1

Time Series Data

1.1 Purpose

Time series are analysed to understand the past and to predict the future, enabling managers or policy makers to make properly informed decisions. A time series analysis quantifies the main features in data and the random variation. These reasons, combined with improved computing power, have made time series methods widely applicable in government, industry, and commerce.

The Kyoto Protocol is an amendment to the United Nations Framework Convention on Climate Change. It opened for signature in December 1997 and came into force on February 16, 2005. The arguments for reducing greenhouse gas emissions rely on a combination of science, economics, and time series analysis. Decisions made in the next few years will affect the future of the planet.

During 2006, Singapore Airlines placed an initial order for twenty Boeing 787-9s and signed an order of intent to buy twenty-nine new Airbus planes, twenty A350s, and nine A380s (superjumbos). The airline's decision to expand its fleet relied on a combination of time series analysis of airline passenger trends and corporate plans for maintaining or increasing its market share.

Time series methods are used in everyday operational decisions. For example, gas suppliers in the United Kingdom have to place orders for gas from the offshore fields one day ahead of the supply. Variation about the average for the time of year depends on temperature and, to some extent, the wind speed. Time series analysis is used to forecast demand from the seasonal average with adjustments based on one-day-ahead weather forecasts.

Time series models often form the basis of computer simulations. Some examples are assessing different strategies for control of inventory using a simulated time series of demand; comparing designs of wave power devices using a simulated series of sea states; and simulating daily rainfall to investigate the long-term environmental effects of proposed water management policies.

P.S.P. Cowpertwait and A.V. Metcalfe, *Introductory Time Series with R*, Use R, DOI 10.1007/978-0-387-88698-5_1,

1.2 Time series

In most branches of science, engineering, and commerce, there are variables measured sequentially in time. Reserve banks record interest rates and exchange rates each day. The government statistics department will compute the country's gross domestic product on a yearly basis. Newspapers publish yesterday's noon temperatures for capital cities from around the world. Meteorological offices record rainfall at many different sites with differing resolutions. When a variable is measured sequentially in time over or at a fixed interval, known as the *sampling interval*, the resulting data form a *time series*.

Observations that have been collected over fixed sampling intervals form a *historical* time series. In this book, we take a *statistical* approach in which the historical series are treated as realisations of sequences of *random variables*. A sequence of random variables defined at fixed sampling intervals is sometimes referred to as a *discrete-time stochastic process*, though the shorter name *time series model* is often preferred. The theory of stochastic processes is vast and may be studied without necessarily fitting any models to data. However, our focus will be more applied and directed towards model fitting and data analysis, for which we will be using R.[1]

The main features of many time series are trends and seasonal variations that can be modelled deterministically with mathematical functions of time. But, another important feature of most time series is that observations close together in time tend to be correlated (*serially dependent*). Much of the methodology in a time series analysis is aimed at explaining this correlation and the main features in the data using appropriate statistical models and descriptive methods. Once a good model is found and fitted to data, the analyst can use the model to forecast future values, or generate simulations, to guide planning decisions. Fitted models are also used as a basis for statistical tests. For example, we can determine whether fluctuations in monthly sales figures provide evidence of some underlying change in sales that we must now allow for. Finally, a fitted statistical model provides a concise summary of the main characteristics of a time series, which can often be essential for decision makers such as managers or politicians.

Sampling intervals differ in their relation to the data. The data may have been aggregated (for example, the number of foreign tourists arriving per day) or sampled (as in a daily time series of close of business share prices). If data are sampled, the sampling interval must be short enough for the time series to provide a very close approximation to the original continuous signal when it is interpolated. In a volatile share market, close of business prices may not suffice for interactive trading but will usually be adequate to show a company's financial performance over several years. At a quite different timescale,

[1] R was initiated by Ihaka and Gentleman (1996) and is an open source implementation of S, a language for data analysis developed at Bell Laboratories (Becker et al. 1988).

time series analysis is the basis for signal processing in telecommunications, engineering, and science. Continuous electrical signals are sampled to provide time series using analog-to-digital (A/D) converters at rates that can be faster than millions of observations per second.

1.3 R language

It is assumed that you have R (version 2 or higher) installed on your computer, and it is suggested that you work through the examples, making sure your output agrees with ours.[2] If you do not have R, then it can be installed free of charge from the Internet site www.r-project.org. It is also recommended that you have some familiarity with the basics of R, which can be obtained by working through the first few chapters of an elementary textbook on R (e.g., Dalgaard 2002) or using the online "An Introduction to R", which is also available via the R help system – type `help.start()` at the command prompt to access this.

R has many features in common with both *functional* and *object oriented* programming languages. In particular, functions in R are treated as objects that can be manipulated or used recursively.[3] For example, the factorial function can be written recursively as

```
> Fact <- function(n) if (n == 1) 1 else n * Fact(n - 1)
> Fact(5)

[1] 120
```

In common with functional languages, assignments in R can be avoided, but they are useful for clarity and convenience and hence will be used in the examples that follow. In addition, R runs faster when 'loops' are avoided, which can often be achieved using matrix calculations instead. However, this can sometimes result in rather obscure-looking code. Thus, for the sake of transparency, loops will be used in many of our examples. Note that R is *case sensitive*, so that `X` and `x`, for example, correspond to different variables. In general, we shall use uppercase for the first letter when defining new variables, as this reduces the chance of overwriting inbuilt R functions, which are usually in lowercase.[4]

[2] Some of the output given in this book may differ slightly from yours. This is most likely due to editorial changes made for stylistic reasons. For conciseness, we also used `options(digits=3)` to set the number of digits to 4 in the computer output that appears in the book.

[3] Do not be concerned if you are unfamiliar with some of these computing terms, as they are not really essential in understanding the material in this book. The main reason for mentioning them now is to emphasise that R can almost certainly meet your *future* statistical and programming needs should you wish to take the study of time series further.

[4] For example, matrix transpose is `t()`, so `t` should not be used for *time*.

The best way to learn to do a time series analysis in R is through practice, so we now turn to some examples, which we invite you to work through.

1.4 Plots, trends, and seasonal variation

1.4.1 A flying start: Air passenger bookings

The number of international passenger bookings (in thousands) per month on an airline (Pan Am) in the United States were obtained from the Federal Aviation Administration for the period 1949–1960 (Brown, 1963). The company used the data to predict future demand before ordering new aircraft and training aircrew. The data are available as a time series in R and illustrate several important concepts that arise in an exploratory time series analysis.

Type the following commands in R, and check your results against the output shown here. To save on typing, the data are assigned to a variable called AP.

```
> data(AirPassengers)
> AP <- AirPassengers
> AP

     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1949 112 118 132 129 121 135 148 148 136 119 104 118
1950 115 126 141 135 125 149 170 170 158 133 114 140
1951 145 150 178 163 172 178 199 199 184 162 146 166
1952 171 180 193 181 183 218 230 242 209 191 172 194
1953 196 196 236 235 229 243 264 272 237 211 180 201
1954 204 188 235 227 234 264 302 293 259 229 203 229
1955 242 233 267 269 270 315 364 347 312 274 237 278
1956 284 277 317 313 318 374 413 405 355 306 271 306
1957 315 301 356 348 355 422 465 467 404 347 305 336
1958 340 318 362 348 363 435 491 505 404 359 310 337
1959 360 342 406 396 420 472 548 559 463 407 362 405
1960 417 391 419 461 472 535 622 606 508 461 390 432
```

All data in R are stored in *objects*, which have a range of *methods* available. The *class* of an object can be found using the `class` function:

```
> class(AP)

[1] "ts"

> start(AP); end(AP); frequency(AP)

[1] 1949    1
[1] 1960   12
[1] 12
```

In this case, the object is of class `ts`, which is an abbreviation for 'time series'. Time series objects have a number of *methods* available, which include the functions `start`, `end`, and `frequency` given above. These methods can be listed using the function `methods`, but the output from this function is not always helpful. The key thing to bear in mind is that *generic* functions in `R`, such as `plot` or `summary`, will attempt to give the most appropriate output to any given input object; try typing `summary(AP)` now to see what happens.

As the objective in this book is to analyse time series, it makes sense to put our data into objects of class `ts`. This can be achieved using a function also called `ts`, but this was not necessary for the airline data, which were already stored in this form. In the next example, we shall create a `ts` object from data read directly from the Internet.

One of the most important steps in a preliminary time series analysis is to plot the data; i.e., create a *time plot*. For a time series object, this is achieved with the generic plot function:

```
> plot(AP, ylab = "Passengers (1000's)")
```

You should obtain a plot similar to Figure 1.1 below. Parameters, such as `xlab` or `ylab`, can be used in `plot` to improve the default labels.

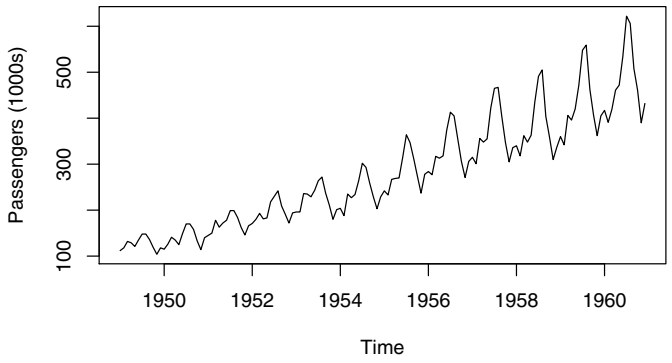


Fig. 1.1. International air passenger bookings in the United States for the period 1949–1960.

There are a number of features in the time plot of the air passenger data that are common to many time series (Fig. 1.1). For example, it is apparent that the number of passengers travelling on the airline is increasing with time. In general, a systematic change in a time series that does not appear to be periodic is known as a *trend*. The simplest model for a trend is a linear increase or decrease, and this is often an adequate approximation.

A repeating pattern within each year is known as *seasonal variation*, although the term is applied more generally to repeating patterns within any fixed period, such as restaurant bookings on different days of the week. There is clear seasonal variation in the air passenger time series. At the time, bookings were highest during the summer months of June, July, and August and lowest during the autumn month of November and winter month of February. Sometimes we may claim there are *cycles* in a time series that do not correspond to some fixed natural period; examples may include business cycles or climatic oscillations such as El Niño. None of these is apparent in the airline bookings time series.

An understanding of the likely causes of the features in the plot helps us formulate an appropriate time series model. In this case, possible causes of the increasing trend include rising prosperity in the aftermath of the Second World War, greater availability of aircraft, cheaper flights due to competition between airlines, and an increasing population. The seasonal variation coincides with vacation periods. In Chapter 5, time series regression models will be specified to allow for underlying causes like these. However, many time series exhibit trends, which might, for example, be part of a longer cycle or be random and subject to unpredictable change. Random, or *stochastic*, trends are common in economic and financial time series. A regression model would not be appropriate for a stochastic trend.

Forecasting relies on extrapolation, and forecasts are generally based on an assumption that present trends continue. We cannot check this assumption in any empirical way, but if we can identify likely causes for a trend, we can justify extrapolating it, for a few time steps at least. An additional argument is that, in the absence of some shock to the system, a trend is likely to change relatively slowly, and therefore linear extrapolation will provide a reasonable approximation for a few time steps ahead. Higher-order polynomials may give a good fit to the historic time series, but they should not be used for extrapolation. It is better to use linear extrapolation from the more recent values in the time series. Forecasts based on extrapolation beyond a year are perhaps better described as scenarios. Expecting trends to continue linearly for many years will often be unrealistic, and some more plausible trend curves are described in Chapters 3 and 5.

A time series plot not only emphasises patterns and features of the data but can also expose *outliers* and *erroneous* values. One cause of the latter is that missing data are sometimes coded using a negative value. Such values need to be handled differently in the analysis and must not be included as observations when fitting a model to data.[5] Outlying values that cannot be attributed to some coding should be checked carefully. If they are correct,

[5] Generally speaking, missing values are suitably handled by `R`, provided they are correctly coded as 'NA'. However, if your data do contain missing values, then it is always worth checking the 'help' on the `R` function that you are using, as an extra parameter or piece of coding may be required.

they are likely to be of particular interest and should not be excluded from the analysis. However, it may be appropriate to consider *robust methods* of fitting models, which reduce the influence of outliers.

To get a clearer view of the trend, the seasonal effect can be removed by aggregating the data to the annual level, which can be achieved in `R` using the `aggregate` function. A summary of the values for each season can be viewed using a boxplot, with the `cycle` function being used to extract the seasons for each item of data.

The plots can be put in a single graphics window using the `layout` function, which takes as input a vector (or matrix) for the location of each plot in the display window. The resulting boxplot and annual series are shown in Figure 1.2.

```
> layout(1:2)
> plot(aggregate(AP))
> boxplot(AP ~ cycle(AP))
```

You can see an increasing trend in the annual series (Fig. 1.2a) and the seasonal effects in the boxplot. More people travelled during the summer months of June to September (Fig. 1.2b).

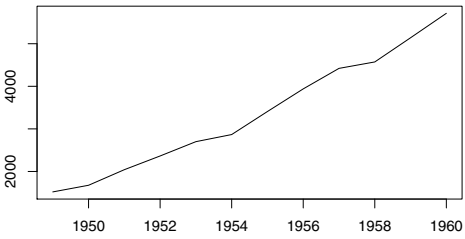
1.4.2 Unemployment: Maine

Unemployment rates are one of the main economic indicators used by politicians and other decision makers. For example, they influence policies for regional development and welfare provision. The monthly unemployment rate for the US state of Maine from January 1996 until August 2006 is plotted in the upper frame of Figure 1.3. In any time series analysis, it is essential to understand how the data have been collected and their unit of measurement. The US Department of Labor gives precise definitions of terms used to calculate the unemployment rate.

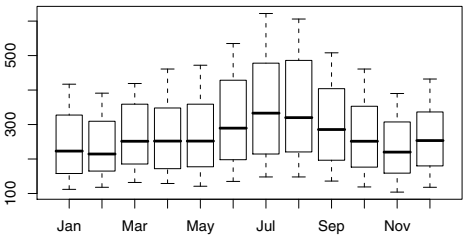
The monthly unemployment data are available in a file online that is read into `R` in the code below. Note that the first row in the file contains the name of the variable (`unemploy`), which can be accessed directly once the `attach` command is given. Also, the `header` parameter must be set to `TRUE` so that `R` treats the first row as the variable name rather than data.

```
> www <- "http://www.massey.ac.nz/~pscowper/ts/Maine.dat"
> Maine.month <- read.table(www, header = TRUE)

> attach(Maine.month)
> class(Maine.month)

[1] "data.frame"
```

When we read data in this way from an ASCII text file, the 'class' is not time series but `data.frame`. The `ts` function is used to convert the data to a time series object. The following command creates a time series object:



(a) Aggregated annual series



(b) Boxplot of seasonal values

Fig. 1.2. International air passenger bookings in the United States for the period 1949–1960. Units on the y-axis are 1000s of people. (a) Series aggregated to the annual level; (b) seasonal boxplots of the data.

```
> Maine.month.ts <- ts(unemploy, start = c(1996, 1), freq = 12)
```

This uses all the data. You can select a smaller number by specifying an earlier end date using the parameter `end`. If we wish to analyse trends in the unemployment rate, annual data will suffice. The average (mean) over the twelve months of each year is another example of aggregated data, but this time we divide by 12 to give a mean annual rate.

```
> Maine.annual.ts <- aggregate(Maine.month.ts)/12
```

We now plot both time series. There is clear monthly variation. From Figure 1.3(a) it seems that the February figure is typically about 20% more than the annual average, whereas the August figure tends to be roughly 20% less.

```
> layout(1:2)
> plot(Maine.month.ts, ylab = "unemployed (%)")
> plot(Maine.annual.ts, ylab = "unemployed (%)")
```

We can calculate the precise percentages in `R`, using `window`. This function will extract that part of the time series between specified start and end points

and will sample with an interval equal to `frequency` if its argument is set to `TRUE`. So, the first line below gives a time series of February figures.

```
> Maine.Feb <- window(Maine.month.ts, start = c(1996,2), freq = TRUE)
> Maine.Aug <- window(Maine.month.ts, start = c(1996,8), freq = TRUE)
> Feb.ratio <- mean(Maine.Feb) / mean(Maine.month.ts)
> Aug.ratio <- mean(Maine.Aug) / mean(Maine.month.ts)

> Feb.ratio
[1] 1.223
> Aug.ratio
[1] 0.8164
```

On average, unemployment is 22% higher in February and 18% lower in August. An explanation is that Maine attracts tourists during the summer, and this creates more jobs. Also, the period before Christmas and over the New Year's holiday tends to have higher employment rates than the first few months of the new year. The annual unemployment rate was as high as 8.5% in 1976 but was less than 4% in 1988 and again during the three years 1999–2001. If we had sampled the data in August of each year, for example, rather than taken yearly averages, we would have consistently underestimated the unemployment rate by a factor of about 0.8.

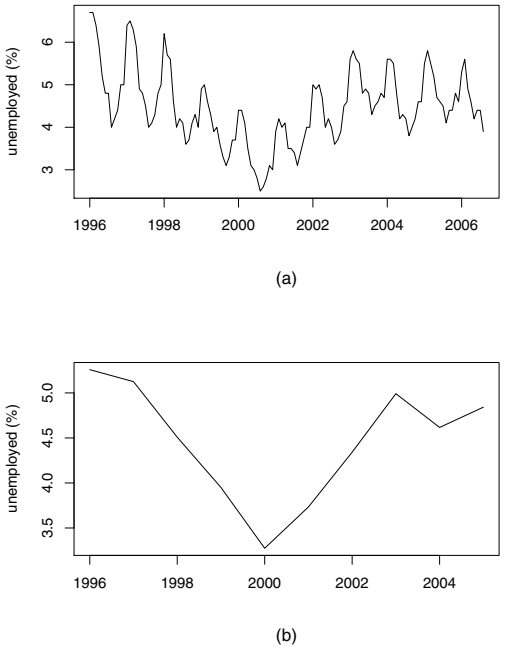


Fig. 1.3. Unemployment in Maine: (a) monthly January 1996–August 2006; (b) annual 1996–2005.

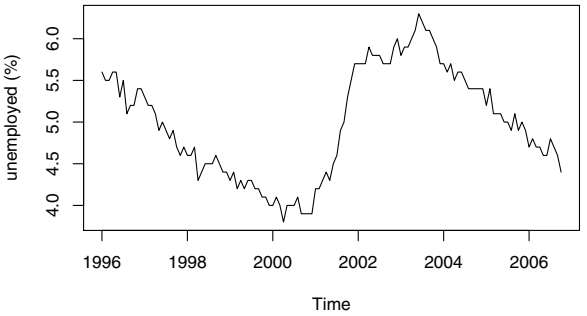


Fig. 1.4. Unemployment in the United States January 1996–October 2006.

The monthly unemployment rate for all of the United States from January 1996 until October 2006 is plotted in Figure 1.4. The decrease in the unemployment rate around the millennium is common to Maine and the United States as a whole, but Maine does not seem to be sharing the current US decrease in unemployment.

```
> www <- "http://www.massey.ac.nz/~pscowper/ts/USunemp.dat"
> US.month <- read.table(www, header = T)
> attach(US.month)
> US.month.ts <- ts(USun, start=c(1996,1), end=c(2006,10), freq = 12)
> plot(US.month.ts, ylab = "unemployed (%)")
```

1.4.3 Multiple time series: Electricity, beer and chocolate data

Here we illustrate a few important ideas and concepts related to *multiple* time series data. The monthly supply of electricity (millions of kWh), beer (Ml), and chocolate-based production (tonnes) in Australia over the period January 1958 to December 1990 are available from the Australian Bureau of Statistics (ABS).[6] The three series have been stored in a single file online, which can be read as follows:

```
www <- "http://www.massey.ac.nz/~pscowper/ts/cbe.dat"
CBE <- read.table(www, header = T)

> CBE[1:4, ]

  choc beer elec
1 1451 96.3 1497
2 2037 84.4 1463
3 2477 91.2 1648
4 2785 81.9 1595
```

[6] ABS data used with permission from the Australian Bureau of Statistics: http://www.abs.gov.au.

```
> class(CBE)

[1] "data.frame"
```

Now create time series objects for the electricity, beer, and chocolate data. If you omit `end`, `R` uses the full length of the vector, and if you omit the month in `start`, `R` assumes 1. You can use `plot` with `cbind` to plot several series on one figure (Fig. 1.5).

```
> Elec.ts <- ts(CBE[, 3], start = 1958, freq = 12)
> Beer.ts <- ts(CBE[, 2], start = 1958, freq = 12)
> Choc.ts <- ts(CBE[, 1], start = 1958, freq = 12)
> plot(cbind(Elec.ts, Beer.ts, Choc.ts))
```

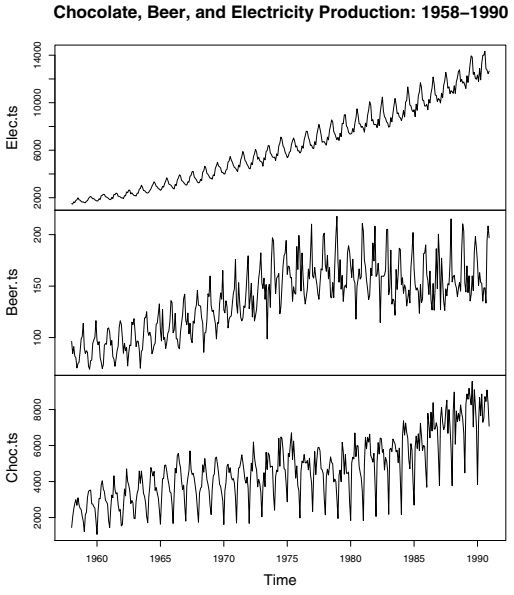


Fig. 1.5. Australian chocolate, beer, and electricity production; January 1958–December 1990.

The plots in Figure 1.5 show increasing trends in production for all three goods, partly due to the rising population in Australia from about 10 million to about 18 million over the same period (Fig. 1.6). But notice that electricity production has risen by a factor of 7, and chocolate production by a factor of 4, over this period during which the population has not quite doubled.

The three series constitute a *multiple* time series. There are many functions in `R` for handling more than one series, including `ts.intersect` to obtain the intersection of two series that overlap in time. We now illustrate the use of the intersect function and point out some potential pitfalls in analysing multiple

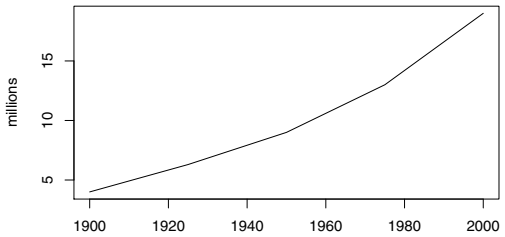


Fig. 1.6. Australia's population, 1900–2000.

time series. The intersection between the air passenger data and the electricity data is obtained as follows:

```
> AP.elec <- ts.intersect(AP, Elec.ts)
```

Now check that your output agrees with ours, as shown below.

```
> start(AP.elec)

[1] 1958    1

> end(AP.elec)

[1] 1960   12

> AP.elec[1:3, ]

      AP Elec.ts
[1,] 340    1497
[2,] 318    1463
[3,] 362    1648
```

In the code below, the data for each series are extracted and plotted (Fig. 1.7).[7]

```
> AP <- AP.elec[,1]; Elec <- AP.elec[,2]

> layout(1:2)
> plot(AP,   main = "", ylab = "Air passengers / 1000's")
> plot(Elec, main = "", ylab = "Electricity production / MkWh")

> plot(as.vector(AP), as.vector(Elec),
                      xlab = "Air passengers / 1000's",
                      ylab = "Electricity production / MWh")
> abline(reg = lm(Elec ~ AP))
```

[7] `R` is case sensitive, so lowercase is used here to represent the shorter record of air passenger data. In the code, we have also used the argument `main=""` to suppress unwanted titles.

```
> cor(AP, Elec)

[1] 0.884
```

In the `plot` function above, `as.vector` is needed to convert the `ts` objects to ordinary vectors suitable for a scatter plot.

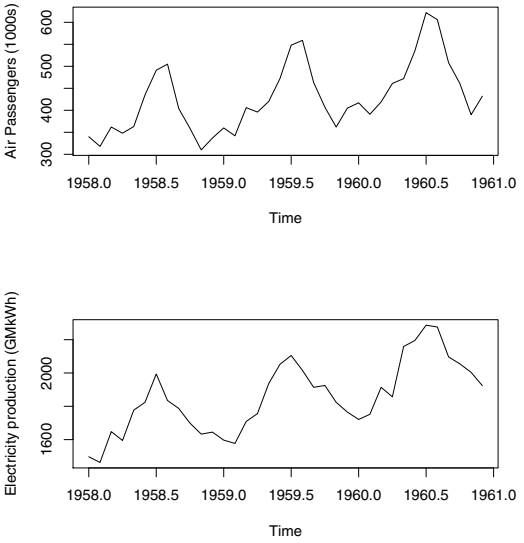


Fig. 1.7. International air passengers and Australian electricity production for the period 1958–1960. The plots look similar because both series have an increasing trend and a seasonal cycle. However, this does not imply that there exists a causal relationship between the variables.

The two time series are highly correlated, as can be seen in the plots, with a correlation coefficient of 0.88. Correlation will be discussed more in Chapter 2, but for the moment observe that the two time plots look similar (Fig. 1.7) and that the scatter plot shows an approximate linear association between the two variables (Fig. 1.8). However, it is important to realise that correlation does not imply causation. In this case, it is not plausible that higher numbers of air passengers in the United States cause, or are caused by, higher electricity production in Australia. A reasonable explanation for the correlation is that the increasing prosperity and technological development in both countries over this period accounts for the increasing trends. The two time series also happen to have similar seasonal variations. For these reasons, it is usually appropriate to remove trends and seasonal effects before comparing multiple series. This is often achieved by working with the residuals of a regression model that has deterministic terms to represent the trend and seasonal effects (Chapter 5).

In the simplest cases, the residuals can be modelled as independent random variation from a single distribution, but much of the book is concerned with fitting more sophisticated models.

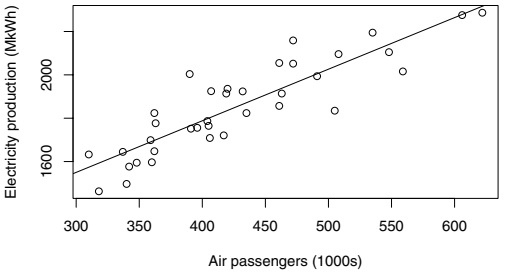


Fig. 1.8. Scatter plot of air passengers and Australian electricity production for the period: 1958–1960. The apparent linear relationship between the two variables is misleading and a consequence of the trends in the series.

1.4.4 Quarterly exchange rate: GBP to NZ dollar

The trends and seasonal patterns in the previous two examples were clear from the plots. In addition, reasonable explanations could be put forward for the possible causes of these features. With financial data, exchange rates for example, such marked patterns are less likely to be seen, and different methods of analysis are usually required. A financial series may sometimes show a dramatic change that has a clear cause, such as a war or natural disaster. Day-to-day changes are more difficult to explain because the underlying causes are complex and impossible to isolate, and it will often be unrealistic to assume any deterministic component in the time series model.

The exchange rates for British pounds sterling to New Zealand dollars for the period January 1991 to March 2000 are shown in Figure 1.9. The data are mean values taken over *quarterly* periods of three months, with the first quarter being January to March and the last quarter being October to December. They can be read into R from the book website and converted to a quarterly time series as follows:

```
> www <- "http://www.massey.ac.nz/~pscowper/ts/pounds_nz.dat"
> Z <- read.table(www, header = T)

> Z[1:4, ]

[1] 2.92 2.94 3.17 3.25

> Z.ts <- ts(Z, st = 1991, fr = 4)
```

```
> plot(Z.ts, xlab = "time / years",
             ylab = "Quarterly exchange rate in $NZ / pound")
```

Short-term trends are apparent in the time series: After an initial surge ending in 1992, a negative trend leads to a minimum around 1996, which is followed by a positive trend in the second half of the series (Fig. 1.9).

The trend seems to change direction at unpredictable times rather than displaying the relatively consistent pattern of the air passenger series and Australian production series. Such trends have been termed *stochastic trends* to emphasise this randomness and to distinguish them from more *deterministic* trends like those seen in the previous examples. A mathematical model known as a *random walk* can sometimes provide a good fit to data like these and is fitted to this series in §4.4.2. Stochastic trends are common in financial series and will be studied in more detail in Chapters 4 and 7.

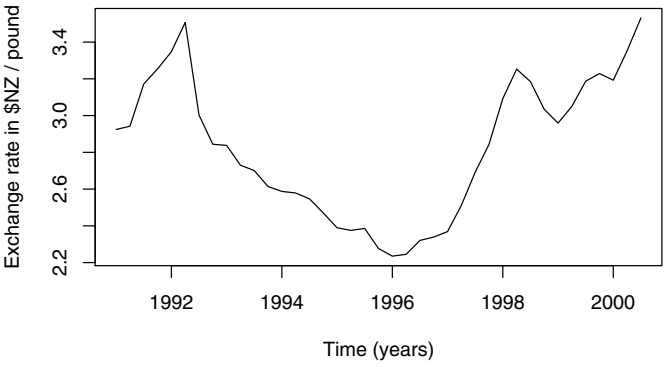


Fig. 1.9. Quarterly exchange rates for the period 1991–2000.

Two local trends are emphasised when the series is partitioned into two subseries based on the periods 1992–1996 and 1996–1998. The `window` function can be used to extract the subseries:

```
> Z.92.96 <- window(Z.ts, start = c(1992, 1), end = c(1996, 1))
> Z.96.98 <- window(Z.ts, start = c(1996, 1), end = c(1998, 1))

> layout (1:2)
> plot(Z.92.96, ylab = "Exchange rate in $NZ/pound",
                xlab = "Time (years)" )
> plot(Z.96.98, ylab = "Exchange rate in $NZ/pound",
                xlab = "Time (years)" )
```

Now suppose we were observing this series at the start of 1992; i.e., we had the data in Figure 1.10(a). It might have been tempting to predict a

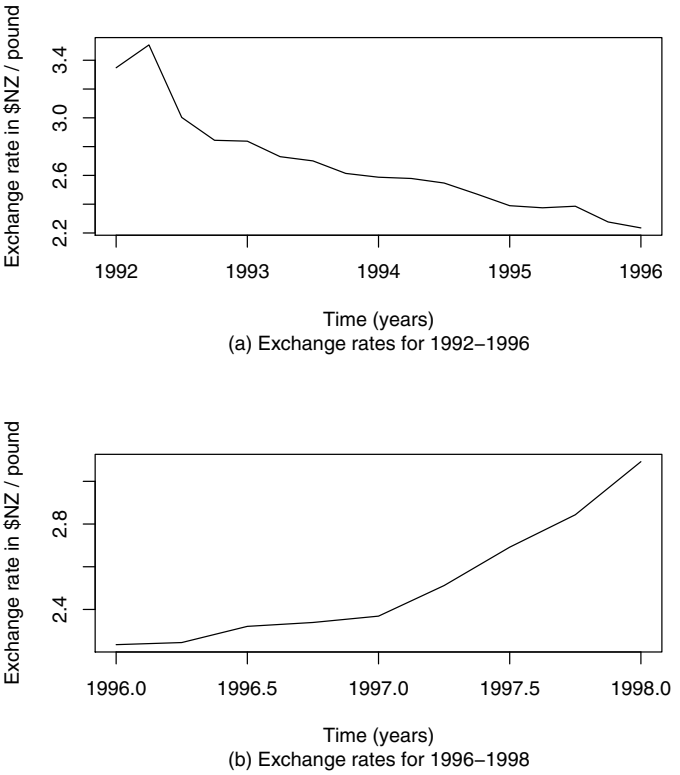


Fig. 1.10. Quarterly exchange rates for two periods. The plots indicate that without additional information it would be inappropriate to extrapolate the trends.

continuation of the downward trend for future years. However, this would have been a very poor prediction, as Figure 1.10(b) shows that the data started to follow an increasing trend. Likewise, without additional information, it would also be inadvisable to extrapolate the trend in Figure 1.10(b). This illustrates the potential pitfall of inappropriate extrapolation of stochastic trends when underlying causes are not properly understood. To reduce the risk of making an inappropriate forecast, statistical tests, introduced in Chapter 7, can be used to test for a stochastic trend.

1.4.5 Global temperature series

A change in the world's climate will have a major impact on the lives of many people, as global warming is likely to lead to an increase in ocean levels and natural hazards such as floods and droughts. It is likely that the world economy will be severely affected as governments from around the globe try

to enforce a reduction in fossil fuel use and measures are taken to deal with any increase in natural disasters.[8]

In climate change studies (e.g., see Jones and Moberg, 2003; Rayner et al. 2003), the following global temperature series, expressed as anomalies from the monthly means over the period 1961–1990, plays a central role:[9]

```
> www <- "http://www.massey.ac.nz/~pscowper/ts/global.dat"
> Global <- scan(www)
> Global.ts <- ts(Global, st = c(1856, 1), end = c(2005, 12),
     fr = 12)
> Global.annual <- aggregate(Global.ts, FUN = mean)
> plot(Global.ts)
> plot(Global.annual)
```

It is the trend that is of most concern, so the `aggregate` function is used to remove any seasonal effects within each year and produce an annual series of mean temperatures for the period 1856 to 2005 (Fig. 1.11b). We can avoid explicitly dividing by 12 if we specify `FUN=mean` in the `aggregate` function.

The upward trend from about 1970 onwards has been used as evidence of global warming (Fig. 1.12). In the code below, the monthly time intervals corresponding to the 36-year period 1970–2005 are extracted using the `time` function and the associated observed temperature series extracted using `window`. The data are plotted and a line superimposed using a regression of temperature on the new time index (Fig. 1.12).

```
> New.series <- window(Global.ts, start=c(1970, 1), end=c(2005, 12))
> New.time <- time(New.series)
> plot(New.series); abline(reg=lm(New.series ~ New.time))
```

In the previous section, we discussed a potential pitfall of inappropriate extrapolation. In climate change studies, a vital question is whether rising temperatures are a consequence of human activity, specifically the burning of fossil fuels and increased greenhouse gas emissions, or are a natural trend, perhaps part of a longer cycle, that may decrease in the future without needing a global reduction in the use of fossil fuels. We cannot attribute the increase in global temperature to the increasing use of fossil fuels without invoking some physical explanation[10] because, as we noted in §1.4.3, two unrelated time series will be correlated if they both contain a trend. However, as the general consensus among scientists is that the trend in the global temperature series is related to a global increase in greenhouse gas emissions, it seems reasonable to

[8] For general policy documents and discussions on climate change, see the website (and links) for the United Nations Framework Convention on Climate Change at http://unfccc.int.

[9] The data are updated regularly and can be downloaded free of charge from the Internet at: http://www.cru.uea.ac.uk/cru/data/.

[10] For example, refer to US Energy Information Administration at http://www.eia.doe.gov/emeu/aer/inter.html.

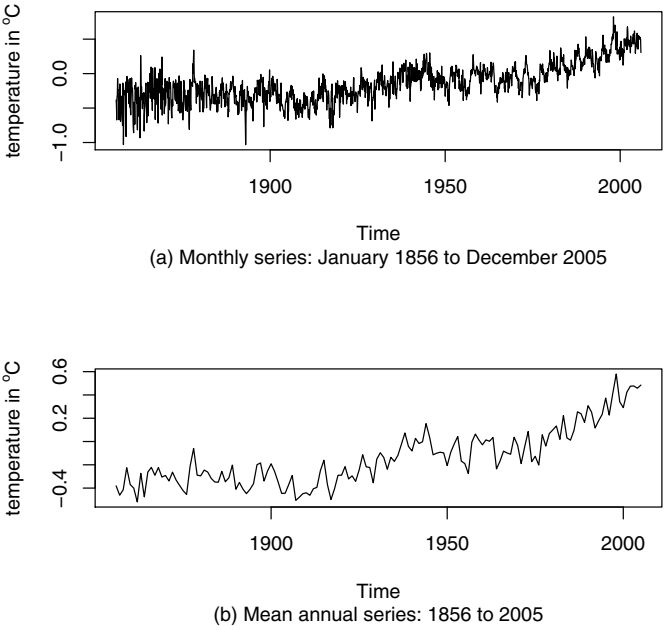


Fig. 1.11. Time plots of the global temperature series (°C).

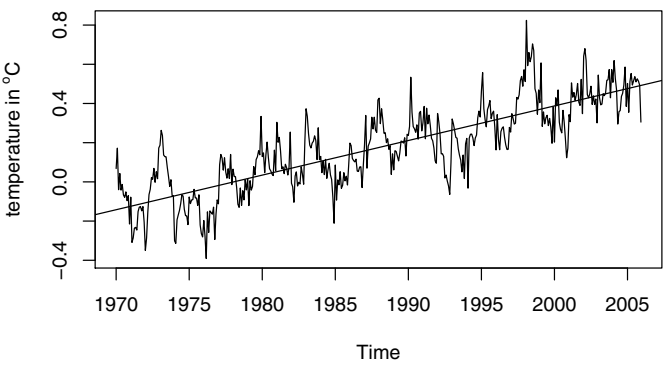


Fig. 1.12. Rising mean global temperatures, January 1970–December 2005. According to the United Nations Framework Convention on Climate Change, the mean global temperature is expected to continue to rise in the future unless greenhouse gas emissions are reduced on a global scale.

acknowledge a causal relationship and to expect the mean global temperature to continue to rise if greenhouse gas emissions are not reduced.[11]

1.5 Decomposition of series

1.5.1 Notation

So far, our analysis has been restricted to plotting the data and looking for features such as trend and seasonal variation. This is an important first step, but to progress we need to fit time series models, for which we require some notation. We represent a time series of length n by $\{x_t : t = 1, \ldots, n\} = \{x_1, x_2, \ldots, x_n\}$. It consists of n values sampled at discrete times $1, 2, \ldots, n$. The notation will be abbreviated to $\{x_t\}$ when the length n of the series does not need to be specified. The time series model is a sequence of random variables, and the observed time series is considered a realisation from the model. We use the same notation for both and rely on the context to make the distinction.[12] An overline is used for sample means:

$$\bar{x} = \sum x_i/n \tag{1.1}$$

The 'hat' notation will be used to represent a *prediction* or *forecast*. For example, with the series $\{x_t : t = 1, \ldots, n\}$, $\hat{x}_{t+k|t}$ is a *forecast* made at time t for a future value at time $t + k$. A forecast is a predicted future value, and the number of time steps into the future is the *lead time* (k). Following our convention for time series notation, $\hat{x}_{t+k|t}$ can be the random variable or the numerical value, depending on the context.

1.5.2 Models

As the first two examples showed, many series are dominated by a trend and/or seasonal effects, so the models in this section are based on these components. A simple *additive decomposition* model is given by

$$x_t = m_t + s_t + z_t \tag{1.2}$$

where, at time t, x_t is the observed series, m_t is the trend, s_t is the seasonal effect, and z_t is an error term that is, in general, a sequence of correlated random variables with mean zero. In this section, we briefly outline two main approaches for extracting the trend m_t and the seasonal effect s_t in Equation (1.2) and give the main `R` functions for doing this.

[11] Refer to http://unfccc.int.

[12] Some books do distinguish explicitly by using lowercase for the time series and uppercase for the model.

If the seasonal effect tends to increase as the trend increases, a multiplicative model may be more appropriate:

$$x_t = m_t \cdot s_t + z_t \tag{1.3}$$

If the random variation is modelled by a multiplicative factor and the variable is positive, an additive decomposition model for $\texttt{log}(x_t)$ can be used:[13]

$$\log(x_t) = m_t + s_t + z_t \tag{1.4}$$

Some care is required when the exponential function is applied to the predicted mean of $\texttt{log}(x_t)$ to obtain a prediction for the mean value x_t, as the effect is usually to bias the predictions. If the random series z_t are normally distributed with mean 0 and variance σ^2, then the predicted mean value at time t based on Equation (1.4) is given by

$$\hat{x}_t = e^{m_t + s_t} e^{\frac{1}{2}\sigma^2} \tag{1.5}$$

However, if the error series is not normally distributed and is negatively skewed,[14] as is often the case after taking logarithms, the bias correction factor will be an overcorrection (Exercise 4) and it is preferable to apply an empirical adjustment (which is discussed further in Chapter 5). The issue is of practical importance. For example, if we make regular financial forecasts without applying an adjustment, we are likely to consistently underestimate mean costs.

1.5.3 Estimating trends and seasonal effects

There are various ways to estimate the trend m_t at time t, but a relatively simple procedure, which is available in `R` and does not assume any specific form is to calculate a *moving average* centred on x_t. A moving average is an average of a specified number of time series values around each value in the time series, with the exception of the first few and last few terms. In this context, the length of the moving average is chosen to average out the seasonal effects, which can be estimated later. For monthly series, we need to average twelve consecutive months, but there is a slight snag. Suppose our time series begins at January ($t = 1$) and we average January up to December ($t = 12$). This average corresponds to a time $t = 6.5$, between June and July. When we come to estimate seasonal effects, we need a moving average at integer times. This can be achieved by averaging the average of January up to December and the average of February ($t = 2$) up to January ($t = 13$). This average of

[13] To be consistent with `R`, we use `log` for the natural logarithm, which is often written `ln`.

[14] A probability distribution is negatively skewed if its density has a long tail to the left.

two moving averages corresponds to $t = 7$, and the process is called centring. Thus the trend at time t can be estimated by the centred moving average

$$\hat{m}_t = \frac{\frac{1}{2}x_{t-6} + x_{t-5} + \ldots + x_{t-1} + x_t + x_{t+1} + \ldots + x_{t+5} + \frac{1}{2}x_{t+6}}{12} \quad (1.6)$$

where $t = 7, \ldots, n-6$. The coefficients in Equation (1.6) for each month are 1/12 (or sum to 1/12 in the case of the first and last coefficients), so that equal weight is given to each month and the coefficients sum to 1. By using the seasonal frequency for the coefficients in the moving average, the procedure generalises for any seasonal frequency (e.g., quarterly series), provided the condition that the coefficients sum to unity is still met.

An estimate of the monthly additive effect (s_t) at time t can be obtained by subtracting $\hat{m}_t$:

$$\hat{s}_t = x_t - \hat{m}_t \quad (1.7)$$

By averaging these estimates of the monthly effects for each month, we obtain a single estimate of the effect for each month. If the period of the time series is a whole number of years, the number of monthly effects averaged for each month is one less than the number of years of record. At this stage, the twelve monthly additive components should have an average value close to, but not usually exactly equal to, zero. It is usual to adjust them by subtracting this mean so that they do average zero. If the monthly effect is multiplicative, the estimate is given by division; i.e., $\hat{s}_t = x_t/\hat{m}_t$. It is usual to adjust monthly multiplicative factors so that they average unity. The procedure generalises, using the same principle, to any seasonal frequency.

It is common to present economic indicators, such as unemployment percentages, as *seasonally adjusted* series. This highlights any trend that might otherwise be masked by seasonal variation attributable, for instance, to the end of the academic year, when school and university leavers are seeking work. If the seasonal effect is additive, a seasonally adjusted series is given by $x_t - \bar{s}_t$, whilst if it is multiplicative, an adjusted series is obtained from $x_t/\bar{s}_t$, where $\bar{s}_t$ is the seasonally adjusted mean for the month corresponding to time t.

1.5.4 Smoothing

The centred moving average is an example of a *smoothing* procedure that is applied retrospectively to a time series with the objective of identifying an underlying signal or trend. Smoothing procedures can, and usually do, use points before and after the time at which the smoothed estimate is to be calculated. A consequence is that the smoothed series will have some points missing at the beginning and the end unless the smoothing algorithm is adapted for the end points.

A second smoothing algorithm offered by `R` is `stl`. This uses a locally weighted regression technique known as *loess*. The regression, which can be a line or higher polynomial, is referred to as local because it uses only some

relatively small number of points on either side of the point at which the smoothed estimate is required. The weighting reduces the influence of outlying points and is an example of robust regression. Although the principles behind `stl` are straightforward, the details are quite complicated.

Smoothing procedures such as the centred moving average and loess do not require a predetermined model, but they do not produce a formula that can be extrapolated to give forecasts. Fitting a line to model a linear trend has an advantage in this respect.

The term *filtering* is also used for smoothing, particularly in the engineering literature. A more specific use of the term filtering is the process of obtaining the best estimate of some variable now, given the latest measurement of it and past measurements. The measurements are subject to random error and are described as being *corrupted by noise*. Filtering is an important part of control algorithms which have a myriad of applications. An exotic example is the Huygens probe leaving the Cassini orbiter to land on Saturn's largest moon, Titan, on January 14, 2005.

1.5.5 Decomposition in R

In R, the function `decompose` estimates trends and seasonal effects using a moving average method. Nesting the function within `plot` (e.g., using `plot(stl())`) produces a single figure showing the original series x_t and the decomposed series m_t, s_t, and z_t. For example, with the electricity data, additive and multiplicative decomposition plots are given by the commands below; the last plot, which uses `lty` to give different line types, is the superposition of the seasonal effect on the trend (Fig. 1.13).

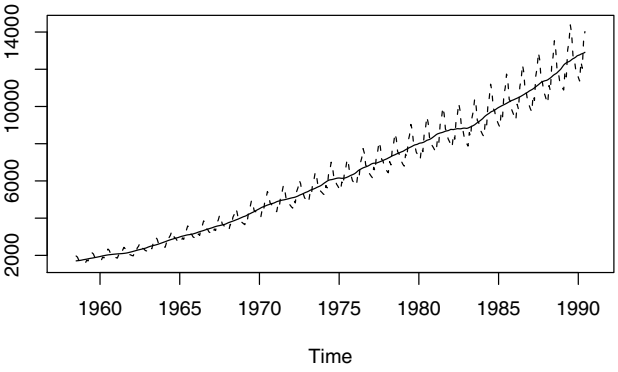


Fig. 1.13. Electricity production data: trend with superimposed multiplicative seasonal effects.

```
> plot(decompose(Elec.ts))
> Elec.decom <- decompose(Elec.ts, type = "mult")
> plot(Elec.decom)
> Trend <- Elec.decom$trend
> Seasonal <- Elec.decom$seasonal
> ts.plot(cbind(Trend, Trend * Seasonal), lty = 1:2)
```

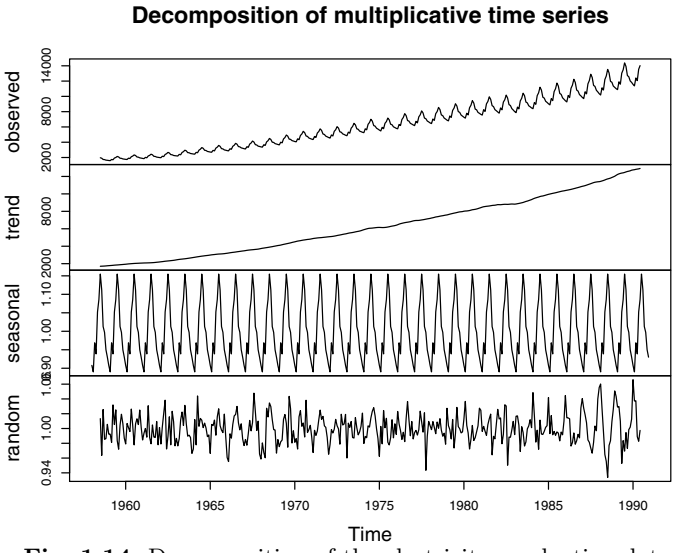


Fig. 1.14. Decomposition of the electricity production data.

In this example, the multiplicative model would seem more appropriate than the additive model because the variance of the original series and trend increase with time (Fig. 1.14). However, the random component, which corresponds to z_t, also has an increasing variance, which indicates that a log-transformation (Equation (1.4)) may be more appropriate for this series (Fig. 1.14). The `random` series obtained from the `decompose` function is not precisely a realisation of the random process z_t but rather an estimate of that realisation. It is an estimate because it is obtained from the original time series using estimates of the trend and seasonal effects. This estimate of the realisation of the random process is a *residual error series*. However, we treat it as a realisation of the random process.

There are many other reasonable methods for decomposing time series, and we cover some of these in Chapter 5 when we study regression methods.

1.6 Summary of commands used in examples

`read.table`	reads data into a data frame
`attach`	makes names of column variables available
`ts`	produces a time series object
`aggregate`	creates an aggregated series
`ts.plot`	produces a time plot for one or more series
`window`	extracts a subset of a time series
`time`	extracts the time from a time series object
`ts.intersect`	creates the intersection of one or more time series
`cycle`	returns the season for each value in a series
`decompose`	decomposes a series into the components trend, seasonal effect, and residual
`stl`	decomposes a series using loess smoothing
`summary`	summarises an `R` object

1.7 Exercises

1. Carry out the following exploratory time series analysis in `R` using either the chocolate or the beer production data from §1.4.3.
   a) Produce a time plot of the data. Plot the aggregated annual series and a boxplot that summarises the observed values for each season, and comment on the plots.
   b) Decompose the series into the components trend, seasonal effect, and residuals, and plot the decomposed series. Produce a plot of the trend with a superimposed seasonal effect.

2. Many economic time series are based on indices. A price index is the ratio of the cost of a basket of goods now to its cost in some base year. In the Laspeyre formulation, the basket is based on typical purchases in the base year. You are asked to calculate an index of motoring cost from the following data. The clutch represents all mechanical parts, and the quantity allows for this.

item (i)	quantity '00 (q_{i0})	unit price '00 (p_{i0})	quantity '04 (q_{it})	unit price '04 (p_{it})
car	0.33	18 000	0.5	20 000
petrol (litre)	2 000	0.80	1 500	1.60
servicing (h)	40	40	20	60
tyre	3	80	2	120
clutch	2	200	1	360

The *Laspeyre Price Index* at time t relative to base year 0 is

$$LI_t = \frac{\sum q_{i0} p_{it}}{\sum q_{i0} p_{i0}}$$

Calculate the LI_t for 2004 relative to 2000.

3. The *Paasche Price Index* at time t relative to base year 0 is

$$PI_t = \frac{\sum q_{it} p_{it}}{\sum q_{it} p_{i0}}$$

   a) Use the data above to calculate the PI_t for 2004 relative to 2000.
   b) Explain why the PI_t is usually lower than the LI_t.
   c) Calculate the *Irving-Fisher Price Index* as the geometric mean of LI_t and PI_t. (The geometric mean of a sample of n items is the nth root of their product.)

4. A standard procedure for finding an approximate mean and variance of a function of a variable is to use a Taylor expansion for the function about the mean of the variable. Suppose the variable is y and that its mean and standard deviation are μ and σ respectively.

$$\phi(y) = \phi(\mu) + \phi'(\mu)(y-\mu) + \phi''(\mu)\frac{(y-\mu)^2}{2!} + \phi'''(\mu)\frac{(y-\mu)^3}{3!} + \ldots$$

   Consider the case of $\phi(.)$ as $e^{(.)}$. By taking the expectation of both sides of this equation, explain why the bias correction factor given in Equation (1.5) is an overcorrection if the residual series has a negative skewness, where the *skewness* γ of a random variable y is defined by

$$\gamma = \frac{E\left[(y-\mu)^3\right]}{\sigma^3}$$