Brief notes on statistics: Part 2 Scatter diagrams, correlation (Kendall’s and Pearson’s) and regression

Michael Wood (Michael.wood@port.ac.uk)

22 October 2012

Introduction and links to electronic versions of this document and the other parts at http://woodm.myweb.port.ac.uk/stats. The data in the tables, and the figures, are in the spreadsheet, http://woodm.myweb.port.ac.uk/stats/StatNotes.xls . For a rough, but still useful, understanding, you can ignore the numbered endnotes, which provide extra detail.

This parts looks at how to deal with **two variables** using concepts that are widely used in research. Do tall people earn more than short people? Does money buy happiness? Answers below. In Part 4 we will see how the ideas can be extended to more than two variables.

The (fictitious) data in Table 1 comes from an experiment in which an organization tried different levels of advertising with a view to seeing the impact on sales. (I have only used six pairs of observations in Table 1 to help you see what is going on. Obviously, useful research in the real world is likely to use much more data.)

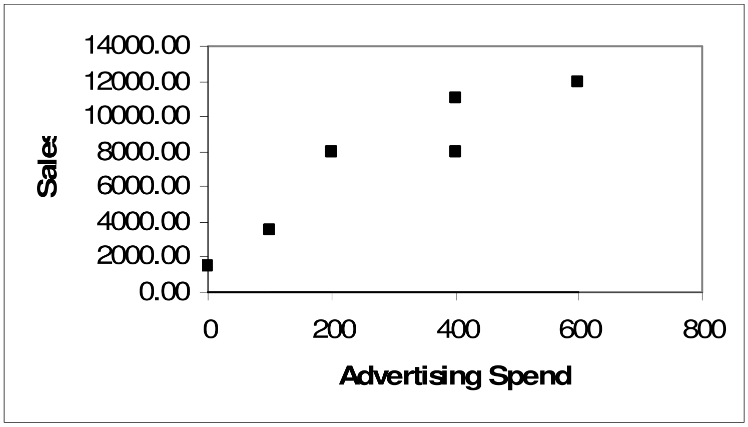
Table 1: Data showing association between Advertising and sales

Advertising Spend (£)	Sales (£)
200	8000
100	3500
400	11000
600	12000
0	1500
400	8000

Scatter diagrams

This data shows a very clear pattern: the best way of seeing this is to draw a *scatter diagram* (easy in Excel: it is one of types of diagram in the Chart Wizard):

Figure 1: Scatter diagram based on Table 1



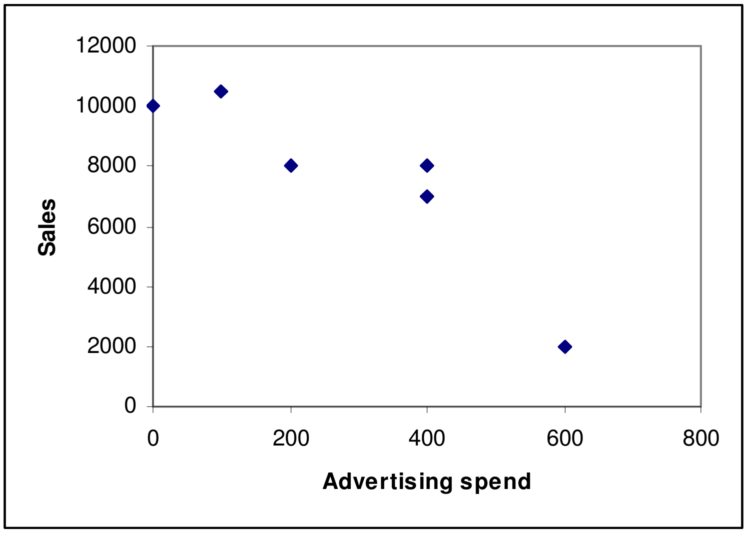
If, on the other hand, Advertising made little difference, or if it reduced sales, this would be obvious from the pattern in the scatter diagram. For example, Table 2 and Figure 1a tell a very different story (which you will explore in Exercise 1 below):

Table 2: Another set of data showing association between Advertising and sales

Observation Number	Advertising Spend (£)	Sales (£)
1	200	8000
2	100	10500
3	400	7000
4	600	2000
5	0	10000
6	400	8000

(The Observation Number column is to clarify the relationship with Table 3 below.)

Figure 1a: Scatter diagram based on Table 2



=

Scatter diagrams make these relationships very obvious. There are some more examples at http://woodm.myweb.port.ac.uk/stats/Scatters.pdf (apologies for the poor quality) and some impressive dynamic scatter diagrams in the video at http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

Correlation coefficients

It is sometimes useful to measure the degree of association between two variables by means of a single number called a correlation coefficient. There are three commonly used correlation coefficients: **Pearson's, Kendall's and Spearman's**.

I'll explain how Kendall's works because this one is the easiest. (It is sometimes called Kendall's *tau* after the Greek letter used as a symbol.) All we do is to take the **observations in pairs**, and ask whether the association is **positive** in the sense that the observation with the higher value of one variable also has the higher value of the other variable (scored as +1), **negative** in the sense that the observation with the higher value of one variable has the lower value of the other (-1) or neither (0). Some of the results for Table 2 are in Table 3.

Table 3: Calculation of Kendall's correlation coefficient for data in Table 2

Observation Number	1	2	3	4	5	6
1		-1	-1	-1	-1	0
2			-1	-1	+1	-1
3						
4						
5						
6						

For example, in Table 3, the entry in *italics (-1)* indicates that the association between the first and second observations (rows) in Table 2 is *negative.* This means that the *higher* value of one variable is associated with the *lower* value of the other. In this case, the first observation has the *higher* advertising spend (200 as opposed to 100) but the *lower* sales figure (8000 as opposed to 10500). You can also see this in Figure 1a. The two points are the second and third from the left, and the line joining them goes *downhill* indicating a negative relationship.

On the other hand the association between rows 2 and 5 is *positive* (+1 in the table) because the *higher* advertising spend is associated with the *higher* sales figure. (Or the lower advertising spend is associated with the lower sales figure – which is saying the same thing.) You can also see this in Figure 1a. These two observations are the two points on the left of Figure 1a. A line joining these two points goes *uphill* indicating a positive relationship.

The 0 indicates that the association is neither positive nor negative because the two observations are equal on one of the variables (both sales figures are 8000).

For obvious reasons we only need to fill in the triangle at the top of the table (the triangle at the bottom would be repeating the same analysis).

The Kendall correlation coefficient is defined as the mean (average) of the numbers in the Table[1].

Quick question 1. Complete Table 3 and work out Kendall's correlation coefficient. Now work out the correlation for Table 1 and Figure 1.

It should be obvious that **positive correlations** correspond to situations where higher values of one variable tend to be associated with higher values of the other, and **negative correlations** to situations where higher values of one variable tend to be associated with lower values of the other. A correlation of zero would indicate that the there are as many positively associated pairs as negatively associated pairs so the two cancel out and the resulting correlation is zero. A small positive correlation (e.g. 0.2) would indicate a slight tendency for higher values of one variable to be associated with higher values of the other. And so on.

A **scatter diagram will show you roughly what the correlation is**. For example, in Figure 1, the correlation is obviously positive because the points with the higher advertising spend also tend to have the higher sales. It's obviously fairly high because the relationship is a strong one, but it's not +1. (It would be +1 if the scatter diagram was an uphill straight line.) On the other hand, the pattern in Figure 1a shows a negative correlation.

In practice, the most commonly used correlation coefficient is **Pearson's**, largely because it meshes neatly with other aspects of statistical theory (e.g. regression, which we come to next). This is the correlation which the Excel function **correl** will find for you. Unfortunately, the formula for working it out is a little complicated[2].

Quick question 2. Have you any idea what this (Pearson's) correlation coefficient will be for Tables 1 and 2? Now check with Excel.

All correlation coefficients are always between +1 and -1. **They measure the direction and consistency of the association between the two number variables**[3]. Many research papers include a table showing correlations between all the variables in the study – e.g. http://woodm.myweb.port.ac.uk/stats/CorrelMatrix.pdf. These are usually Pearson correlations, but Tables 1 and 2 in Moutafi et al (2007) (http://woodm.myweb.port.ac.uk/stats/KendallCorrel.pdf) use Kendall coefficients[4].

Quick question 2a. What do you think a scatter diagram corresponding to a Pearson correlation coefficient of –0.23 would look like. What about a correlation of +1 and –1?

The next section is on regression. This is an idea which can be used for analyzing the relation between two number variables, but it can also be extended to deal with more than two variables.

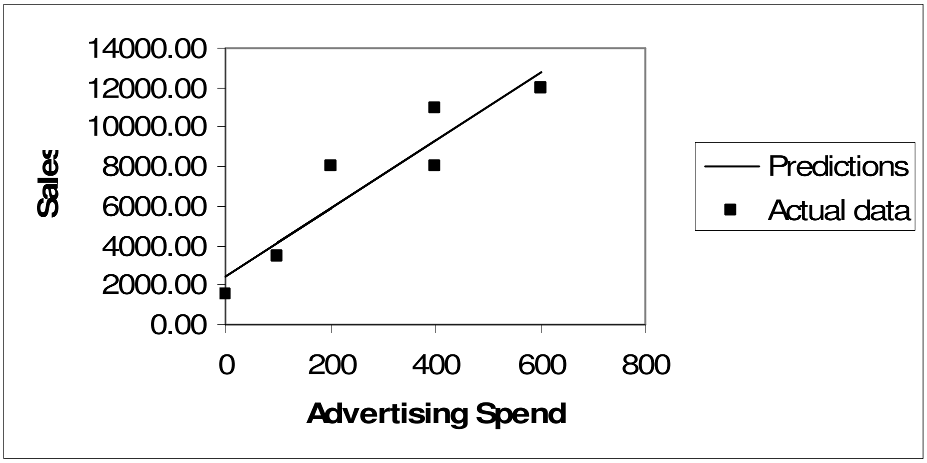
Regression

Figure 1 suggests that the more we advertise, the more we are likely to sell. This relationship means that we should be able to predict (roughly) sales from the amount spent on advertising. We can do this with a *regression model*.

A *regression model* is an equation (i.e. an arithmetical rule) for predicting values of one variable from values of one or more other variables. In practice, as well as being used to make predictions, regression is also often used to help understand or explain the situation, or to assess the impact of one variable on another – we will return to this after we have seen how regression works.

Figure 1 shows a clear relationship between the two variables and suggests that we could superimpose a line to predict sales from advertising spend:

Figure 2: Regression model based on Table 1 and Figure 1



In practice you could probably draw this line "by eye", but it is useful to be able to do it automatically – because then we can get a computer to do it, and because the same method can be used for more complicated situations.

In Figure 2, Sales is the *dependent* variable because we are assuming it depends on the Advertising Spend. Advertising Spend is the *independent* variable because we are assuming it is not dependent on Sales. Alternative terms are ***predicted* (dependent)** and ***predictor* (independent)** variables. Excel uses the terms **Y** (dependent) and **X** (independent).

The method for fitting a regression line: least squares

The first step is to see how we can describe a line. You may remember the equation

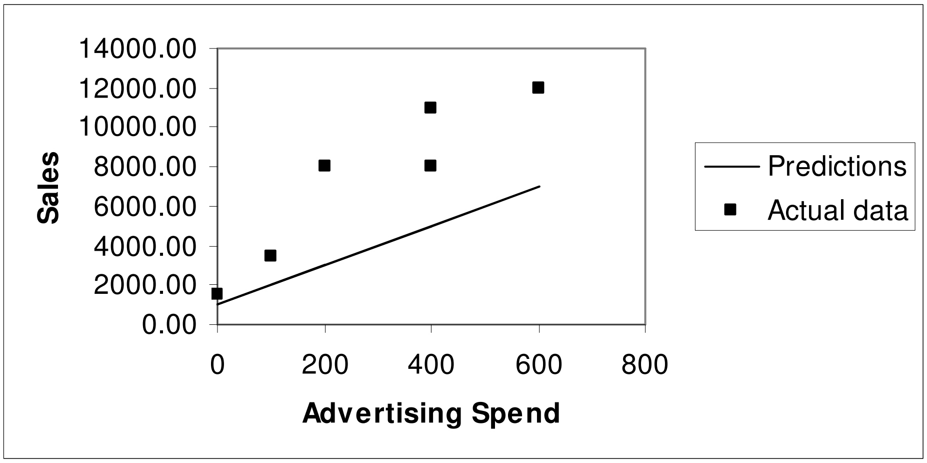
$$y = mx + c$$

from school. If not here it is in words (*y* is PredictedSales, *m* is Slope, etc):

PredictedSales = Slope x AdvertisingSpend + Constant (Equation 1)

To clarify the method, I'll start with a line which is obviously not the best one – say with the Constant = 1000, and the Slope = 10. The line we get is (click on the Graph tab to see this):

Figure 3: A rather poor prediction line (from the Graph sheet of pred1var)



The line is simply where all the predictions lie. To see this, let's take any point on the Advertising Spend axis – say Advertising Spend = 400.

Quick question 3. What is the PredictedSales for AdvertisingSpend = 400 (and Constant=1000 and Slope=10)?

Your answer should be on the line. Whatever value of Advertising Spend you choose you will find that the prediction is on the line. That's what the line is!

As well as making predictions, it's also helpful to see what the constant and the slope mean. First the constant. **If Advertising Spend is 0**, Equation 1 says simply

PredictedSales = Constant

In other words the constant is the predicted sales when Advertising Spend = 0. In the graph, this corresponds to where the line crosses, or intersects, the vertical axis (since here AdvertisingSpend is 0). For this reason, another word for the constant is ***intercept****.*

As the name suggests, **the slope tells you how steep the line is**. To see why, let's work out the PredictedSales for an AdvertisingSpend of 401. This is:

PredictedSales = 10 x 401 + 1000 = £5010

This is 10 more than the PredictedSales for an AdvertisingSpend of 400. In other words, if the AdvertisingSpend goes up by one unit, the PredictedSales will go up by 10 units – the value of slope. ***This is what the slope tells us: it's the increase in the prediction for the dependent variable if the independent variable is increased by one.***

Quick question 4. What would a slope of –30 indicate? What will the graph look like?

Quick question 4a. What do you think the slope and constant are for the best fit regression line in Figure 2? (The computed answer is just below, but don't look till you've tried to estimate it.)

This is how we can describe any line – all we need is to say what the constant (or intercept) and the slope are. (There's a video on this, and the next bit, at http://youtu.be/RIAcq0NMGtA .)

Now to find the line that fits best. To do this we use an approach known as the ***method of least squares***. This is a way of choosing the slope and constant (intercept) so that the line fits the data as closely as possible. There is a detailed explanation of how this works at http://woodm.myweb.port.ac.uk/stats/LeastSquares.pdf. (It's not essential to read this but it should help you understand the basic principle behind simple regression and many other statistical techniques.)

In our case the **slope = 17.25**, and the **constant = 2446**. These are the values that give the line in Figure 2. **The model predicts that spending an extra £1 on advertising will increase sales by £17.25.**

Interpreting and using a regression model

You can now use this value of the constant and slope to make predictions just like we did above for the rather poor model in Figure 3. The interpretation of the slope is just like before, only now it is more useful because it refers to the best prediction. **The slope is the additional Sales that are predicted for an increase in the AdvertisingSpend of one unit (pound) – a very useful piece of information. The slope is also called the *regression coefficient* or the *x coefficient.***

Quick question 5. Use the model to predict the sales if £500 is spent on advertising. How much extra sales would be predicted if one more pound was spent on advertising?

There is another essential bit of information provided by the least squares method. This is ***R squared***. This is a measure of how well the model fits the data – R squared 0 representing a useless prediction where the independent variable is of no help, and 1 representing a perfect prediction with no errors. In the present case R squared is 0.87.

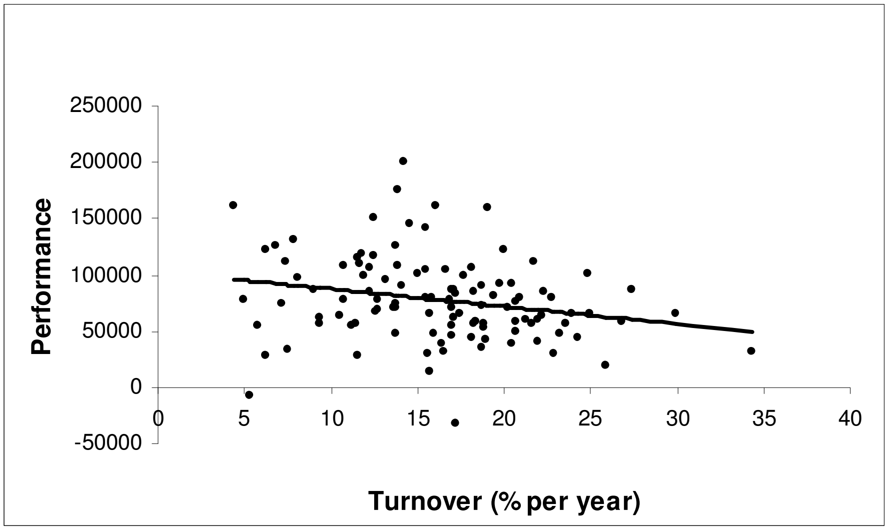
The reason for the term R squared is that it is equal to the **square of the Pearson correlation coefficient**, and the conventional symbol for this is *r.* This should make sense if you remember that if the correlation is zero the prediction from the regression line will be useless (the independent variable is not related to the dependent and so is no help), whereas if it is +1 or –1 the prediction will be perfect because all the data points lie in a straight line.

Quick question 6. What is the Pearson correlation between variables used in the regression model in Figure 2 if R squared is 0.87.

Another way in which R squared is sometimes interpreted is by saying that **AdvertisingSpend *explains* or *accounts for* 87% of the variation**[5] in Sales figures. I'll look at this idea in more detail in Part 4 of these notes.

Figure 4 provides another example. This is based on data on staff turnover and a measure of performance from 110 branches of an organization in Europe. The detailed results of this research are described in Glebbeek and Bax (2004). I am grateful to Dr Arie Glebbeek for permission to put some of this data in the Turnover sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls .

Figure 4: Actual performance and predicted performance of 110 branches



Quick question 7. What do you think the regression coefficient (slope), R squared and the Pearson correlation coefficient are here?(You should be able to estimate these very roughly.)

Figure 4 corresponds to a model with a low value of R squared. The scatter in the diagram shows clearly that a prediction of performance based only on staff turnover will not be accurate.

Quick question 8. What would the equivalent graph to Figure 4 look like if R squared was 0, … or 1?

Software for regression

Mathematicians have derived formulae[6] for the slope and intercept which give identical results to the method using the solver described at http://woodm.myweb.port.ac.uk/stats/LeastSquares.pdf. These formulae are incorporated in many computer packages: **Excel, for example, has worksheet formulae for the slope and intercept, if you right click on the points of a scatter diagram it will offer various trend lines**

(and an equation and a value for R squared) and there is also a Regression Tool – which we will look at when we come on to multiple regression.

Regression in practice

Regression is very widely used in research, in business and in many other walks of life. We will return to it later, when we'll see how we can incorporate more than one independent variable. This makes the method far more powerful (see Ayres, 2007, for an evangelistic account of what regression can do). However, even with one variable it can still be very useful. For example, there is a well known way of calculating maximum heart rate by subtracting a person's age from 220. So for a 70 year-old the maximum heart rate should be 220 – 70 or 150 beats per minute. Obviously this will vary from individual to individual, but this is a standard way of making a rough prediction.

Quick question 9. Thinking of this as a regression equation with maximum heart rate as the dependent variable and age as the independent variable, what are the slope and the constant?

According to Robergs and Landwhr (2002) this formula is not based on any systematic research. A better prediction can be made using a slope and constant derived from empirical data on the age and maximum heart rate of a large sample of people (using the least squares method described above, of course) – this gives a slope of –0.685 and a constant of 205.8.

Quick question 10. What prediction for a 70 year old's maximum heart rate does this slope and constant give?

Another result from a regression study is that "each inch of height is associated with a 1.5 percent increase in wages, for both men and women [in the UK]" (Case et al, 2008). According to this research tall people do earn more than short people!

Summary

Scatter diagrams are a useful way of showing the relationship between two number variables. You can also summarize this relationship with a correlation coefficient – these are always between +1 and -1 and measure the direction and consistency of the association between the two number variables – or a regression model which uses one variable to try to predict the other by means of a straight line. (There is more on regression in Part 4.)

Exercises

1 Using the data in *Table 2* above (which you can find on from the Part2Data sheet of http://woodm.myweb.port.ac.uk/stats/StatNotes.xls)

a) Produce a scatter diagram. Highlight the data and click on the Chart wizard or Insert in Excel 2007 and find Scatter …

b) What advice would you give to the organization about its advertising on the basis of this data?

c) Calculate the Pearson correlation coefficient. (**C**lick f_x, then find correl.**)**

d) Now right click on one of the points of the scatter diagram and click on Add Trendline. Then tick Linear, Display equation and Display R squared. This is regression line. Make sure you understand everything. Check that R squared is the square of the Pearson correlation (the answer to the previous question).

e) What are the predicted sales for an Advertising spend of £500, and £2000? Estimate these from the line, and then calculate them from the equation. Are the answers sensible?

f) What does the slope tell you about the impact of advertising on sales?

g) Finally, change the data so that there is now a correlation of roughly zero. Check that you can see how the correlation coefficient and the regression line relate to the scatter diagram. Now change the data again to give a correlation of exactly +1, and then – 1.

2 The excel workbook at http://woodm.myweb.port.ac.uk/nms/drink.xls contains data on the smoking habits of a sample of students. Use this to investigate the relationship between age and the amount students said they drank on Saturday (Satunits). You should produce a scatter diagram, work out a correlation coefficient, and a regression line for predicting the amount drunk (dependent, Y variable) from the age (independent, X variable) by right clicking on the points of the scatter diagram. What is the slope and R squared, and what do they mean? Is the regression line sensible?

3 See what you can work out from the data on American cars at http://www.amstat.org/publications/jse/datasets/kuiper.xls

4 Click on the Exam worksheet in http://woodm.myweb.port.ac.uk/stats/StatNotes.xls and work out the correlation between the exam marks and the assignment marks. Is the answer what you would have expected?

Now divide each of the assignment marks by 5, and work out the correlation between the new (reduced) assignment marks and the exam marks. Is the answer what you would have expected?

Now suppose, that with another set of marks, the correlation was –0.8. What would this suggest about the exam and the assignment?

Additional exercises

1 The table below shows the ages and earnings of a sample of people living in a town.

Age	Earnings (£)
56	49000
40	9500
56	17500
41	34500
53	8500
53	12000
29	8000
20	1500
44	19500
34	55000
49	30000
34	10500
42	11000
22	1000
27	9000
26	4500

Set up a regression model to see how earnings depend on age. What earnings would you predict for a 30 year old, a 50 year old, and 100 year old? How accurate do you expect your answers to be? Explain the reasons for any inaccuracies. What is the slope and what does it mean?

2 The marks for a group of four students in two examinations were:

Mathematics *Ann: 60, Bill: 75, Sue: 85, Dan: 90*

English: *Ann: 50, Bill: 48, Sue: 49, Dan: 46*

The standard deviation (*stdevp)* of the marks for maths is 11.5, and for English is 1.5. (You should be able to work these out with Excel, and without.) The Pearson coefficient between the marks in the two subjects is -0.8.

Work out Kendall's correlation coefficient. Is it similar to the Pearson coefficient?

Assuming that these marks are reasonably typical of all students doing the two examinations, what comments would you make about the difference between the two standard deviations, and about the correlation? Would it be fair to add the marks in the two examination to give students an overall mark?

You should be able to answer the following questions without a computer or calculator:

(a) What would the standard deviation of the maths marks be if the marks were halved (ie they became 30, 37.5, 42.5, 45)?

(b) What if 30 was deducted from each mark?

(c) What effect would both of these changes have on the correlation? Would it still be -0.8 in both cases?

(d) Find the standard deviations of these marks:

60, 60, 75, 75, 85, 85, 90, 90 (compare with the maths marks above)

(e) And the standard deviation of these:

70, 70, 70, 70.

Use a computer available to check these.

3 Can you find a sample of four pairs of numbers for which the Kendall correlation is -1, but the Pearson correlation is slightly greater than this?

4 Find some data of your own where regression may be useful or interesting, and produce a regression model and see if it leads to any interesting results.

5 Dissanaike (1999) produced a regression model to predict the return which investors would receive from investing in a particular security for a period of four years, from the return they would have received if they had invested in the same security in the previous four years. The data on which the model was based were the returns for a sample of large companies over consecutive periods of four years.

The regression coefficient cited was -0.112, and the value of R squared was 0.0413.

Suppose you were considering investing in two shares: A or B. A has produced a return over the last four years of -5%, and B has produced +5%. Use the regression model to predict which share is likely to produce the better returns over the next four years, and by how much. How sure would you be?

References

Ayres, I. (2007). *Super crunchers: how anything can be predicted.* London: John Murray.

Case, A.; Paxson, C. & Islam, M. (2008). Making Sense of the Labor Market Height Premium: Evidence From the British Household Panel Survey. *NBER Working Paper No. 14007*. Retrieved from http://www.princeton.edu/~accase/downloads/w14007_Case_Paxson_Islam.pdf on 14 August 2011.

Dissanaike, G. (1999). Long term stock price reversals in the UK: evidence from regression tests. *British Accounting Review, 31,* 373-385.

Glebbeek, A. C. & Bax, E. H. (2004). Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal, 47,* No 2, 277-286.

Moutafi, J, Furnham, A, & Crump, J. (2007). Is managerial level related to personality? *British Journal of Management, 18*, 272-280.

Robergs, R. A. & Landwehr, R. (2002). The surprising history of the "Hrmax=220-age" equation. *Journal of Exercise Physiology Online, 5*, 2.

Answers to Quick questions and notes on some exercises

1 There is one +1, two 0s and twelve –1s. Adding these up gives 11, and the average (divide by 15) is –0.73. The correlation for Table 1 is +0.87 (13/15). The easiest way to see this may be to look at a scatter diagram: look at each pair of points in turn and check if the line between them is "uphill" (+1), or "downhill" (– 1).

2 You should expect the Pearson correlations to be similar (because it's another measure of correlation) but not identical (otherwise there wouldn't be another name for it!). Using Excel, the Pearson correlations are +0.93 for Table 1, and –0.90 for Table 2.

2a Figure 4 above is a scatter diagram with a Pearson correlation of –0.23. There a slight tendency for high values of one variable to be associated with low values of the other, but it only very slight. A correlation of zero 0 would indicate no relationship at all: high values of one variable are equally likely to be associated with high and low values of the other variable.

The other extremes are correlations of +1, which would indicate an uphill straight line, and –1, which indicates a downhill straight line. (The steepness of the slopes of these lines is irrelevant.)

3. The prediction from the line for an AdvertisingSpend of 400 is (click on the Predictions tab to see this in pred1var):

PredictedSales = 10 x 400 + 1000 = 4000 + 1000 = £5000

4 A slope of -30 would be a *downhill* slope.

5 The prediction for £500 advertising is sales of £11,070. An extra one pound advertising expenditure (making it £501) leads to an extra £17 (the slope) in predicted sales – i.e. £11,087.

6 As R squared is 0.87, the correlation coefficient, r, is the square root of 0.87 which is +0.93. (The square root could be -0.93, but not in this case because the correlation is obviously positive.)

7 Regression coefficient (slope) = – 1552; R squared = 0.055; Pearson correlation coefficient = – 0.23. These are worked out from the spreadsheet, but you should be able to guess rough answers from the graph. To work out the slope you need to look at how much the

Performance changes for each 1% change in Turnover. The correlation is clearly negative and fairly small, so R squared (the square of the correlation) will also be small.

8. If R squared were 0, the scatter graph would show no correlation, and the prediction graph would be a horizontal line through the middle (average) of the performance scale. If R squared were 1, the data points would all lie on a straight line which would be the prediction line – either sloping up (r = +1), or down (r = – 1)?

9. The slope is –1 and the constant is 220.

10 The predicted maximum heart rate for a 70-year-old is 158.

Exercise 1

a) The scatter diagram should show a clear negative relation.

b) … so the advertising should obviously be stopped because more advertising leads to reduced sales!

c) The Pearson correlation is -0.898.

e) The predicted sales with Advertising Spend = £500 is £4930 (=-12.248*500+11054), and with Advertising Spend = £2000 the prediction is -£13442. The first answer is sensible; the second is not. This illustrates the principle that regression predictions may be wildly inaccurate where the value of the independent variable is right outside the range covered by the data (all of which are between £0 and £600).

f) The slope is -12.248 which tell us that each extra £1 spent on advertising leads to a predicted *fall* in sales of £12.248.

Endnotes – these include more detail which can be ignored for a rough understanding

[1] This method is slightly simpler than the standard method implemented by statistics packages such as SPSS in the way it deals with ties (pairs of observations which are equal on one of the variables). However, the difference between the answers is usually small and not worth worrying about.

[2] see http://woodm.myweb.port.ac.uk/dra/VarianceCovariance.doc.

[3] This is a rather vague interpretation. To understand the interpretation of Pearson correlations in more detail you need to square them to produce the statistic called R^2 which has a precise interpretation explained in the section on regression. For Kendall

coefficients, you need to think through how they are derived – e.g. if the Kendall correlation is –0.5 the proportion of negatively related pairs must have been 75% and the proportion of positively related pairs must have been 25%.

[4] The advantage of the Kendall correlation is that it ignores the size of the differences between the values of the two variables - if some of these sizes are much bigger than the rest, the Pearson correlation can be misleading. Also, Kendall's is much simpler to understand. Pearson's has the advantage that it ties in neatly with a lot of statistical theory (especially regression - R squared is the square of the Pearson correlation). For more detail try typing "Kendall correlation advantages" into Google (without the "...", obviously).

[5] as measured by the variance.

[6] See http://woodm.myweb.port.ac.uk/dra/VarianceCovariance.doc .