infty} \pi_j(x_{t-j} - \mu).$$

We then can proceed as discussed in (8.24) and (8.25).

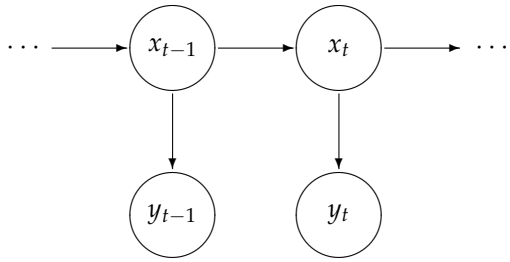


Figure 8.7 *Diagram of a state space model.*

8.4 State Space Models

A very general model that subsumes a whole class of special cases of interest in much the same way that linear regression does is the state space model that was introduced in Kalman (1960) and Kalman and Bucy (1961). The model arose in the space tracking setting, where the state equation defines the motion equations for the position or state of a spacecraft with location x_t and the data y_t reflect information that can be observed from a tracking device. Although it is typically applied to multivariate time series, we focus on the univariate case here.

In general, the state space model is characterized by two principles. First, there is a hidden or latent process x_t called the state process. The unobserved state process is assumed to be an AR(1),

$$x_t = \alpha + \phi x_{t-1} + w_t\,, \tag{8.31}$$

where $w_t \sim \text{iid } \mathrm{N}(0, \sigma_w^2)$. In addition, we assume the initial state is $x_0 \sim \mathrm{N}(\mu_0, \sigma_0^2)$. The second condition is that the observations, y_t, are given by

$$y_t = Ax_t + v_t\,, \tag{8.32}$$

where A is a constant and the observation noise is $v_t \sim \text{iid } \mathrm{N}(0, \sigma_v^2)$. In addition, x_0, $\{w_t\}$ and $\{v_t\}$ are uncorrelated. This means that the dependence among the observations is generated by states. The principles are displayed in Figure 8.7.

A primary aim of any analysis involving the state space model, (8.31)–(8.32), is to produce estimators for the underlying unobserved signal x_t, given the data $y_{1:s} = \{y_1, \ldots, y_s\}$, to time s. When $s < t$, the problem is called *forecasting* or *prediction*. When $s = t$, the problem is called *filtering*, and when $s > t$, the problem is called *smoothing*. In addition to these estimates, we would also want to measure their precision. The solution to these problems is accomplished via the *Kalman filter* and *smoother*.

First, we present the Kalman filter, which gives the prediction and filtering equations. We use the following notation,

$$x_t^s = E(x_t \mid y_{1:s}) \quad \text{and} \quad P_t^s = E(x_t - x_t^s)^2\,.$$