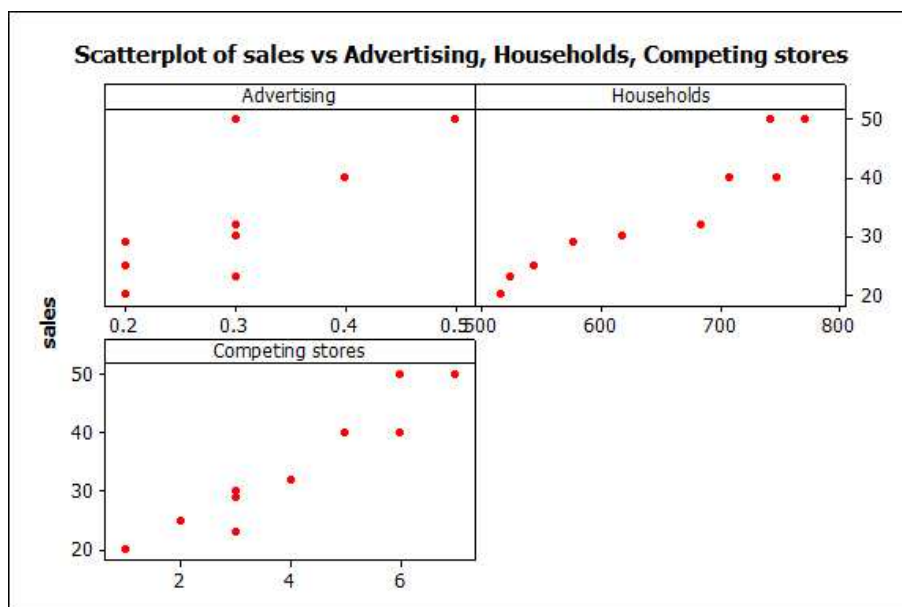


Solutions to selected practical questions

Practical 1 – Question 2

- (a) The plots below show the relationship between sales and each of the three candidate predictor variables:



- There seems to be a weak/moderate, positive, linear relationship between amount the spent on advertising and sales;
 - the relationship between number of households and sales seems a little stronger – and again is positive – but possibly non-linear;
 - there is a moderate/strong, positive, linear relationship between the number of competing stores and sales.
- (b) The table below shows the sample correlation coefficients between sales and each of the three potential predictor variables (along with their p -values):

	Advertising	Households	Competing stores
Sales	0.754 (0.012)	0.944 (0.000)	0.953 (0.000)

The number of competing stores seems to show the strongest linear association with sales – the correlation coefficient between sales and competing stores is closer to 1 than that for the other two pairings (though it is worthwhile to note that all *three* potential predictors seems to be significantly correlated with sales, as all three p -values are less than 0.05).

(c) From Minitab we get:

Regression Analysis: sales versus Advertising, Households, Competing stores

The regression equation is

sales = - 0.8 - 19.8 Advertising + 0.0363 Households + 4.37 Competing stores

Predictor	Coef	SE Coef	T	P
Constant	-0.80	16.18	-0.05	0.962
Advertising	-19.77	22.85	-0.86	0.420
Households	0.03634	0.03261	1.11	0.308
Competing stores	4.374	2.187	2.00	0.092

S = 3.25887 R-Sq = 93.8% R-Sq(adj) = 90.7%

giving the “full model” as

$$Y = -0.8 - 19.77X_1 + 0.03634X_2 + 4.374X_3 + \epsilon,$$

where $\epsilon \sim N(0, 3.25887^2)$.

(d) We now exclude X_1 (advertising), as this has the largest p -value exceeding 0.05. Doing so, gives:

Regression Analysis: sales versus Households, Competing stores

The regression equation is

sales = - 7.6 + 0.0456 Households + 3.06 Competing stores

Predictor	Coef	SE Coef	T	P
Constant	-7.65	13.86	-0.55	0.598
Households	0.04565	0.03023	1.51	0.175
Competing stores	3.058	1.543	1.98	0.088

S = 3.19970 R-Sq = 93.0% R-Sq(adj) = 91.0%

(e) From the output in part (d), we see that both the number of households and the number of competing stores appear to be insignificant in the model, since both have p -values greater than 0.05 (though competing stores is quite close the the cut-off). Since the number of households is the *least* significant (as it has the largest p -value), we exclude it and re-fit, giving the output shown overleaf:

Regression Analysis: sales versus Competing stores

The regression equation is

$$\text{sales} = 13.0 + 5.24 \text{ Competing stores}$$

Predictor	Coef	SE Coef	T	P
Constant	12.959	2.603	4.98	0.001
Competing stores	5.2353	0.5910	8.86	0.000

$$S = 3.44633 \quad R\text{-Sq} = 90.7\% \quad R\text{-Sq}(\text{adj}) = 89.6\%$$

- (f) (i) The estimated regression equation for the output in part (e) is

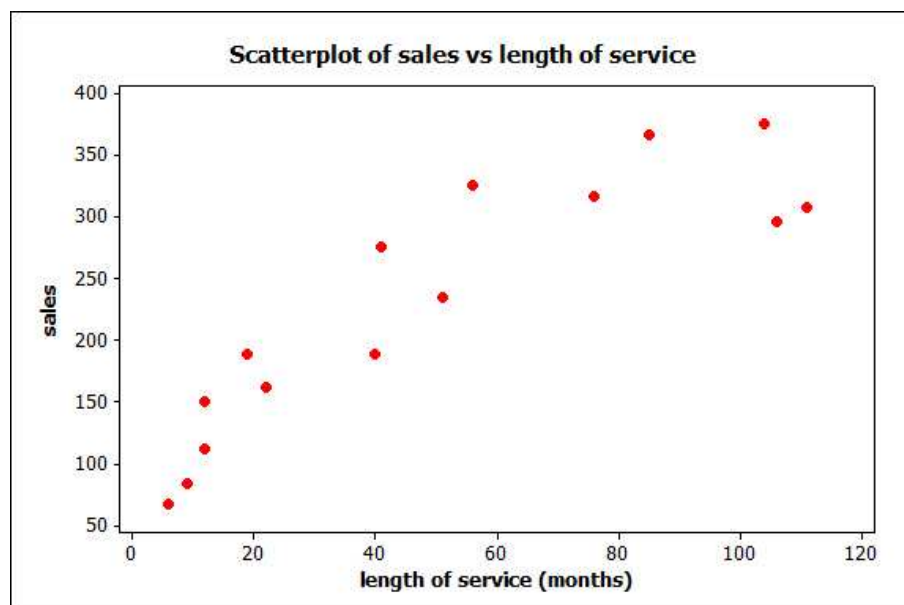
$$Y = 12.959 + 5.2353X_3 + \epsilon,$$

where $\epsilon \sim N(0, 3.44633^2)$.

- (ii) The p -value for X_3 (competing stores) is now 0.000 (to 3 d.p.). This means that we would now *reject* the null hypothesis $H_0 : \beta_3 = 0$, suggesting that X_3 (number of competing stores) is now an important predictor of Y (sales).
- (iii) We now have $R^2 = 90.7\%$, which means that just under 91% of the variability in Y is explained by X_3 . Notice that this hasn't reduced that much given that we've removed two predictor variables (from 93.8% when all three predictors were included).

Practical 2 – Question 2

- (a) The graph below shows that there is a fairly strong, positive *non-linear* relationship between length of service and sales.



- (b) Performing a quadratic regression in Minitab gives the following output:

Regression Analysis: sales versus service, servicesquared

The regression equation is

sales = 45.3 + 6.34 service - 0.0345 servicesquared

Predictor	Coef	SE Coef	T	P
Constant	45.35	22.77	1.99	0.070
service	6.345	1.058	6.00	0.000
servicesquared	-0.034486	0.008948	-3.85	0.002

S = 34.4528 R-Sq = 90.2% R-Sq(adj) = 88.6%

- (i) The quadratic regression equation is:

$$Y = 43.35 + 6.345X - 0.034486X^2 + \epsilon,$$

where $\epsilon \sim N(0, 34.4528^2)$.

- (ii) Clearly β_1 and β_2 are both significant in the model – we would reject both

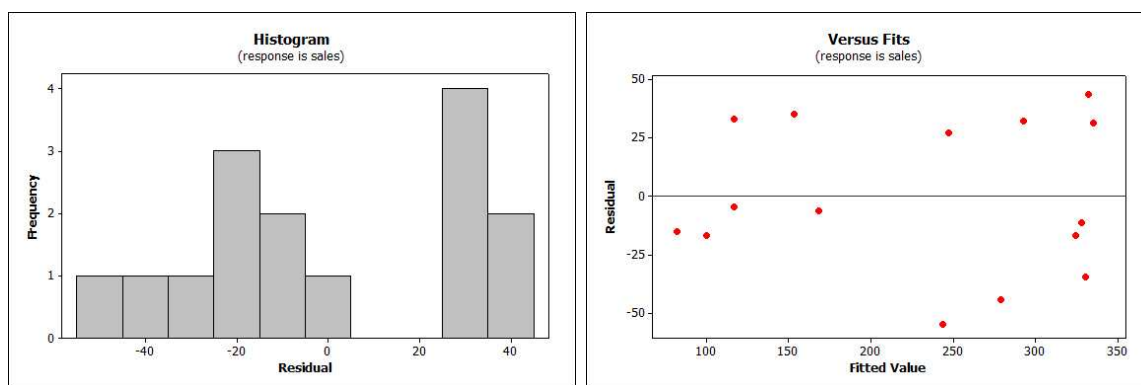
$$H_0 : \beta_1 = 0 \quad \text{and}$$

$$H_0 : \beta_2 = 0$$

since both have p -values less than 0.05 (0.000 and 0.002 respectively).

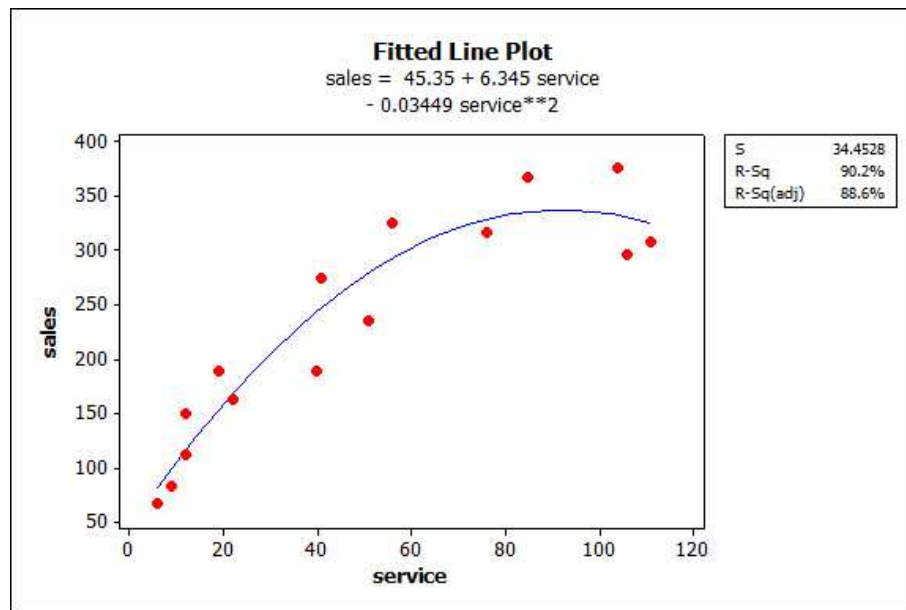
- (iii) The R^2 statistic is promising – just over 90% of the variability in Y (sales) is explained by both X and X^2 .

- (c) Below is a histogram of the residuals (left) and a plot of the residuals against the fitted values (right) for the quadratic regression model.



The histogram shows that the assumption of normality for the residuals might be questionable, since the sample distribution doesn't seem to follow the bell-shaped curve of the Normal distribution. However, this assumption is quite difficult to check from such a small sample. The plot to the right shows little association between the residuals and the fitted values, which is good!

- (d) Below is the scatterplot constructed in part (a), but now with the fitted quadratic regression curve overlaid.



- (e) We can expect Laura, who has worked at the *Holiday Hypermarket* for 18 months, to sell

$$\begin{aligned} Y &= 43.35 + 6.345 \times 18 - 0.034486 \times 18^2 \\ &= 146.3865, \end{aligned}$$

i.e. about 146 holidays.

Practical 2 – Question 3

- (a) **Logistic regression** – as the response variable (Y) is binary (can only take the values 0 or 1).
- (b) Using the output given on the practical sheet:
- (i) Our estimates of β_0 and β_1 are $\hat{\beta}_0 = -9.81397$ and $\hat{\beta}_1 = 0.727783$;
 - (ii) We would reject $H_0 : \beta_1 = 0$ since the p -value associated with β_1 is (only just!) less than 0.05 (it is 0.049!);
 - (iii) The logistic regression equation is given by

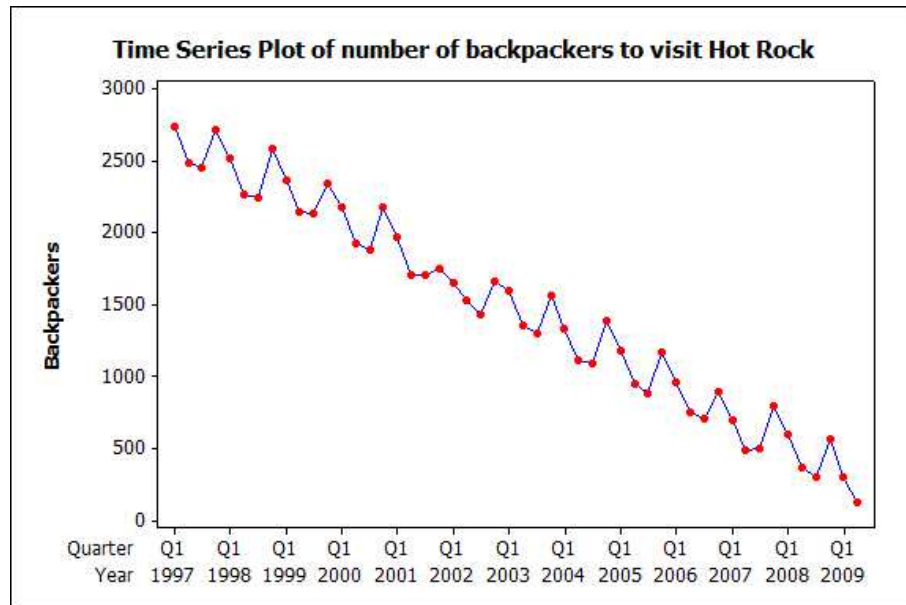
$$E[Y] = \frac{e^{-9.81397+0.727783X}}{1 + e^{-9.81397+0.727783X}}.$$

- (iv) If Alexandra is given a 12.5mg dose of this hangover cure, the probability that her hangover will be completely cured (according to our model) will be

$$\frac{e^{-9.81397+0.727783 \times 12.5}}{1 + e^{-9.81397+0.727783 \times 12.5}} = 0.328.$$

Practical 3

1. The graph below is a time series plot showing the number of backpackers visiting the *Hot Rock Hostel* in Australia.



2. From the graph in (1), we can see there is clear negative, linear trend, with seasonal variability. There is also something weird happening in 2001 – the cycle doesn't seem so defined here!
4. From Minitab we get:

Regression Analysis: MA versus Timepoint

The regression equation is

MA = 2736 - 49.4 Timepoint

46 cases used, 4 cases contain missing values

Predictor	Coef	SE Coef	T	P
Constant	2736.09	8.86	308.95	0.000
Timepoint	-49.4394	0.3080	-160.49	0.000

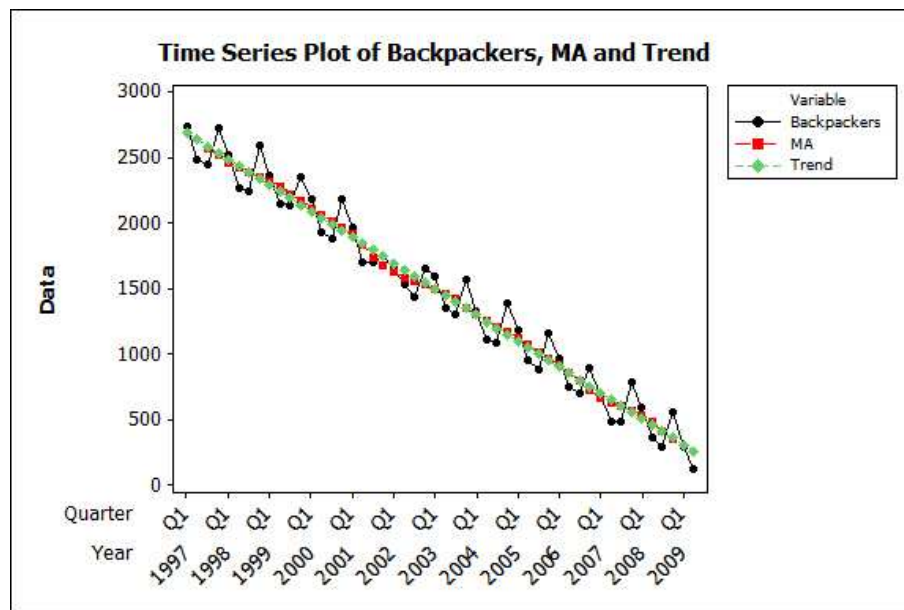
S = 27.7373 R-Sq = 99.8% R-Sq(adj) = 99.8%

giving

$$Y^* = 2736.09 - 49.4394T + \epsilon,$$

where $\epsilon \sim N(0, 27.7373^2)$. The slope term β_1 is significant as its p -value is less than 0.05 (0.000 to 3 d.p.).

5. The graph below shows the time series plot from (1), but now with the moving averages and fitted linear trend superimposed:



6. From Minitab the estimated seasonal effects are

$$\begin{aligned}\hat{s}_{Q1} &= +52.141 \\ \hat{s}_{Q2} &= -132.984 \\ \hat{s}_{Q3} &= -121.984 \\ \hat{s}_{Q4} &= +202.828\end{aligned}$$

7. (a) October–December 2009 is the fourth quarter in 2009, which would be time–point $t = 52$. This gives an estimated trend of

$$Y^* = 2736.09 - 49.4394 \times 52 = 165.2412.$$

Applying the seasonal effect for the fourth quarter gives a forecast of

$$165.2412 + 202.828 = 368.0692$$

i.e. about 368 backpackers.

- (b) April–June 2010 is the second quarter in 2010, which would be time–point $t = 54$. This gives an estimated trend of

$$Y^* = 2736.09 - 49.4394 \times 54 = 66.3624.$$

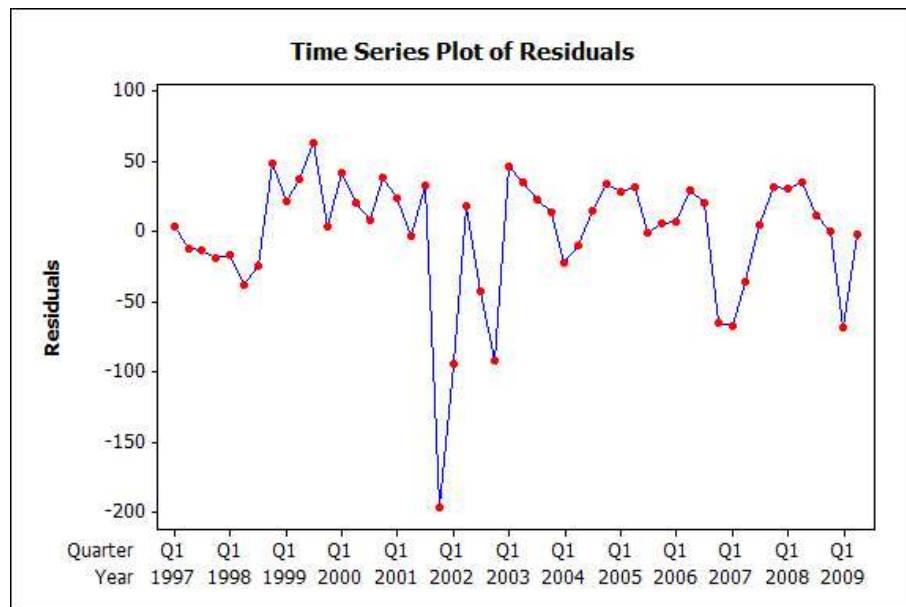
Applying the seasonal effect for the second quarter gives a forecast of

$$66.3624 - 132.984 = -66.6216$$

i.e. about –67 backpackers!

- (c) We have assumed our negative (linear) trend will continue forever – which in this application is nonsense, because at some point this will give a negative number of backpackers! And this has happened in the forecast in part (b)! You can't have a negative number of people!

9. The graph below is a time series plot of the resulting residual series. From this, we can see we have successfully removed any trend and seasonality – we are pretty much left with a *stationary* series. However, we can now more clearly see that there is an outlier in this series – in the fourth quarter of 2001.



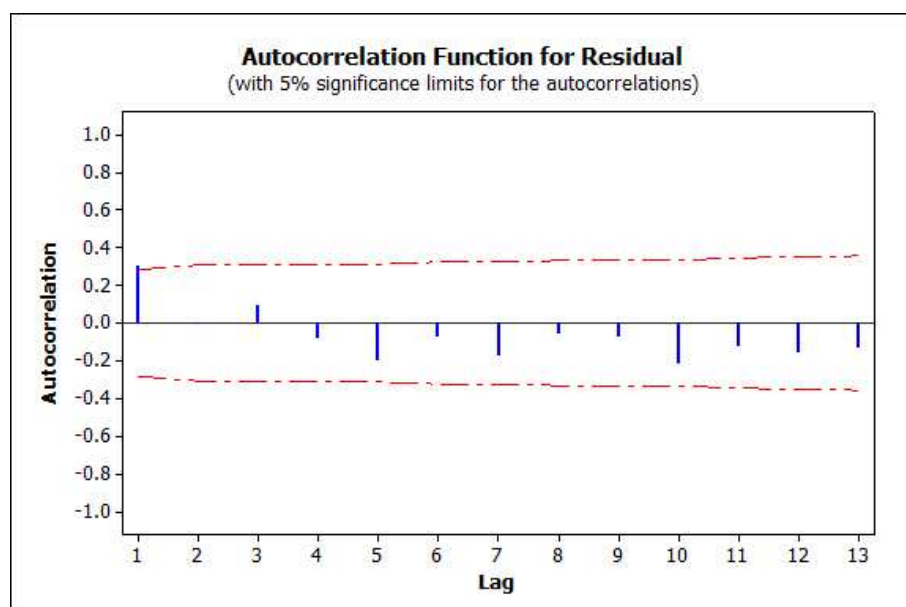
10. For the residual series, we get

$$r_1 = 0.298$$

$$p\text{-value} = 0.038$$

suggesting that although the lag 1 autocorrelation coefficient is small (0.298), it *is* significant, as the associated p -value is less than 0.05 (and so we would reject $H_0 : \rho_1 = 0$).

11. Below is the resulting correlogram for the series of residuals. The first bar matches with the value of r_1 found in question 10 (0.298). This is the only significant autocorrelation – the autocorrelations for all other lags are insignificant, as they do not exceed the significance bands in the plot.



12. From Minitab, we get:

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	0.2977	0.1378	2.16	0.036
Constant		-0.758	6.243	-0.12	0.904
Mean		-1.079	8.889		

This gives

$$\begin{aligned}Y_t &= -1.079 + 0.2977(Y_{t-1} - (-1.079)) + \epsilon \quad \text{i.e.} \\Y_t &= -1.079 + 0.2977(Y_{t-1} + 1.079) + \epsilon.\end{aligned}$$

The autoregressive parameter α is significant here, since the associated p -value is less than 0.05 (it's actually 0.036).

13. (a) Using the above fitted model, and the fact that $Y_{50} = -2.134$, we get

$$Y_{51} = -1.079 + 0.2977(-2.134 + 1.079) = -1.393073.$$

(b) The trend at time-point 51 is given by

$$Y^* = 2736.09 - 49.4394 \times 51 = 214.6806.$$

(c) Adding in the trend (214.608 from part (b)) and the seasonal effect for the third quarter (-121.984 from question 6) to the forecasted residual (-1.393073 from part (a)) gives a full forecast for the third quarter of 2009 of

$$-1.393073 + 214.6806 - 121.984 = 91.30353,$$

or about 91 backpackers.