Chapter 8: Working with Possible Relationships

Extra questions

1. What would you say are the characteristics of a good forecast?

2. Given the data below, determine the correlation coefficient and the regression of *y* on *x*. Comment on your results.

x	15	17	19	20	21	24
y	83	83	85	86	89	90

3. Given the data below, determine the correlation coefficient and the regression of *y* on *x*. Comment on your results.

x	15	17	19	20	21	24
y	83	90	86	85	89	83

4. A manager of a Fitness Club has collected the following data on typical weekly attendance and distance travelled:

Typical weekly attendance	Distance travelled (miles)
2	2
3	3
7	1
5	4
3	2
4	3
1	3

(a) Plot this data and comment on the observed relationship

(b) Determine the correlation coefficient and give an interpretation to its value

Discuss whether regression is worthwhile.

5. You have been given the following data on how y relates to x_1 and x_2.

y	x_1	x_2
53	17	5.4
87	17	5.8
238	19	7.9
457	25	14.5
456	27	14.5
356	22	12.8
442	27	15.9
600	35	17.3
235	21	16.0
156	19	12.4
375	19	12.4
376	21	12.6

Use multiple regression software to show the relationship and explain your results.

6. A company ranks applicants on the basis of their application form and their interview performance. The results for the last advertised post are given below:

Applicant	A	B	C	D	E	F	G
Application ranking	3	5	1	6	2	4	7
Interview ranking	4	6	2	5	1	3	7

Calculate Spearman's coefficient for this data and comment on the result.

Extra answers

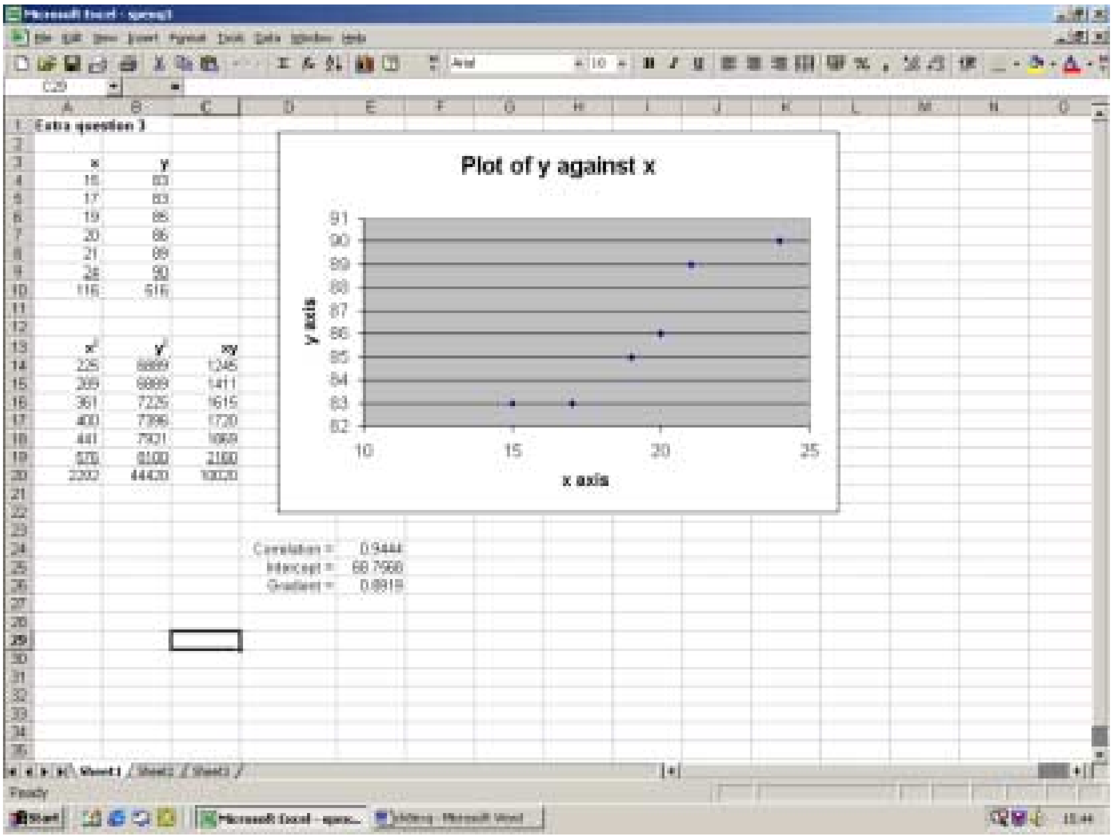
1. Essentially we would want a forecast to give us figures that are reasonably close to those likely to occur. The problem is that we never know at a particular point in time. We can, of course, look at the track record of any forecasting method and ask the question ‘how good has it been?’ We can look back over the data and examine the difference between what happened and what we predicted would happen.

We do talk about the error being the difference between the actual value and the forecasted value. We would want a selected forecasting method to have relatively low errors. It is also important to recognise that there are a number of forecasting methods and you will want the one most appropriate for your problem.

A characteristic of a good forecasting method is that the errors don’t go in any particular direction, referred to as bias. Clearly, we would not want to be making forecasts that were consistently too high or too low.

In all forecasting there is a balance between technical ‘know how’ and judgment. You we want a forecasting method that is scientifically sound but you will also want to use your knowledge of any problem situation. It might be that you know that poor weather over the next two weeks will affect short-term sales or recent changes to government policy will make a particular service more popular.

2. Summary information is given on the following spreadsheet:



Extra question 3

x	y	
15	83	
17	83	
19	85	
20	86	
21	89	
24	90	
116	516	
x^2	y^2	xy
225	6889	1245
289	6889	1411
361	7225	1615
400	7396	1720
441	7921	1869
576	8100	2160
2292	44420	10020

Correlation =	0.9444
Intercept =	68.7966
Gradient =	0.8819

Substituting into the formula given in the text:

$$r = \frac{6\times 10020 - 116\times 516}{\sqrt{\left(6\times 2292 - (116)^2\right)\left(6\times 44420 - (516)^2\right)}}$$

$$r = \frac{264}{\sqrt{296\times 264}} = 0.9444$$

The value of r = 0.9444 confirms a strong positive relationship

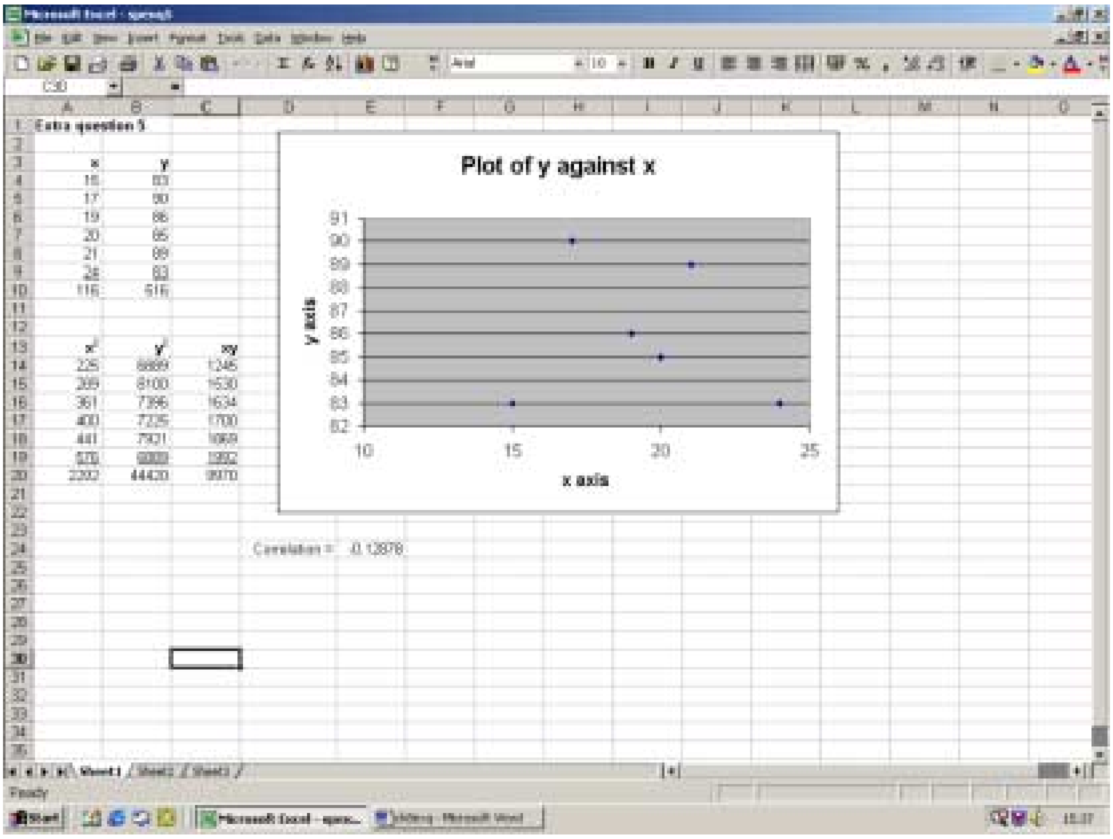
$$b = \frac{6\times 10020 - 116\times 516}{6\times 2292 - (116)^2} = \frac{264}{296} = 0.8919$$

$$a = \frac{516}{6} - 0.8919\times\frac{116}{6} = 68.7566$$

This gives a regression equation of $y = 68.7566 + 0.8919x$.

The correlation and regression results are also given on the spreadsheet using the function wizard (the slight discrepancy on the intercept is due to rounding when using the formula). We may have some concerns about the highest and lowest values on this plot but would generally feel able to use linear regression within the range of the x values (x = 15 to 24)

3. Summary information is given on the following spreadsheet:



Extra question 5

x	y	
15	83	
17	90	
19	86	
20	85	
21	89	
24	83	
116	516	
x^2	y^2	xy
225	6889	1245
289	8100	1530
361	7396	1634
400	7225	1700
441	7921	1869
576	6889	1992
2292	44420	9970

Correlation = -0.12878

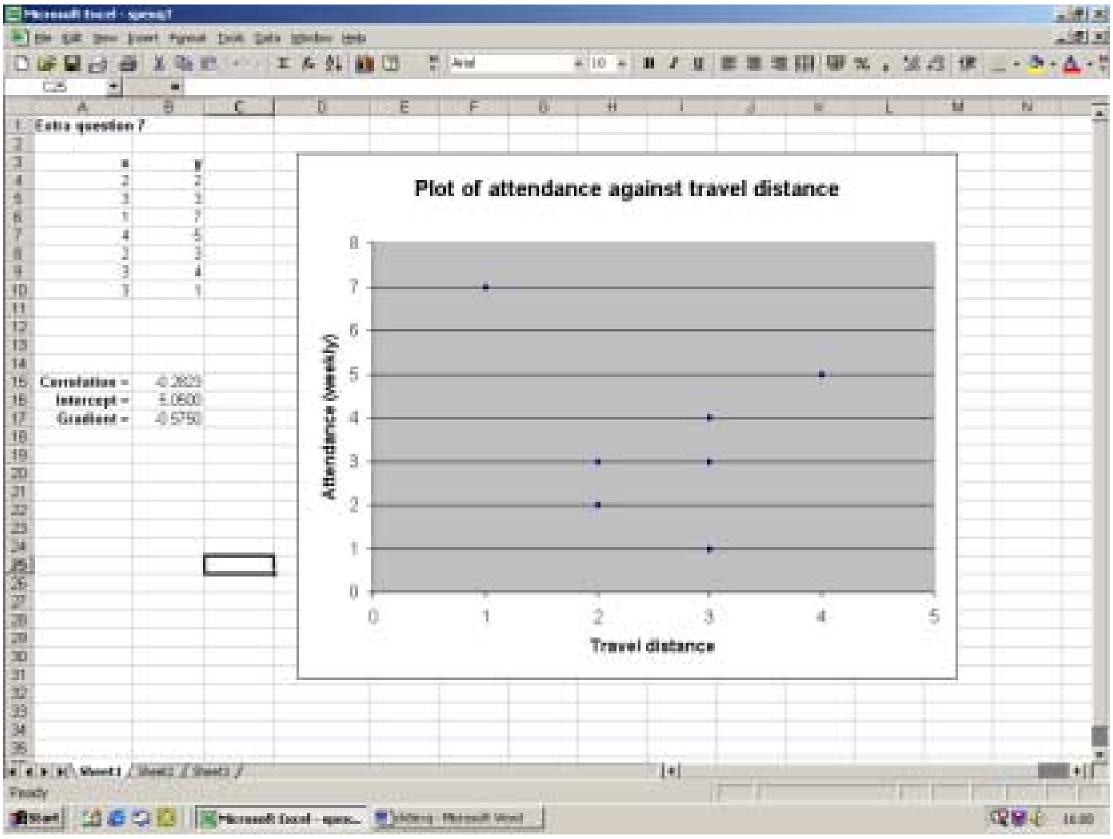
Substituting into the formula given in the text:

$$r = \frac{6\times 9970 - 116\times 516}{\sqrt{\left(6\times 2292 - (116)^2\right)\left(6\times 44420 - (516)^2\right)}}$$

$$r = \frac{-36}{\sqrt{296\times 264}} = -0.12878$$

The closeness of this correlation coefficient to 0 does suggest that fitting a straight line to this data is not worthwhile and for this reason we have not included the regression coefficients. In fact, you only need to look at the graph to see a fairly random scatter of points. However, the spreadsheet will give regression coefficients if you request them. You should always examine the graph plot and the closeness of the correlation coefficient to 0 before you decide whether to proceed with regression.

4. We can again use a spreadsheet to examine the graph and consider the calculated values.



Extra question 7

x	y
2	2
3	3
1	7
4	5
2	3
3	4
3	1

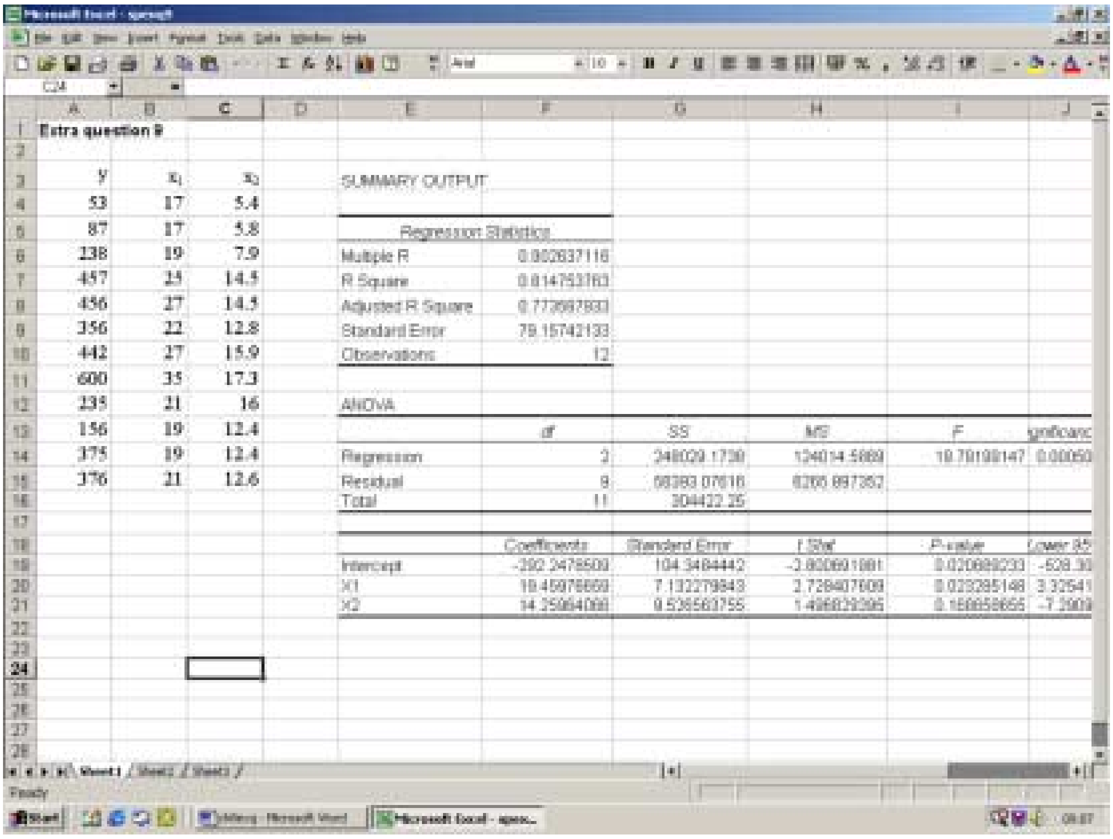
Correlation =	-0.2823
Intercept =	5.0500
Gradient =	-0.5750

(a) Attendance has been plotted on the y axis and travel distance on the x axis and this is suggestive of the direction of the relationship. There is no very distinctive pattern emerging from the graph. It is clear with such a small data set, that individual values can make a big difference. It is always important to get as much quality data as you can. In this case we might also be concerned that distance has been measured in miles. The customer might see big differences between fractions of a mile.

(b) The correlation is close to 0 but negative. Also one observation might be making all the difference here. One the basis of the observed relationship (just a small scatter of points) and the correlation, there is no real justification for fitting a straight line.

(c) Given that fitting a straight line would not be seen as worthwhile, there is little point in calculating the regression coefficients. To give the regression line can indeed by misleading if it is used for prediction purposes.

5. The printout shown below was produced on Excel using Regression within the Data Analysis tools. If this is installed you will find it in the Tools menu. If you need to install, go to Add-ins under the Tools menu and tick for Analysis Toolpak.



Extra question 9

y	x_1	x_2
53	17	5.4
87	17	5.8
238	19	7.9
457	25	14.5
456	27	14.5
356	22	12.8
442	27	15.9
600	35	17.3
235	21	16
156	19	12.4
375	19	12.4
376	21	12.6

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.902637116
R Square	0.814753763
Adjusted R Square	0.773587933
Standard Error	79.15742133
Observations	12

ANOVA

	df	SS	MS	F	gnificanc
Regression	2	248029.1738	124014.5869	19.79198147	0.00050
Residual	9	56393.07616	6265.897352		
Total	11	304422.25			

	Coefficients	Standard Error	t Stat	P-value	Lower 95
Intercept	-292.2478509	104.3484442	-2.800691881	0.020689233	-528.30
X1	19.45976669	7.132279843	2.728407609	0.023285148	3.32541
X2	14.25964088	9.526563755	1.496829396	0.168858665	-7.2909

The closeness of Multiple R to +1 does suggest that this multiple regression exercise is worthwhile.

The resulting equation is y = -292.2 + 19.5x_1 + 14.3x2

This is only a starting point with this kind of modelling. What we would what to do now is test the importance of each of the variables and make a judgement of what variables to include in our predictive equation. If you are interested in this methodology, you are recommended to look at Curwin and Slater *'Quantitative Methods for Business Decisions*' or other more advanced books.

6. The sum of squared difference of ranks is shown below:

Applicant	Application ranking	Interview ranking	d	d^2
A	3	4	-1	1
B	5	6	-1	1
C	1	2	-1	1
D	6	5	1	1
E	2	1	1	1
F	4	3	1	1
G	7	7	0	0
				6

$$r = 1 - \frac{6 \times 6}{7(49-1)} = 1 - \frac{36}{336} = 0.8929$$

This value of r shows a high level of agreement between the ranking of application forms and the ranking at interview.