Chapter 8

Time Series and Forecasting

8.1 Introduction

A time series is a collection of observations made sequentially in time. When observations are made continuously, the time series is said to be continuous; when observations are taken only at specific time points, the time series is said to be discrete. In this course we consider only discrete time series, where the observations are taken at equal intervals. The observations $x_1, x_2, \ldots, x_n$ should be regarded as the realisation of random quantities $X_1, X_2, \ldots, X_n$.

The first step in the analysis of time series is usually to plot the data against time, in a time series plot.

8.1.1 Time series plots in Minitab

Once the data has been entered in the Minitab worksheet (or loaded from a data file), you can plot the data, or analyse it, using one of Minitab’s built-in time series analysis functions. Selecting Graph – Time Series Plot – Simple, and then entering the column which contains your data in Series will produce a simple plot of your data against time. However, if the Time/Scale option is chosen, you can select the period over which your data has been collected (e.g. hourly, quarterly, monthly); you can also enter the start year (e.g. 1970) and the start month (e.g. 7 for July) to produce better labels for your time axis. Using the Labels option will also allow you to include a relevant title for your time series plot. Figure 8.1 (left-hand-side) shows a time series plot of monthly average air temperatures in England and Wales between 1970 and 1978 (inclusive). The right-hand-side plot in the same figure shows the same time series plot, but with the time axis labelled correctly (using the Time/Scale option in Minitab). The plot on the right-hand-side also has a title added.

Figures 8.2, 8.3 and 8.4 show time series plots of some other datasets; figure 8.2 shows a time series plot of the monthly means of daily relative sunspot numbers, figure 8.3 shows a time series plot of the (four-monthly) sales of a department store (between 1994 and 1997), and figure 8.4 shows the total precipitation in New York City each year from 1870
to 1968.

Figure 8.1: Left: Time series plots of monthly average air temperatures in England and Wales (right-hand-side plot shows edited time axis and inserted title

8.1.2 Example: sales figures for a department store

Suppose we have the following four-monthly sales figures (in thousands of pounds) for a department store:

<table>
<thead>
<tr>
<th></th>
<th>Jan–Apr</th>
<th>May–Aug</th>
<th>Sep–Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>8</td>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>1995</td>
<td>10</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>1996</td>
<td>10</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>1997</td>
<td>11</td>
<td>16</td>
<td>13</td>
</tr>
</tbody>
</table>

Figure 8.3 shows a time series plot of these data; from this plot, we can see that the sales figures clearly exhibit seasonal patterns, with sales peaking in the second “season” (May–August) and dipping either side of this. We can also see that, overall, sales have increased through time, with 1997 showing the highest sales figures for all seasons. Thus, there is also a clear increasing trend in the series. The next two sections will demonstrate how both the trend and seasonal patterns can be estimated.
Figure 8.2: Time series plot of the monthly means of daily relative sunspot numbers

Figure 8.3: Time series plot of four-monthly sales for a department store

Figure 8.4: Time series plot of yearly total precipitation in New York
8.2 Isolating the trend

There are several methods we could use for isolating the trend. The method we will study is based on the notion of moving averages. To calculate a moving average, we simply average over the cycle around an observation. For example, for the department store sales data, we have three “seasons” (Jan–Apr, May–Aug and Sep–Dec) and so a full cycle consists of three observations. Thus, to calculate the first moving average we would take the first three values of the time series and calculate their mean, i.e.

\[
\frac{8 + 13 + 10}{3} = 10.33.
\]

Similarly, the second moving average is

\[
\frac{13 + 10 + 10}{3} = 11.
\]

The rest of the moving averages have been calculated in this way, and are shown in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Jan–Apr</th>
<th>May–Aug</th>
<th>Sep–Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>10.33</td>
<td>11.00</td>
<td>*</td>
</tr>
<tr>
<td>1995</td>
<td>11.33</td>
<td>11.67</td>
<td>11.67</td>
</tr>
<tr>
<td>1996</td>
<td>12.00</td>
<td>12.00</td>
<td>12.33</td>
</tr>
<tr>
<td>1997</td>
<td>12.67</td>
<td>13.33</td>
<td>*</td>
</tr>
</tbody>
</table>

Obviously, there’s no moving average associated with the first and last data points, as there’s no observation before the first, or after the last, in order to calculate the moving average at these points! The length of the cycle over which to average is often obvious; for example, much data is presented quarterly or monthly, and that can provide a natural cycle around which to base the process. In our example, we have three clearly defined “seasons”, and so a cycle of length 3 would seem like the obvious choice. Figure 8.5 shows a plot of the original data with the moving averages overlaid. This clearly shows the upward rise in sales over the four years.

These moving averages represent the trend in our time series. We might be interested only in this trend, and wish to project this trend into the future – in our example, we might want to forecast sales for Jan–Apr 1998, or maybe even further ahead. Note the linearity in the plot of our moving averages in figure 8.5; in chapter 7 you were introduced to a technique known as linear regression, which calculates the equation of the line of best fit for a scatter plot of paired data. If we consider our units of time to be the explanatory variable, and if a linear relationship exists between our time variable and the data (which it seems to in our example), we could use linear regression techniques to estimate a straight line trend in our data. This straight line trend could then be extended into the future to make forecasts of sales at future time points.

Recall from chapter 6 that the simple linear regression model is given by

\[
Y = \alpha + \beta X + \epsilon,
\]
where $Y$ is the response variable, $X$ is the explanatory variable, and $\epsilon$ is a random error. We now reformulate this model so that

$$Y = \alpha + \beta T + \epsilon,$$

where $T$ is the “time point”. The first observation in the series has a time point of 1 (i.e. $T = 1$), the second has a time point of $T = 2$, and so on. So if we wanted to predict sales in the period Jan–Apr 1998, we would substitute $T = 13$ into the regression equation, since the last observed time point was for Sep–Dec 1997 and was the 12th time point. But, before we do this, we need to estimate $\alpha$ and $\beta$!

The formulae for $\alpha$ and $\beta$ are exactly the same as in chapter 7, but now $X$ is replaced with $T$, i.e.

$$\hat{\beta} = \frac{S_{TY}}{S_{TT}} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{t},$$

where

$$S_{TY} = \left( \sum ty \right) - n\bar{t}\bar{y} \quad \text{and} \quad S_{TT} = \left( \sum t^2 \right) - n\bar{t}^2.$$ 

Remember, the easiest way to calculate these quantities is to draw up a table! note that for $y$ we use the moving average values, not the actual observations! Notice also that we don’t have any observations at time points 1 and 12, since we were unable to calculate moving averages here.
Thus,

\[ \bar{t} = \frac{\sum t}{n} \]

\[ \bar{y} = \frac{\sum y}{n} \]

Similarly,

\[ S_{TY} = \left( \sum ty \right) - n \bar{t} \bar{y} \]

\[ S_{TT} = \left( \sum t^2 \right) - n \bar{t}^2 \]

Thus,

\[ \hat{\beta} = \frac{S_{TY}}{S_{TT}} \]

\[ \hat{\alpha} = \bar{y} - \hat{\beta} \bar{t} \]

So, the regression equation for our trend is

\[ y = \]
8.3 Isolating the seasonal effects

We have seen how we can use simple averaging over the cycle around an observation to isolate the trend in our series, and how to estimate this trend using the linear regression techniques discussed in chapter 7. We can use the linear trend equation to make forecasts of future sales, but these forecasted values will assume sales increase linearly only (since the regression equation is an equation of a straight line). Going back to the time series plot of our data (figures 8.3 and 8.5), we can see that the underlying level of the sales data does increase in a linear fashion, but there are clear cycles around this increase. In this section we will discuss how to estimate these so-called seasonal effects.

First of all, we calculate the seasonal deviations by subtracting the moving average for each observation from the original observation. Again, we cannot calculate values for the first and last observations. For example, the seasonal deviation for May–Aug 1994 is found as

\[ 13 - 10.33 = 2.67. \]

Similarly, for Sep–Dec 1994, we have

\[ 10.00 - 11.00 = -1. \]

The other seasonal deviations, along with the seasonal means, are shown in the table below:

<table>
<thead>
<tr>
<th></th>
<th>Jan–Apr</th>
<th>May–Aug</th>
<th>Sep–Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>*</td>
<td>2.67</td>
<td>-1</td>
</tr>
<tr>
<td>1995</td>
<td>-1.33</td>
<td>2.33</td>
<td>-0.67</td>
</tr>
<tr>
<td>1996</td>
<td>-2</td>
<td>3</td>
<td>-1.33</td>
</tr>
<tr>
<td>1997</td>
<td>-1.67</td>
<td>2.67</td>
<td>*</td>
</tr>
<tr>
<td>means</td>
<td>-1.67</td>
<td>2.6675</td>
<td>-1</td>
</tr>
</tbody>
</table>

We can now calculate the seasonal effects. The seasonal effect for each season is the seasonal mean for that season minus the overall mean. The overall mean from the table above is found as

\[
\frac{2.67 - 1 - 1.33 + \ldots + 2.67}{10} = \frac{2.67}{10} = 0.267.
\]

Thus, the seasonal effects for each season are

\[ \hat{s}_1 = -1.67 - 0.267 = -1.937 \]

\[ \hat{s}_2 = 2.6675 - 0.267 = 2.4005 \]

\[ \hat{s}_3 = -1 - 0.267 = -1.267. \]
Note that the seasonal effects should add up to give zero. Ours don’t – we have
\[ \hat{s}_1 + \hat{s}_2 + \hat{s}_3 = -1.937 + 2.4005 - 1.267 \]
\[ = -0.8035. \]

Thus, we have to make an adjustment so they do add up to give zero. To do this, we find
the mean of our seasonal effects, and then subtract this from each of the seasonal effects.
In this example, the mean of our seasonal effects is
\[ \frac{-1.937 + 2.4005 - 1.267}{3} = \frac{-0.8035}{3} = -0.26783. \]

Thus, if we subtract this from each of the seasonal effects, the adjusted seasonal effects
will then add up to give zero. Thus, the adjusted seasonal effects are
\[
\hat{s}_1 = -1.937 - (-0.26783) \\
= -1.66917 \\

\hat{s}_2 = 2.4005 - (-0.26783) \\
= 2.66833 \\

\hat{s}_3 = -1.267 - (-0.26783) \\
= -0.99917.
\]

Just be careful with double negatives! Now the seasonal effects do sum to give zero!

### 8.4 Forecasting

There are many ways in which we can forecast future observations. One way is to use
the linear regression equation for the trend in our series. For the department store sales
data, recall that this was
\[ y = 10.043 + 0.275t. \]

To predict average sales in Jan–Apr 1998, we would substitute \( t = 13 \) into the above
equation, since this would be our 13th observation. Doing so, gives
\[ y = 10.043 + 0.275 \times 13 \]
\[ = 10.043 + 3.575 \]
\[ = 13.618. \]

However, we’re not quite done yet! This assumes that our data follow a straight line;
looking at the time series plot in figure 8.3, clear cycles around an increasing trend can be
seen. Thus, we now need to add in our seasonal effect. The seasonal effect for Jan–Apr
is \( \hat{s}_1 = -1.66917 \). Thus, our “full” forecast for sales in Jan–Apr 1998 is
\[ 13.618 + (-1.66917) = 11.949, \]
i.e. £11,949, or just under £12,000.
8.5 Exercises

Consider the following data for sales (in thousands of pounds) at a newly opened sandwich shop on a large industrial estate. The sandwich shop only opens five days a week as many of the factories shut at weekends.

<table>
<thead>
<tr>
<th>Week beginning</th>
<th>Mon</th>
<th>Tues</th>
<th>Wed</th>
<th>Thurs</th>
<th>Fri</th>
</tr>
</thead>
<tbody>
<tr>
<td>21/02/05</td>
<td>28</td>
<td>16</td>
<td>23</td>
<td>44</td>
<td>54</td>
</tr>
<tr>
<td>28/02/05</td>
<td>31</td>
<td>22</td>
<td>27</td>
<td>51</td>
<td>60</td>
</tr>
<tr>
<td>07/03/05</td>
<td>35</td>
<td>25</td>
<td>30</td>
<td>51</td>
<td>65</td>
</tr>
</tbody>
</table>

(a) Produce a time series plot of the sales data, either by hand or using Minitab. Remember to label your axes and give your plot a suitable title.

(b) Comment on your time series plot in part (a).

(c) Calculate the moving averages for the sales data, based on a five–observation cycle (Hint: Your first moving average will be for Wednesday 23/02/05, and is \((28 + 16 + 23 + 44 + 54)/5 = 33\)).

(d) Using the moving averages obtained in part (c), find the linear trend

\[ y = \hat{\alpha} + \hat{\beta}t. \]

(e) Calculate the seasonal deviations by subtracting the moving average for each observation from the original observation, and calculate the seasonal means.

(f) Calculate the overall mean, and hence calculate the seasonal effects for Monday, \ldots, Friday (Remember, if your seasonal effects don’t sum to give zero, you need to make a suitable adjustment).

(g) Use the regression equation obtained in part (d), and the seasonal effects in part (f), to forecast sales at the sandwich shop on Monday 14th March 2005 and Tuesday 15th March 2005.