

Chapter 2

Time series and Forecasting

2.1 Introduction

Data are frequently recorded at regular time intervals, for instance, daily stock market indices, the monthly rate of inflation or annual profit figures. In this Chapter we think about how to display and model such data. We will consider how to detect trends and seasonal effects and then use these to make forecasts. As well as review the methods covered in MAS1403, we will also consider a class of time series *models* known as *autoregressive moving average* models. Why is this topic useful? Well, making forecasts allows organisations to make better decisions and to plan more efficiently. For instance, reliable forecasts enable a retail outlet to anticipate demand, hospitals to plan staffing levels and manufacturers to keep appropriate levels of inventory.

2.2 Displaying and describing time series

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 first step in the analysis of time series is usually to plot the data against time, in a *time series plot*. Suppose we have the following four-monthly sales figures for *Turner's Hangover Cure* as described in Practical 2 (in thousands of pounds):

	Jan–Apr	May–Aug	Sep–Dec
2006	8	10	13
2007	10	11	14
2008	10	11	15
2009	11	13	16

We could enter these data into a single column (say column C1) in Minitab, and then click on **Graph–Time Series Plot–Simple–OK**; entering C1 in **Series** and then clicking **OK** gives the graph shown in figure 2.1.

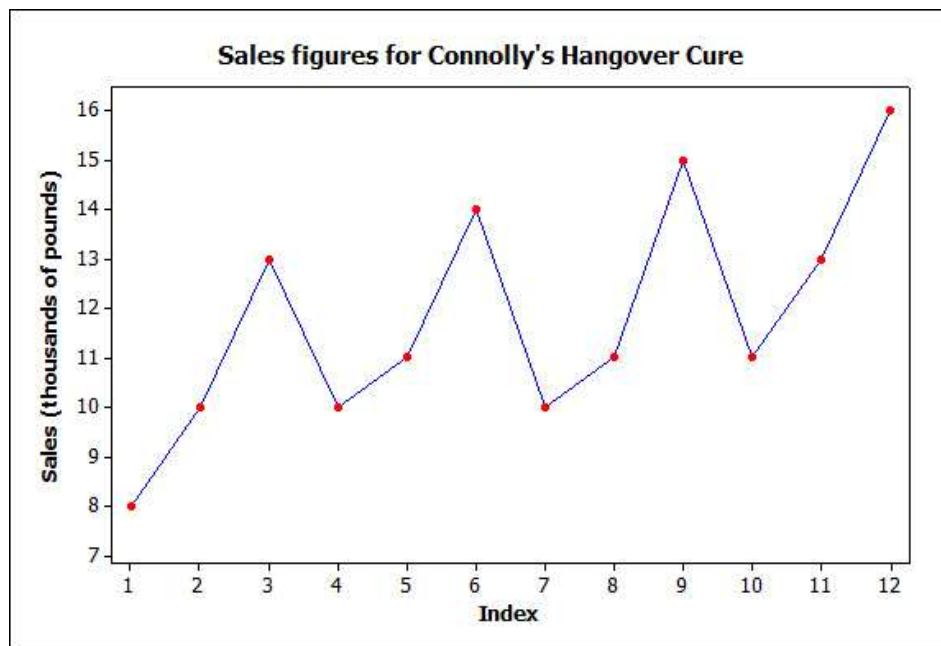


Figure 2.1: Time series plot showing sales figures for *Turner's Hangover Cure*

Notice that this is very similar to a scatterplot; however,

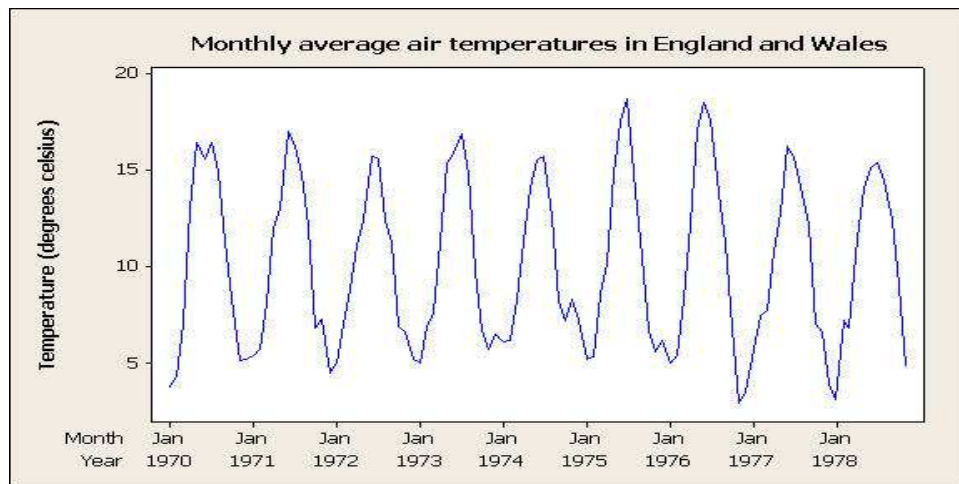
- the x -axis now represents time;
- we join together successive points in the plot.

Also notice that the time axis is not conveniently labelled; for example, it doesn't show the years. We will look at how to change the appearance of such plots in Minitab in Practical 3.

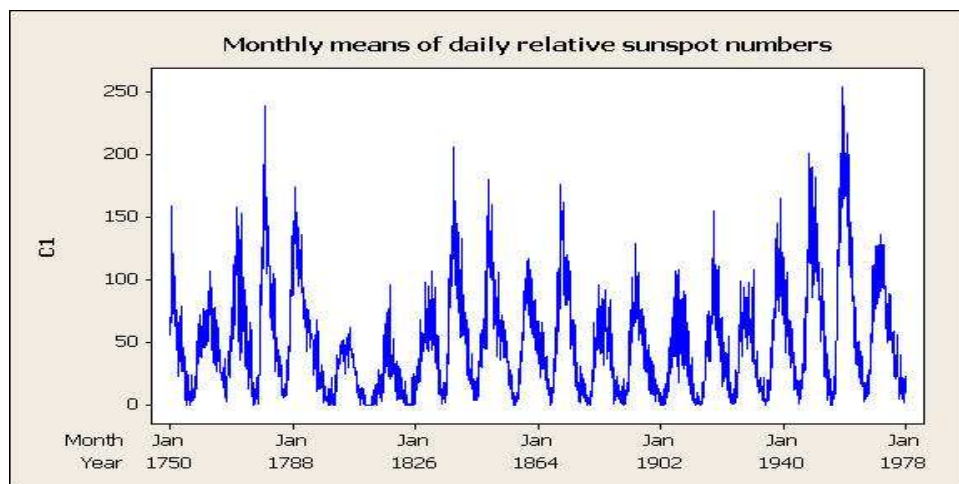
So what can we *say* about the sales figures for *Turner's Hangover Cure*?



Look at the time series plots shown below. How could you describe these?



Comments:

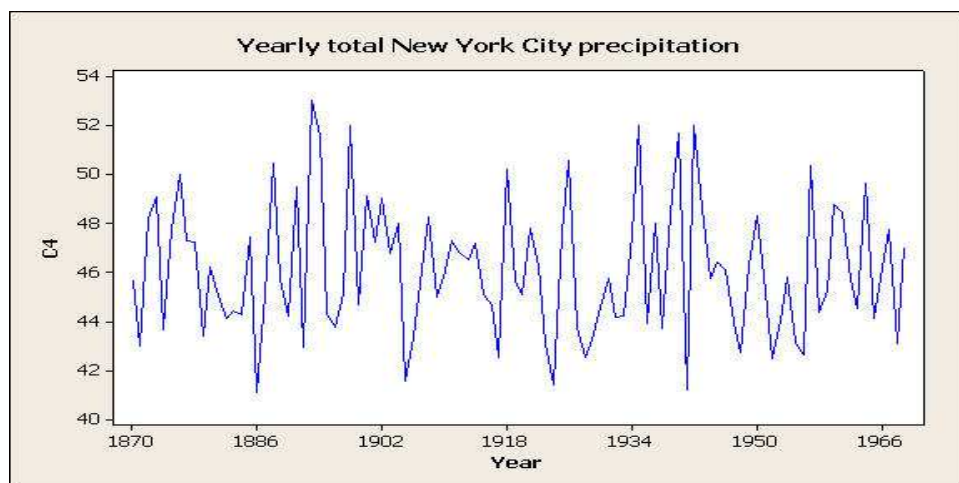


Comments:





Comments:



Comments:



2.3 Isolating the trend

2.3.1 MAS1403 review

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 *Turner's* sales figures, 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.

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

Similarly, the second moving average is

$$\frac{10 + 13 + 10}{3} = 11.$$

The rest of the moving averages can be calculated in this way, and should be entered into table 2.1 below.

	Moving averages		
	Jan–Apr	May–Aug	Sep–Dec
2006	*	10.33	11.00
2007			
2008			
2009			*

Table 2.1: Moving averages for *Turner's Hangover Cure* sales figures

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. You should be able to calculate such moving averages by hand; however, as with most of the material in this course, **Minitab** can do this for us, which is very useful for larger datasets!

In **Minitab**, you would click on **Stat–Time Series–Moving Average**; you would enter **C1** in the **Variable** box and enter the **MA length** as 3 (since we have a cycle length of 3). You should **Center the moving averages**; click on **Storage** and select **Moving Averages** (and then **OK**); select **Graphs** and choose the box that says **Plot smoothed vs. actual**. Doing so will store the moving averages you calculated in table 2.1 in the next available column in **Minitab** and you should also get the plot shown in Figure 2.3. Figure 2.2 is a **Minitab** screenshot illustrating the process described above.

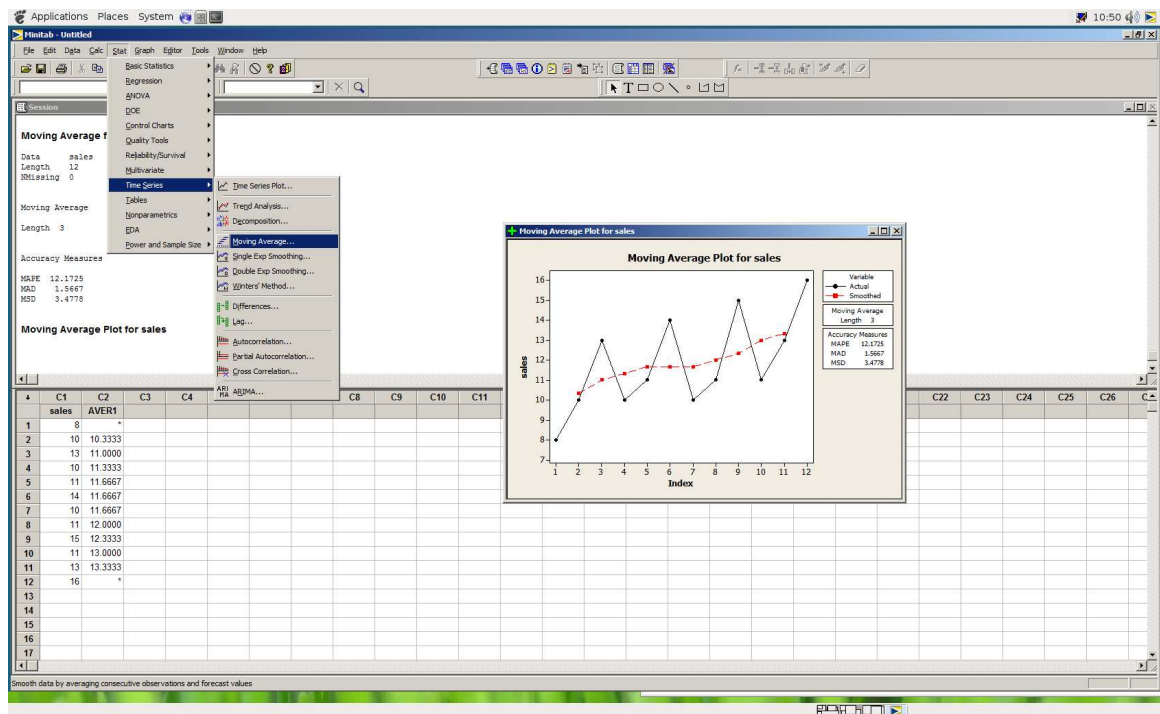


Figure 2.2: Minitab screenshot showing the moving average option

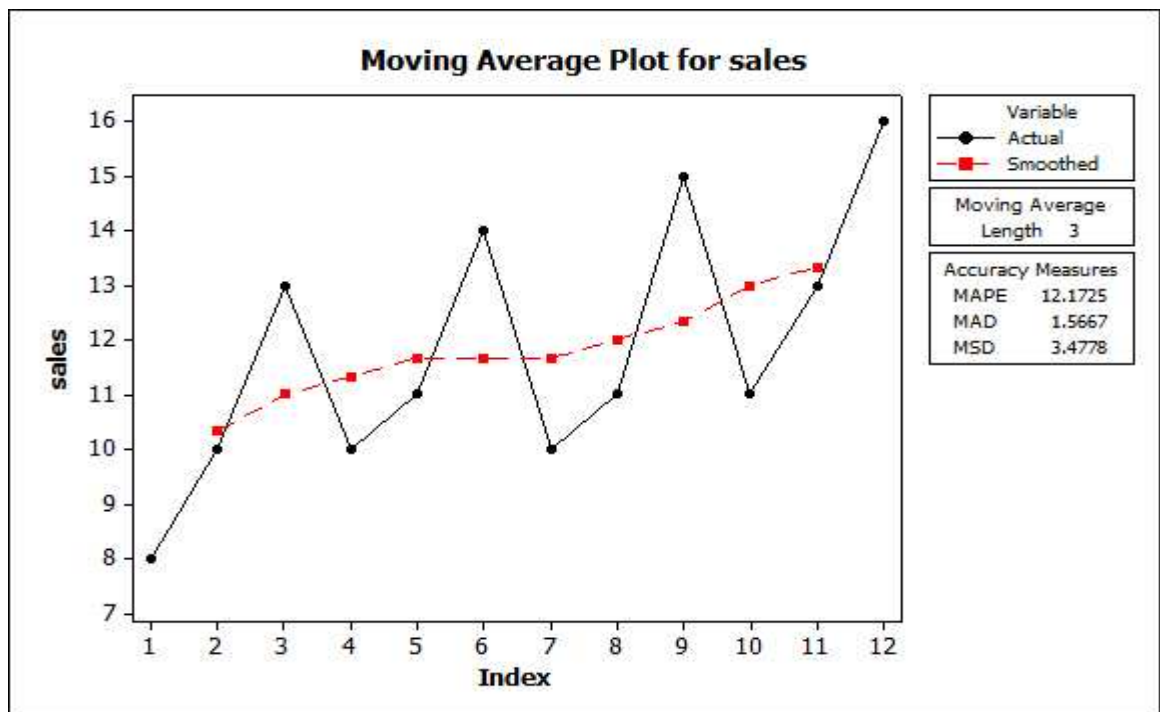


Figure 2.3: Time series plot with moving averages superimposed

2.3.2 Quarterly and monthly data

In MAS1403 we considered the calculation of moving averages when the cycle length was a convenient number, i.e. an odd number. For instance, in the last example, the cycle length was 3; taking the average over every consecutive triple is easy to do, and *centres* the moving average around the middle observation.

Let $Y_1, Y_2, \dots, Y_n$ be our time series of interest, and so $y_t, t = 1, \dots, n$ are the *observed* values at time t . Then, for a cycle of length 3, the *three-point moving average* at time t is given by

$$y_t^* = \frac{y_{t-1} + y_t + y_{t+1}}{3},$$

and this is *centred* around time point t . What if we have quarterly data?

Moving averages for quarterly data

Suppose we have 3-monthly (quarterly) data, so a cycle consists of 4 observations, e.g.

2007	1
	2
	3
	4
2008	1
	2
	3
	4

Now simple averaging over a cycle around an observation cannot be used as this would span four quarters and would not be centred on an integer value of t .

For example, if we take $t = (2007, 4)$ and calculate the mean of the quarters 2, 3 and 4 of 2007 and the first quarter of 2008, this gives us *not* an estimate for the trend at time $t = (2007, 4)$, but it gives us an estimate for the trend somewhere between $t = (2007, 3)$ and $t = (2007, 4)$. A simple average over 5 quarters cannot be used, as this would give twice as much weight to the quarter appearing at both ends. Therefore, we use the following formula as an estimate for the moving average at time t :

$$y_t^* = \frac{y_{t-2} + 2(y_{t-1} + y_t + y_{t+1}) + y_{t+2}}{8}.$$

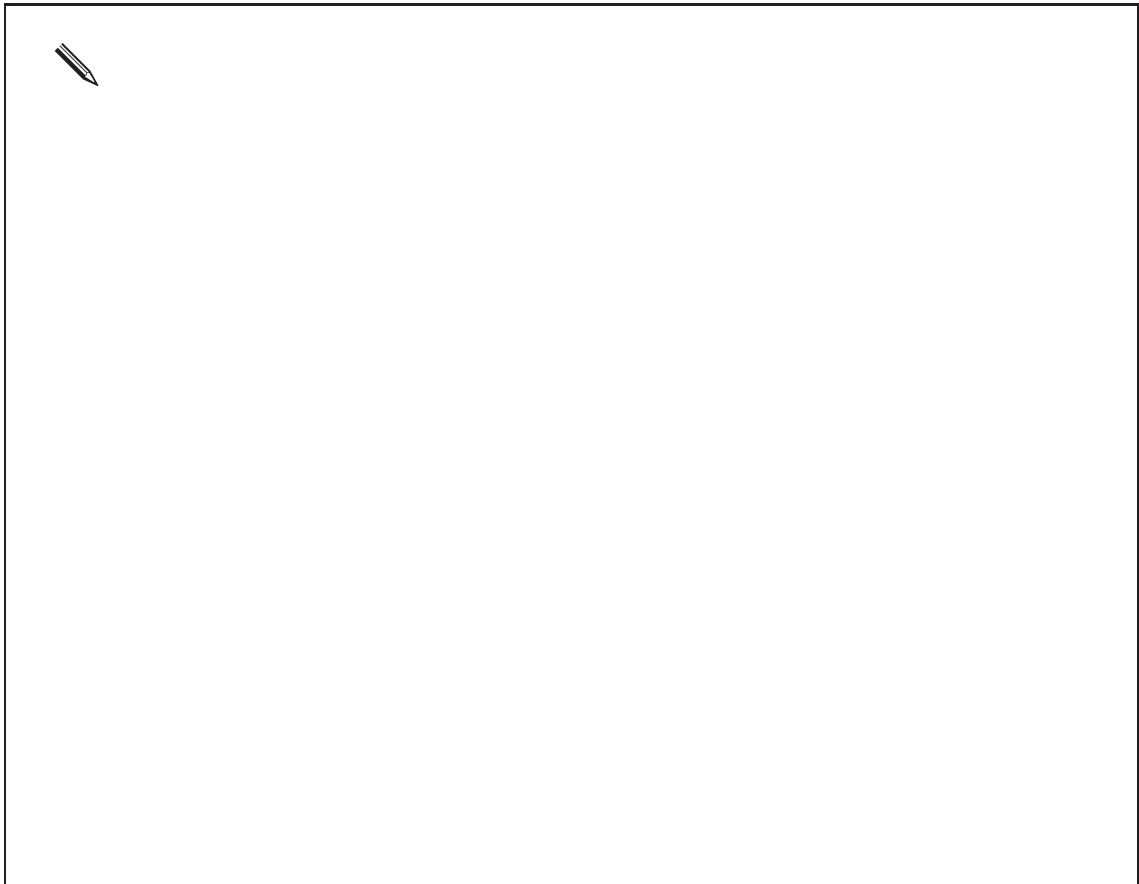
Example

Table 2.2 shows the quarterly passenger figures (rounded, in Millions) for British Airways between 2006–2008 (inclusive). Calculate the series of *quarterly moving averages* and enter your results in the correct cells of table 2.3. The first one is done for you.

	Q1 (Jan–Mar)	Q2 (Apr–Jun)	Q3 (Jul–Sep)	Q4 (Oct–Dec)
2006	12	6	8	10
2007	14	7	8	13
2008	16	9	10	13

Table 2.2: British Airways passenger figures, 2006–2008

$$\begin{aligned}
 y_3^* &= \frac{12 + 2(6 + 8 + 10) + 14}{8} \\
 &= \frac{12 + 48 + 14}{8} \\
 &= 9.25
 \end{aligned}$$



	Q1 (Jan–Mar)	Q2 (Apr–Jun)	Q3 (Jul–Sep)	Q4 (Oct–Dec)
2006	*	*	9.25	
2007				
2008			*	*

Table 2.3: British Airways quarterly moving averages, 2006–2008

As before, we can get *Minitab* to do this for us, as well as produce a time series plot with the moving averages superimposed; such a plot is shown in Figure 2.4.

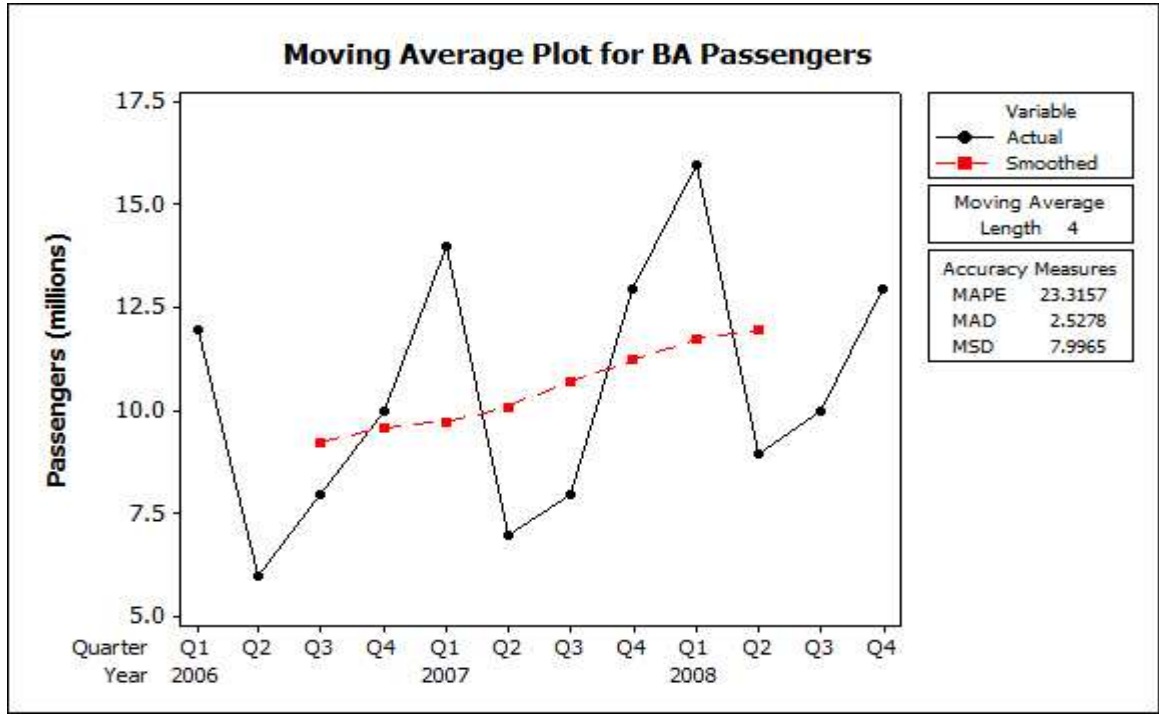


Figure 2.4: Time series plot with moving averages superimposed for the BA passenger data

Moving averages for monthly data

By similar reasoning, i.e. to ensure our moving averages are centred around an integer time value and to avoid undue weight being given to a particular “season”, we use the following formula to obtain moving averages for monthly data:

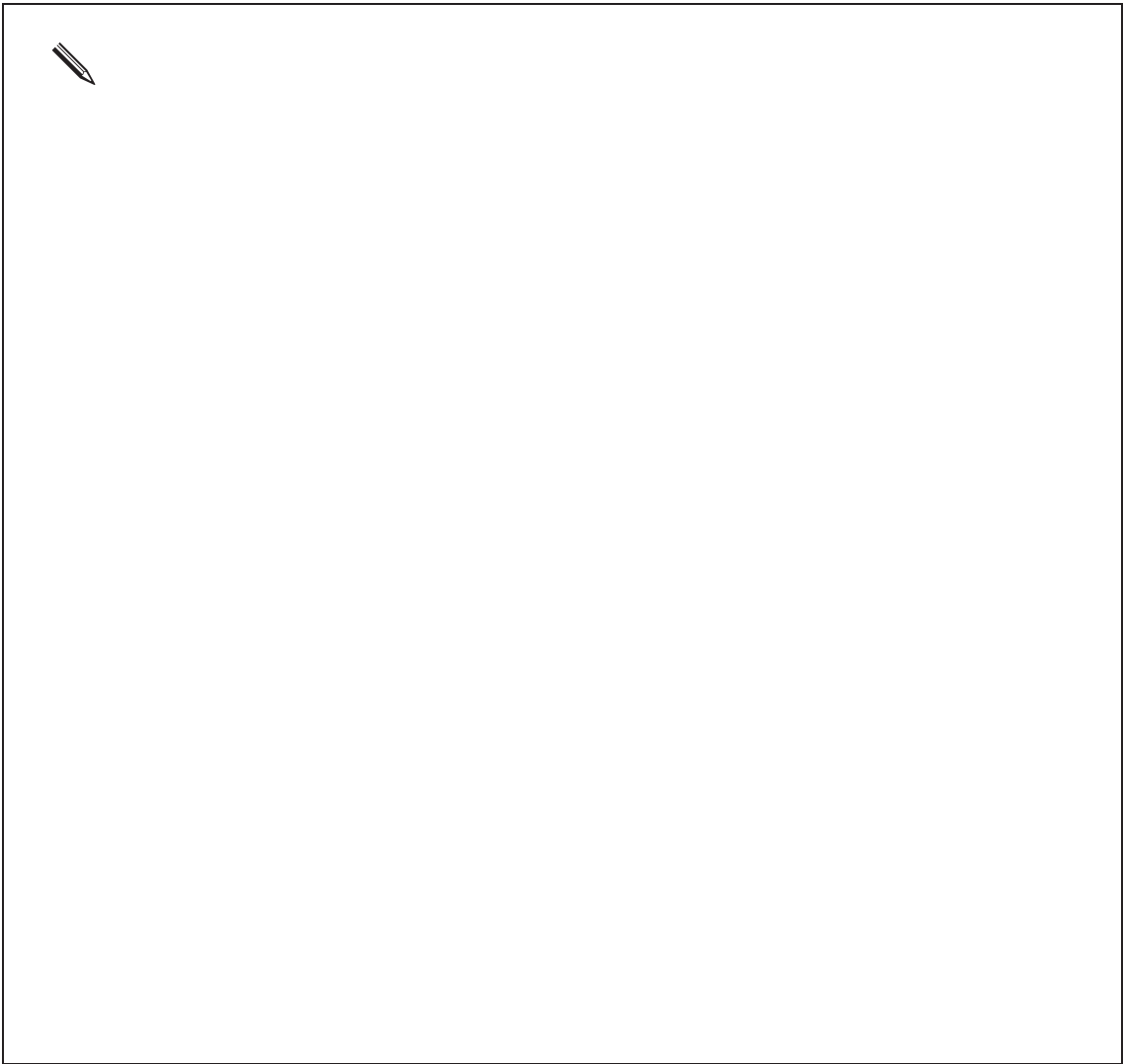
$$y_t^* = \frac{y_{t-6} + 2(y_{t-5} + \dots + y_{t-1} + y_t + y_{t+1} + \dots + y_{t+5}) + y_{t+6}}{24}.$$

Table 2.4 shows the number of British visitors, in thousands per month, to the Spanish island of Menorca (kindly provided by the Spanish Tourist Board). Obtain the series of *monthly moving averages* and enter your results in table 2.5; the first one has been done for you (in fact, to save time, I’ve left space for some of your calculations but have entered the answers into Table 2.5 for you). Again, this can be done in *Minitab*; Figure 2.5 shows a time series plot for these data, with the calculated moving averages superimposed.

	J	F	M	A	M	J	J	A	S	O	N	D
2003	5	3	4	8	10	12	14	20	19	14	6	3
2004	7	4	8	10	15	16	17	21	20	16	8	4
2005	8	5	8	10	16	18	20	22	21	17	9	5

Table 2.4: British tourists to Menorca, 2003–2005

$$\begin{aligned} y_7^* &= \frac{5 + 2(3 + 4 + 8 + 10 + 12 + 14 + 20 + 19 + 14 + 6 + 3) + 7}{24} \\ &= \frac{238}{24} \\ &= 9.917. \end{aligned}$$



	J	F	M	A	M	J	J	A	S	O	N	D
2003	*	*	*	*	*	*	9.92	10.04	10.25	10.50	10.79	11.17
2004	11.46	11.63	11.71	11.83	12.00	12.13	12.21	12.29	12.33	12.33	12.38	12.50
2005	12.71	12.88	12.96	13.04	13.13	13.21	*	*	*	*	*	*

Table 2.5: British tourists to Menorca, 2003–2005: moving averages

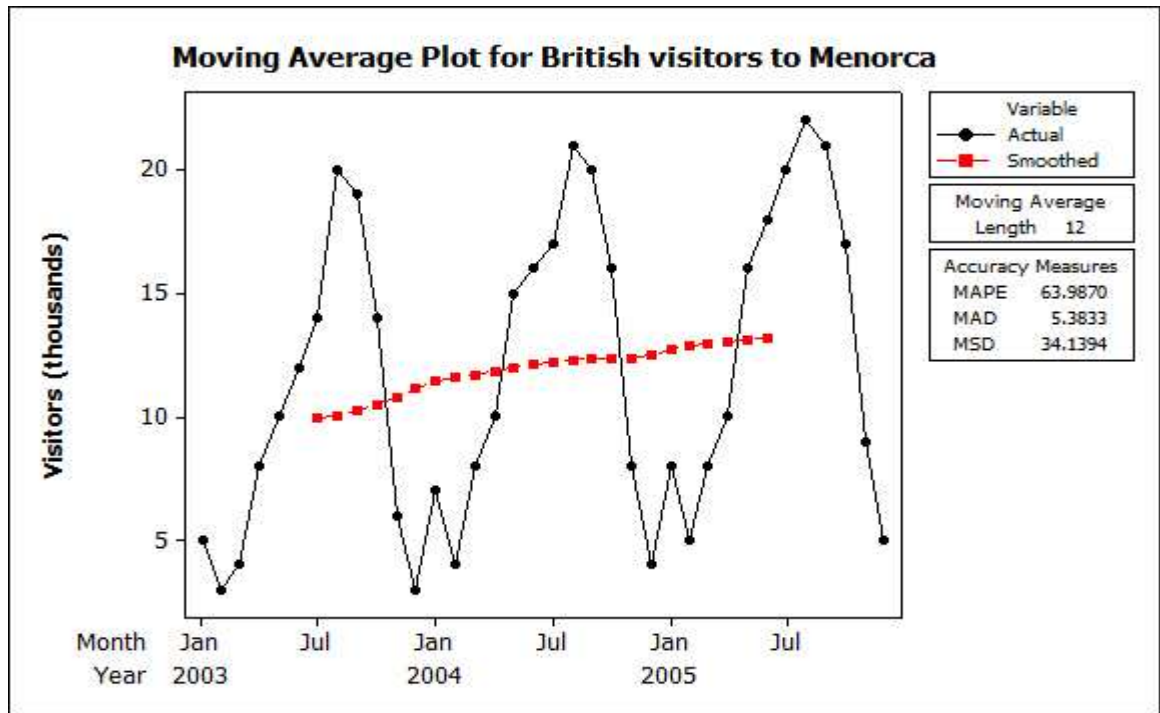


Figure 2.5: Time series plot with moving averages superimposed for the Menorca visitors data

2.3.3 Using simple linear regression for the trend

Look at the plots in Figures 2.3, 2.4 and 2.5. Notice that, once we've smoothed out the data by calculating moving averages, these moving averages seem to follow (roughly) a straight line. From a forecasting point-of-view, this is great, since we can use some of the ideas from the last chapter in this course to *model* this straight line relationship! In fact, even if the moving averages *did not* follow a straight line, it might be possible to employ, for example, quadratic regression here.

Example: BA passengers data

Look again at the data in Table 2.2 and the time series plot in Figure 2.4, showing the changes in quarterly passenger numbers for British Airways between 2006 and 2008. How could we use this information to predict passenger numbers in the first quarter of 2009? Or the second quarter of 2010? One approach is to fit a regression line to the series of moving averages and then extend this line to predict future moving averages. Since the moving averages in Figure 2.4 seem to show a reasonably linear pattern, we could use simple linear regression here, where the predictor variable is time and the response variable is the series of moving averages. Putting the moving averages calculated on page 53 (and shown in Table 2.3), and the corresponding time indices, in a table, gives:

t	y^*	t^2	ty^*
3	9.25	9	27.75
4	9.625	16	38.5
5	9.75	25	48.75
6	10.125	36	60.75
7	10.75	49	75.25
8	11.25	64	90
9	11.75	81	105.75
10	12	100	120
52	84.5	380	566.75

Why have we drawn a table up like this? Well, we are simply replacing the simple linear regression equation from Section 1.2.2 (page 10), with

$$Y^* = \beta_0 + \beta_1 T + \epsilon,$$

where Y^* represents our moving averages and T represents time. Thus, we now have

$$\hat{\beta}_1 = \frac{S_{TY^*}}{S_{TT}} \quad \text{and}$$

$$\hat{\beta}_0 = \bar{y}^* - \hat{\beta}_1 \bar{t},$$

where

$$S_{TY^*} = \sum_{i=3}^{10} t_i y_i^* - n \bar{t} \bar{y}^* \quad \text{and}$$

$$S_{TT} = \sum_{i=3}^{10} t_i^2 - n \bar{t}^2.$$

Using the sums from the above table gives:

$$S_{TY^*} = 566.75 - 8 \times \frac{52}{8} \times \frac{84.5}{8}$$

$$= 17.5,$$

$$S_{TT} = 380 - 8 \times \left(\frac{52}{8} \right)^2$$

$$= 42.$$

Thus, we have

$$\begin{aligned}\hat{\beta}_1 &= \frac{17.5}{42} \\ &= 0.417 \quad \text{and} \\ \hat{\beta}_0 &= \frac{84.5}{8} - 0.417 \times \frac{52}{8} \\ &= 7.852.\end{aligned}$$

So the regression equation is given by

$$Y^* = 7.852 + 0.417T + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$. Of course, you could also find this regression equation using **Minitab**; with the original data in column **C1** and the moving averages in column **C2** (I tell you how to obtain moving averages in **Minitab** on page 50 of these notes), you should also set up a time index column from 1 up to 12 (perhaps in column **C3**). Then the options **Stat-Regression-Regression** can be used, specifying the moving averages (column **C2**) as the **Response** variable and the time index column (column **C3**) as the **Predictor**. If you click on **Storage** and check the box that says **Fits**, the fitted values from the linear regression will also be stored in the **Minitab** worksheet. This is illustrated in the screenshot of Figure 2.6. With the fitted values stored, a time series plot with the moving averages *and* regression line superimposed can now be produced. This is shown in Figure 2.7, and you will see how to do this for yourself in Practical 3. Shown below is the **Minitab** output for the regression analysis, confirming our calculations above: notice that from **Minitab** we also have an estimate of σ , the standard deviation of the residuals, and so our fully specified model for the trend in passenger numbers is

$$Y^* = 7.852 + 0.417T + \epsilon, \quad \epsilon \sim N(0, 0.156^2).$$

Regression Analysis: AVER1 versus C3

The regression equation is

AVER1 = 7.85 + 0.417 C3

8 cases used, 4 cases contain missing values

Predictor	Coef	SE Coef	T	P
Constant	7.8542	0.1658	47.37	0.000
C3	0.41667	0.02406	17.32	0.000

S = 0.155902 R-Sq = 98.0% R-Sq(adj) = 97.7%

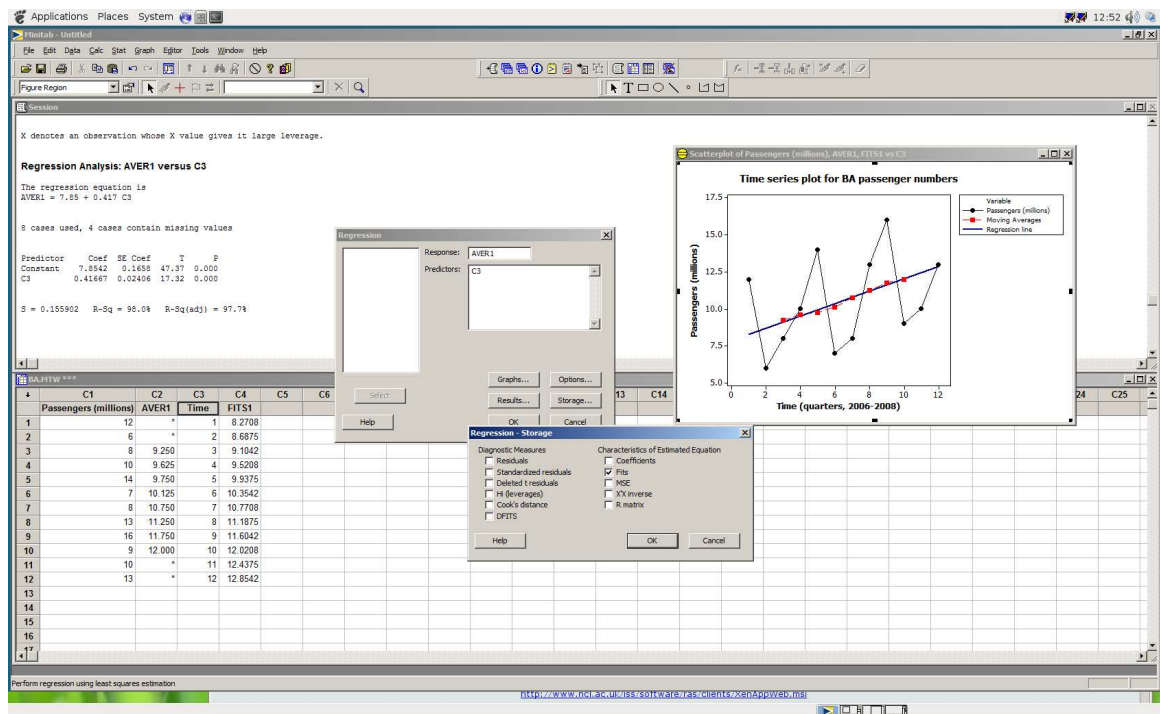


Figure 2.6: Minitab screenshot showing how to fit a simple linear regression to the British Airways moving averages

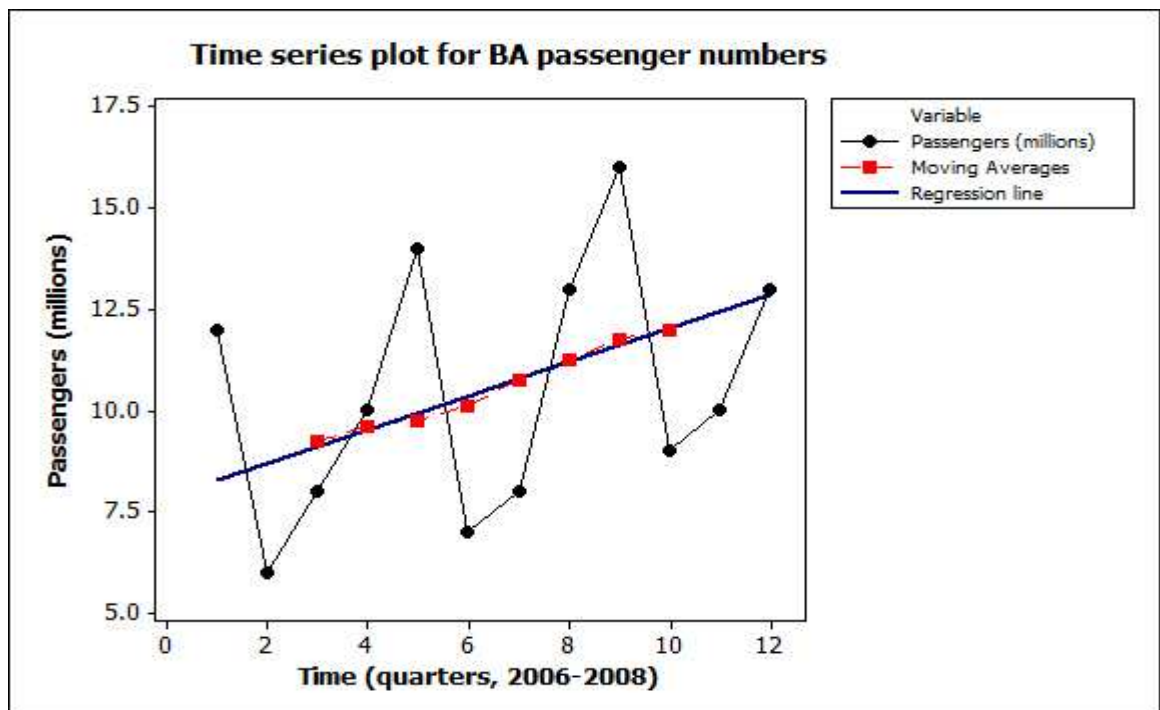


Figure 2.7: Time series plot with moving averages *and* regression line superimposed for the BA passengers data

Questions

Use the estimated regression equation to forecast total BA passenger numbers in Jan–March 2009.



Why might the global economic situation in 2009–2010 invalidate this forecast?



What else have we not accounted for here?



2.4 Isolating the seasonal effects

In the last section we examined how to isolate trend in our time series data. We did this by

- “smoothing out” the data by finding moving averages (for cycle lengths of 3, 4 and 12; a cycle length of 4 could represent quarterly data and a cycle length of 12 could represent monthly data);
- fitting a regression line to the series of moving averages.

However, as we noted in the last example, any forecasts we make based on the regression line alone do not take into account the *seasonal cycles* around that line. We will now review the methods used in MAS1403 to identify *seasonal effects*, but will also see this in action in Minitab.

2.4.1 MAS1403 review

In MAS1403 we used several steps to obtain our seasonal effects:

1. Find the *seasonal deviations* (original data minus moving averages or, in our new notation, $y_t - y_t^*$, $t = 1, \dots, n$);
2. Calculate the *seasonal means*, which are just the mean of the seasonal deviations for each season;
3. Calculate the *seasonal effects*, which are the seasonal means minus the mean of all the seasonal deviations;
4. Obtain the *adjusted seasonal effects* by adjusting the seasonal effects found in step (4) so that they sum to give zero (only do this if they don’t sum to zero in the first place).

Example: BA passenger data

Recall from table 2.2 and 2.3 the quarterly British Airways passenger figures (in millions for 2006–2008), and the corresponding moving averages, respectively:

	Q1 (Jan–Mar)	Q2 (Apr–Jun)	Q3 (Jul–Sep)	Q4 (Oct–Dec)
2006	12	6	8	10
2007	14	7	8	13
2008	16	9	10	13
	Q1 (Jan–Mar)	Q2 (Apr–Jun)	Q3 (Jul–Sep)	Q4 (Oct–Dec)
2006	*	*	9.25	9.625
2007	9.75	10.125	10.75	11.25
2008	11.75	12	*	*

Step 1: Seasonal deviations

	Q1 (Jan–Mar)	Q2 (Apr–Jun)	Q3 (Jul–Sep)	Q4 (Oct–Dec)
2006	*	*		
2007				
2008			*	*
Seasonal means				

Table 2.6: Seasonal deviations for Brisith Airways data

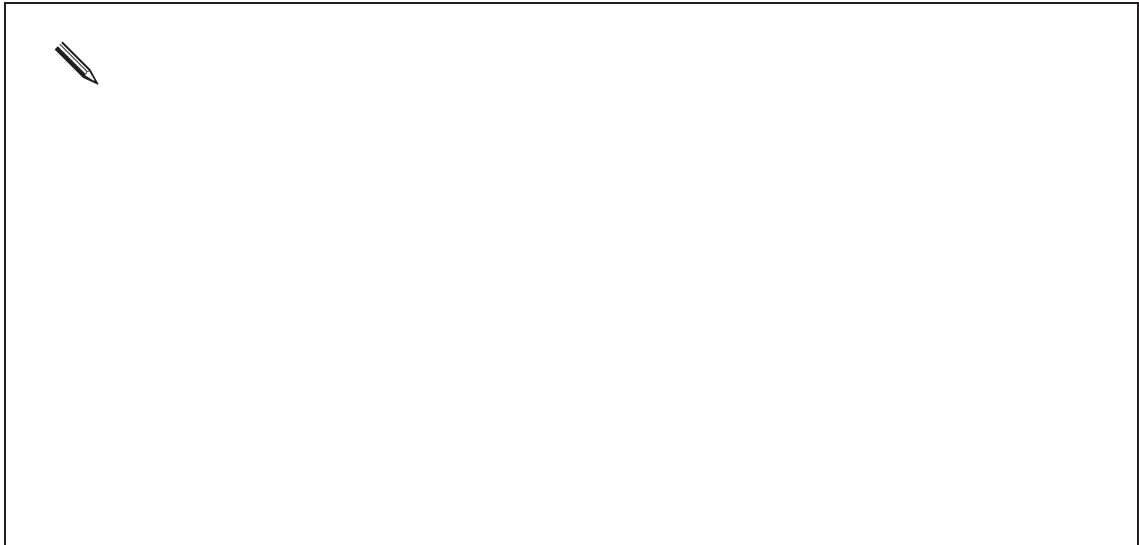
Step 2: Seasonal means

Now calculate the seasonal means, and enter them in table 2.6 above. Use the space below to show your working, if you need to.



Step 3: Seasonal effects



Step 4: Adjusted seasonal effects**2.4.2 Seasonal effects in Minitab**

As always, we can find the seasonal effects for our time series data using **Minitab**, which is just as well – imagine how long this process would take if you had monthly data, or even daily data, collected over many years!? With the entire time series in a single column of a **Minitab** worksheet (say column **C1**), we would click on **Stat–Time Series–Decomposition**. We would enter the **Variable** as **C1** (if that’s where our data are), enter the **Seasonal length** as 4 (as we have quarterly data here); select **Trend plus seasonal** as that’s what we have in this example; select **Additive** for the **Model type**; and then finally, before clicking on **OK**, we can get **Minitab** to store the results in the next available column of the worksheet by clicking on **Storage** and selecting **Seasonals**. This is illustrated in the **Minitab** screenshot shown in figure 2.8, and you will be trying this for yourself in next week’s practical session. Notice the values **Minitab** has stored in column **C2** here are very close the values we calculated by hand; our calculations are obviously prone to rounding error.

2.4.3 Using the seasonal effects to make forecasts

Recall the question at the top of page 60 in these notes:

Use the estimated regression equation to forecast total BA passenger numbers in Jan–March 2009.

We can now do this more realistically by adjusting our forecast obtained via the regression equation for the seasonal effect for Jan–March. Recall that the regression equation for the moving averages was found to be:

$$Y^* = 7.852 + 0.417T + \epsilon.$$

January–March 2009 would be time–point 13, and so using this regression equation gave us a forecast of

$$\begin{aligned} Y^* &= 7.852 + 0.417 \times 13 \\ &= 13.273, \end{aligned}$$

or just over 13 million passengers. However, you’ll notice from figure 2.7 that the first quarter of each year always seems to record higher than average passenger figures; so we now adjust this initial forecast by the seasonal effect for January–March, which was found to be +4.1875, giving a *full forecast* of

$$13.273 + 4.1875 = 17.4605,$$

or just under 17.5 million passengers. Note that this has still not taken into account the global financial situation of late!

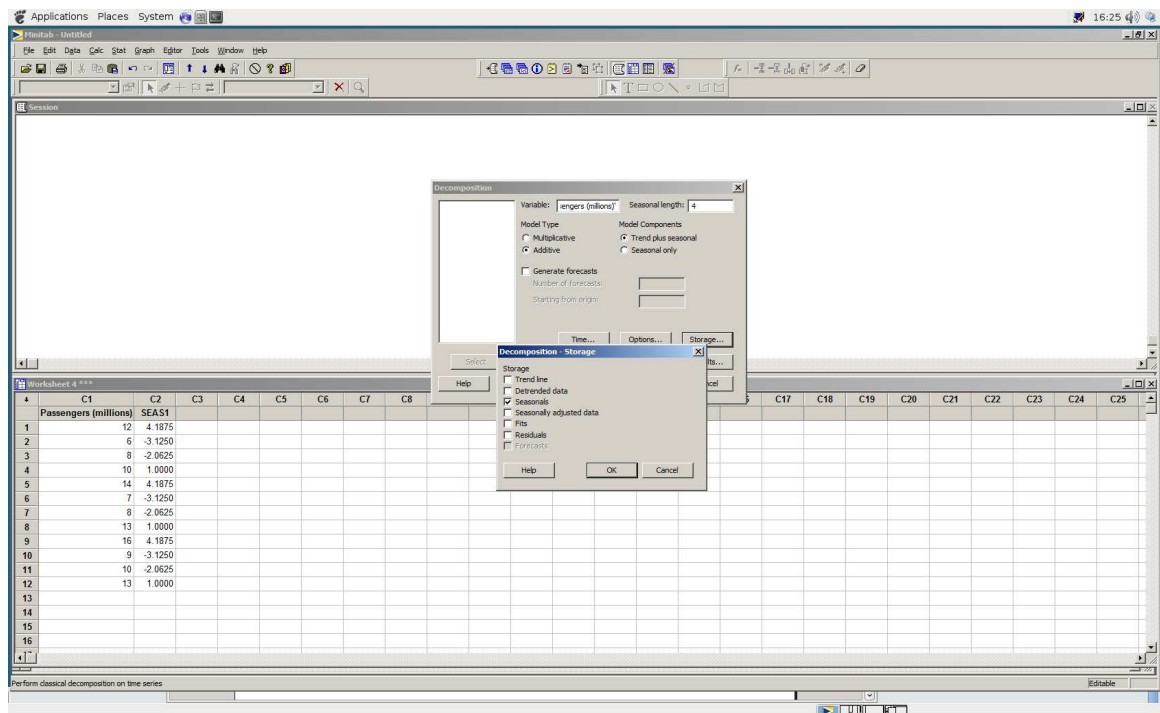


Figure 2.8: Minitab screenshot showing how to obtain seasonal effects in Minitab

2.5 Obtaining the residual series

In the next section, we will consider some special *probability models* for time series data. These models assume that our data are *stationary*, i.e. have no trend or seasonality. Most of the time, our time series data exhibit either trend, or seasonality, or *both* – in fact, this is what makes time series data so interesting – and so these probability models are not immediately useable. However, if we can estimate the trend and seasonal components of our data – and we have shown how to do this in the previous two sections – we can attempt to *make* our time series stationary by *de-trending* and *de-seasonalising*.

Example: British Airways passenger data

Table 2.7 shows the original BA passenger data in the first column; the *fitted* values from the simple linear regression model for the trend in the second column (you will see how to obtain these values in *Minitab*, though you should be able to see how to get these by hand), and our calculated seasonal effects in the third column. The fourth column shows our *de-trended*, *de-seasonalised* data, obtained by subtracting the trend and the seasonal effects from the original data. The resulting series is often called the *residual series*, i.e. this is what's left over when we've taken out the trend and seasonality, and it is series such as this that we can model using *time series models* (see next Section). A plot of this residual series is shown in figure 2.9.

BA passenger data	Trend (fitted values)	Seasonal effects	Residual series
12	8.269	+4.1875	-0.4565
6	8.686	-3.125	0.439
8	9.103	-2.0625	0.9595
10	9.520	+1	-0.52
14	9.937	+4.1875	-0.1245
7	10.354	-3.125	-0.229
8	10.771	-2.0625	-0.7085
13	11.188	+1	0.812
16	11.605	+4.1875	0.2075
9	12.022	-3.125	0.103
10	12.439	-2.0625	-0.3765
13	12.856	+1	-0.856

Table 2.7: Obtaining the *residual series* for the BA passenger data

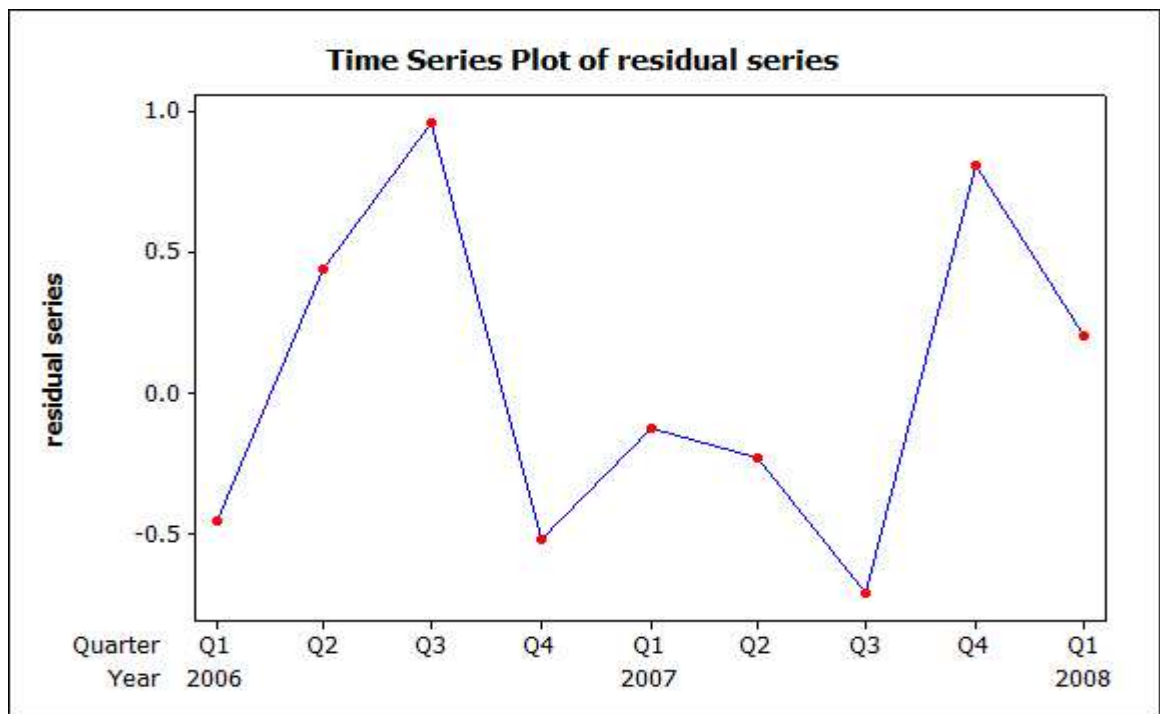


Figure 2.9: Time series plot of the residuals series for the BA passenger data