

7 Correlation and Regression

7.1 Learning outcomes

On completing this chapter you should be able to:

- calculate the correlation coefficient between two variables, and explain the value;
- calculate the rank correlation coefficient between two sets of data, and explain the value;
- demonstrate the use of regression analysis between two variables to find the line of best fit, and explain its meaning;
- calculate a forecast of the value of the dependent variable, given the value of the independent variable.

7.2 Introduction

In the next two chapters we look at one of the major applications of statistics, namely forecasting. Although there are a number of ways of producing forecasts that involve little or no mathematics, we shall concentrate here on two of the most important quantitative approaches, causal and extrapolative. A *causal* approach is based on the assumption that changes in the variable that we wish to forecast are *caused* by changes in one or more other variables. With an *extrapolative* approach, we examine past data on the variable that is to be forecast, in order to determine any patterns the data exhibits. It is then assumed that these patterns will continue into the future: in other words, they are *extrapolated*.

Let us begin with the causal approach, and, to simplify matters, we shall deal with the case in which the variable to be forecast (the dependent variable, y) depends on only one other variable (the independent variable, x). Such data is called bivariate because in each situation we have two pieces of information, denoted x and y . The dependent variable, y , is also referred to as the *response variable* and the independent variable, x , as the *influencing variable*. So x is assumed to influence a response in y . Before actually looking at how to produce forecasts in such situations, we must consider the question as to how we know that changes in y are caused by changes in x (alternatively: how we know that y depends on x). The answer to this involves the study of *correlation*.

7.3 Correlation

Two variables are said to be *correlated* if they are related to one another, or, more precisely, if changes in the value of one tend to accompany changes in the other. Now, we have already used the (x, y) notation of Chapter 2, and this initially suggests a graphical approach: if there are pairs of data available on the variables x and y , then these can be plotted as points against a set of x - and y -axes. The result is known as a *scatter diagram* or *scatter graph*.

Example 7.A

A company is investigating the effects of its advertising on sales. Consequently, data on monthly advertising and sales in the following month are collated to obtain:

Advertising expenditure in month (£'000)	Total sales in following month (£'000)
1.3	151.6
0.9	100.1
1.8	199.3
2.1	221.2
1.5	170.0

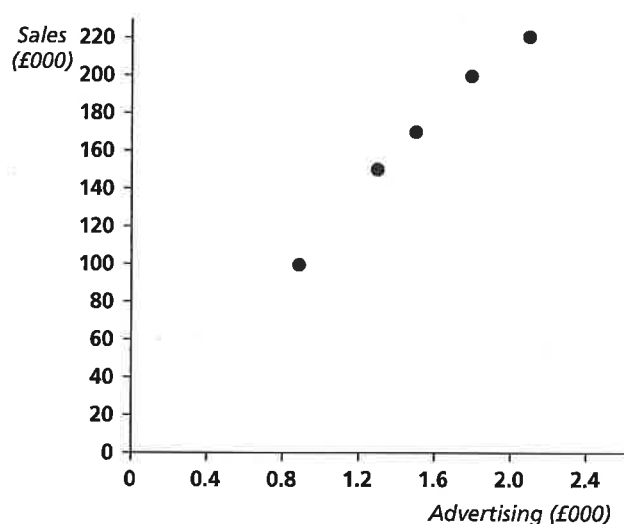
Plot these data on a scatter diagram.

Solution

Since the company is interested in how advertising affects sales, it is clear that sales should be the dependent variable, y , and advertising the independent, x . The scatter diagram is shown in Figure 7.1.

The five pairs of data points are marked as points on the graph.

Figure 7.1 Scatter diagram (Example 7.A)



Since, unlike in Chapter 2, we have not been told that y is a function of x here, the points have not been joined up. In particular, although the points appear to be *close* to lying on a straight line, they do not lie *exactly* on a line; we do not know that a linear function is involved and so none has been drawn in.

The scatter diagram in the above example seems to show a case where the two variables *are* related to one another. Further, the relationship seems to be of an approximately *linear* nature: it is an example of linear correlation. Since the approximation is so good, and the points are close to a straight line, we talk of *strong* linear correlation. Finally, as the gradient of the 'line' is positive, it is *positive linear* (or direct) correlation.



Exercise 7.1

Most of the exercises in this chapter relate to the following table. A company owns six sales outlets in a certain city. The sales last year of two of its key products are given below, together with the sizes of each outlet:

Outlet	Floor space m^2	Sales of L '000 units	Sales of M '000 units
A	75	22.4	30.7
B	60	21.1	12.9
C	108	29.6	47.1
D	94	27.1	38.8
E	92	27.0	41.5
F	130	36.9	79.0

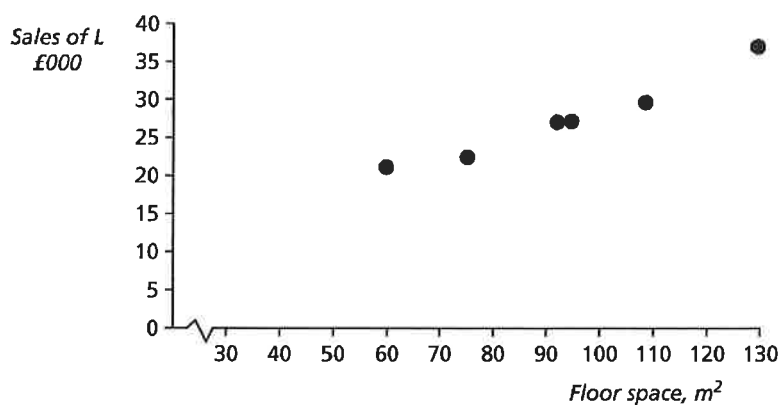
The company is investigating the effects of outlet size on sales.
Plot scatter diagrams of:

- sales of L against size;
- sales of M against size.

Solution

Did you plot size of outlet as x in both cases? We are investigating the effect of size on sales, so sales must be the dependent variable.

(a) **Figure 7.2** Scatter diagram (Exercise 7.1(a))



(b) **Figure 7.3** Scatter diagram (Exercise 7.1(b))

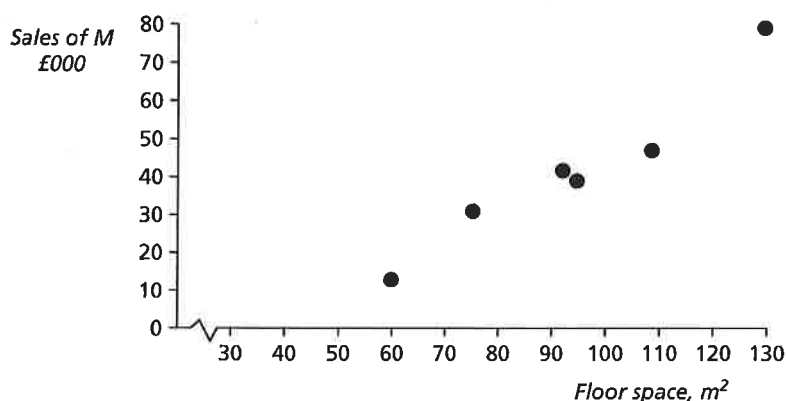
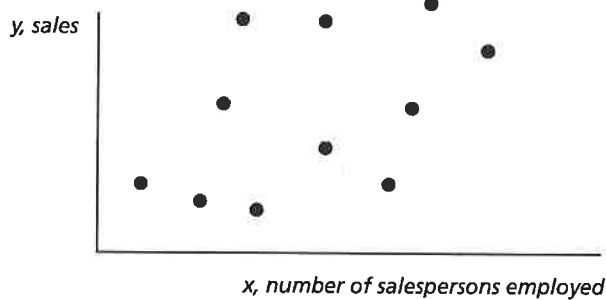


Figure 7.4 shows further examples of:

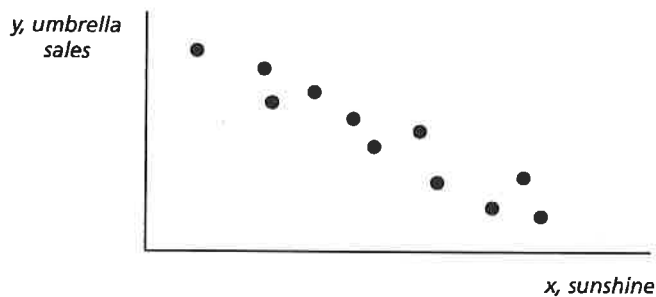
- very weak positive linear correlation, in which y shows a slight tendency to increase as x does;
- strong negative linear (or inverse) correlation, in which y shows a strong tendency to decrease as x increases; and
- non-linear correlation, in which x and y are clearly related, but not in a linear fashion.

Figure 7.4 Examples of correlation

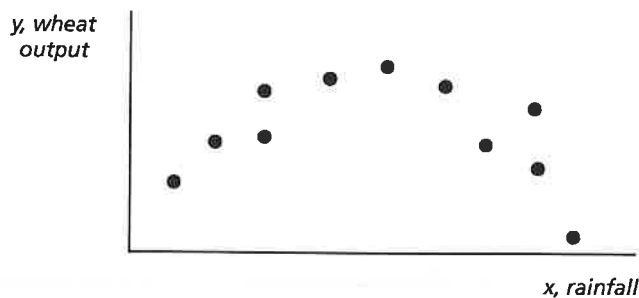
(a) Very weak positive linear



(b) Strong negative linear



(c) Non-linear



Although such scatter diagrams are useful for getting a feel for the presence or otherwise of correlation, it is often difficult to judge what is 'weak' and what is 'strong', and, indeed, whether a large number of points on a diagram constitute any correlation at all. Therefore an objective *measure* of correlation is needed.



Exercise 7.2

From your own experience, try to think of pairs of variables that might have the different degrees of correlation, from weak to strong and from negative to positive.

Solution

These are just some examples:

- costs probably have a strong positive correlation with the number of units produced;
- number of deaths on the roads probably has a middling positive correlation with traffic levels;
- the level of street crime is often thought to relate to the level of visible policing, so the correlation would be negative but probably not strong;
- a strong negative correlation would probably be found if almost any measure of bodily function, such as the condition of the heart, were compared with age in adults, although the graph is unlikely to be perfectly linear.

7.4 Pearson's correlation coefficient

The statistician Pearson developed a measure of the amount of linear correlation present in a set of pairs of data. Pearson's correlation coefficient, denoted r , is defined as:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

where n is the number of data points.



This formula is given in your exam so you do not need to worry about remembering it.

This measure has the property of always lying in the range -1 to $+1$, where:

- $r = +1$ denotes perfect positive linear correlation (the data points lie *exactly* on a straight line of positive gradient);
- $r = -1$ denotes *perfect* negative linear correlation (again the data points lie on a straight line but with a negative gradient); and
- $r = 0$ denotes no *linear* correlation.

The strength of a correlation can be judged by its proximity to $+1$ or -1 : the nearer it is (and the further away from zero), the stronger is the linear correlation. A common error is to believe that negative values of r cannot be strong. They can be just as strong as positive values except that y is decreasing as x increases.

Example 7.B

Evaluate Pearson's correlation coefficient for the data on sales and advertising spend in Example 7.A, and interpret its value.

Solution

As with previous calculations involving summations, we facilitate the calculations by setting them out in columns:

x	y	x^2	y^2	xy
1.3	151.6	1.69	22,982.56	197.08
0.9	100.1	0.81	10,020.01	90.09
1.8	199.3	3.24	39,720.49	358.74
2.1	221.2	4.41	48,929.44	464.52
1.5	170.0	2.25	28,900.00	255.00
7.6	842.2	12.40	150,552.50	1,365.43

$$\text{Thus, } r = \frac{(5 \times 1,365.43) - (7.6 \times 842.2)}{\sqrt{[(5 \times 12.4) - 7.6^2] [(5 \times 150,552.5) - 842^2]}} = \frac{426.43}{\sqrt{4.24 \times 43,461.66}} = 0.993$$

The value of Pearson's correlation coefficient in this case is 0.993. The arithmetic in such a calculation can be seen to be potentially very tedious. It is worthwhile investigating the availability of any computer packages or special functions on a calculator in order to ease the computation of correlation coefficients.

The value of the coefficient in this case is clearly very close to the value 1, indicating a very strong positive linear correlation, and reflecting the close proximity of the points in Figure 7.1 to a straight line.



Exercise 7.3

Using the data on floor space and sales from Exercise 7.1, evaluate Pearson's correlation coefficients for:

- (a) sales of L and size;
- (b) sales of M and size.

Solution

- (a) The necessary summations are $n=6$; $\Sigma x=559$; $\Sigma y=164.1$; $\Sigma x^2=55,089$; $\Sigma y^2=4,648.15$; $\Sigma xy=15,971.2$. Hence:

$$r = \frac{(6 \times 15,971.2) - (559 \times 164.1)}{\sqrt{(6 \times 55,089 - 559^2)(6 \times 4,648.15 - 164.1^2)}} = \frac{4,095.3}{\sqrt{(18,053 \times 960.09)}} = \frac{4,095.3}{\sqrt{(17,332,504.77)}} = \frac{4,095.3}{4,163.23} = 0.984$$

This is a very strong positive correlation between outlet size and sales of L.

- (b) For M, the summations are $n=6$; $\Sigma x=559$; $\Sigma y=250$; $\Sigma x^2=55,089$; $\Sigma y^2=12,796$; $\Sigma xy=25,898.5$.

These result in $r=0.974$, which similarly shows a very strong positive correlation between outlet size and sales of M.

Many students initially find these calculations very difficult. Even if you got the right answer, you may find it useful to run through the calculations once more. Correlation is a very important topic in business mathematics.

7.5 Interpreting correlation coefficients

In general, it is not always as straightforward to interpret a value of r as in the above case. Although it would be inappropriate for the purpose of this text to go into detailed theory, it must be noted that the sample size (n) has a crucial effect: the smaller the value of n , the 'easier' it is for a large value of r to arise purely by accident.

Very rough guidelines are that, with a sample of ten data points, a minimum correlation of about 0.6 is needed before you can feel confident that any sort of linear relationship holds. With twenty data points, the minimum correlation needed is about 0.4.

Extrapolation is a further danger in the interpretation of r . If your x -values range from 0.9 to 2.1, then $r=0.993$ tells you that there is a near-perfect linear relationship between x and y in that range. However, you know nothing at all about the relationship outside that range. It may or may not continue to be linear. The process of drawing conclusions outside the range of the data is called extrapolation. It often cannot be avoided but it leads to unreliable conclusions.

It is possible that an apparently high correlation can occur *accidentally or spuriously* between two unconnected variables. There is no mathematical way of checking when this is the case, but common sense can help. In the case under discussion, it seems plausible that sales and the advertising spend *are* connected, and so it would seem reasonable to assume that this is not an accidental or spurious correlation.

More importantly here, two variables can be correlated because they are separately correlated to a *hidden third variable*. The size of the region could well be such a variable: larger regions would tend to have larger sales figures and the management of larger regions would tend to have larger advertising budgets. It is therefore *possible* that this high correlation coefficient may have arisen because the variable 'sales' is highly correlated with size of region, advertising expenditure is highly correlated with size of region, but sales and advertising spend are not themselves directly connected.

Even if this third variable effect is not present, we still cannot conclude that y depends on x . The strong correlation lends support to the *assumption* that this is so, but does not *prove* it. *Correlation cannot be used to prove causation.*

7.6 Rank correlation: Spearman's coefficient

There are occasions when the degree of correlation between two variables is to be measured but one or both of them is not in a suitable quantitative form. In such circumstances, Pearson's coefficient cannot be used, but an alternative approach – *rank* correlation – might be appropriate. The most common measure of this type is *Spearman's rank correlation coefficient*, R :

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

where d denotes the difference in ranks, and n the sample size.



You do not need to remember this formula because it will be given in your exam.

The arithmetic involved in calculating values of this coefficient is much easier than that for Pearson's coefficient, as the following example illustrates.

Example 7.C

As part of its recruitment procedures, a company awards applicants ratings from A (excellent) to E (unsatisfactory) for their interview performance, and marks out of 100 for a written test. The results for five interviewees are as follows.

Interviewee	Interview grade	Test score
a	A	60
b	B	61
c	A	50
d	C	72
e	D	70

Calculate the Spearman's rank correlation coefficient for this data, and comment on its value.

Solution

In order to apply the formula, the grades and scores are ranked, with the best scores given a rank of 1. Notice how interviewees a and c share the best interview grade. They therefore share the ranks 1 and 2 to give 1.5 each.

Interviewee	Rank of interview grade	Rank of test score	<i>d</i>	<i>d</i> ²
a	1.5	4	-2.5	6.25
b	3	3	0	0.00
c	1.5	5	-3.5	12.25
d	4	1	3	9.00
e	5	2	3	9.00
				<u>36.50</u>

Hence:

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 36.50}{5(25 - 1)} = -0.825$$

The high negative value (near to -1) indicates that interview grades and test scores almost totally disagree with each other – good interview grades go with the lowest test scores and vice-versa. This should concern the company, as it may mean that one or both methods of judging applicants is faulty. The interpretation of *R*-values is similar to that for *r*. Warnings similar to those in section 7.5 also apply when judging values of *R*.



Exercise 7.4

An expert was asked to rank, according to taste, eight wines costing below £4. Her rankings (with 1 being the worst taste and 8 the best) and the prices per bottle were as follows:

Sample	Rank of taste	Price £
A	1	2.49
B	2	2.99
C	3	3.59
D	4	2.99
E	5	3.59
F	6	3.99
G	7	3.99
H	8	2.99

Calculate Spearman's rank correlation coefficient for this data and interpret your result. What result would you expect if the best-tasting wine were ranked 1 and the worst 8?

Solution

Sample	Rank of taste	Rank of price	<i>d</i>	<i>d</i> ²
A	1	1	0	0
B	2	3	-1	1
C	3	5	-2	4
D	4	3	1	1
E	5	6	-1	1
F	6	7.5	-1.5	2.25
G	7	7.5	-0.5	0.25
H	8	3	5	25
				<u>34.50</u>

$$\text{Hence, } R = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 34.5}{8(64 - 1)} = 1 - \frac{207}{504} = 1 - 0.41 = 0.59$$

There seems to be some positive correlation between price and taste, with the more expensive wines tending to taste better. Given the sample size the result is not really reliable and it cannot be extrapolated to wines costing more than £4. Had the taste rankings been allocated in the opposite order, the correlation would be -0.59 .

Students often find this calculation difficult and it is worth running through it again if you had problems. Probably the most common errors are either forgetting to subtract from 1 or subtracting the numerator from 1 prior to dividing by the denominator.

Tied rankings can also be difficult. B, D and H all cost £2.99. Had they been marginally different they would have been ranked 2, 3 and 4. Since they are identical, they each have the rank of 3 (the average of 2, 3 and 4). Similarly F and G share the ranks 7 and 8 by giving them an average 7.5 each.

The d column is obtained by subtracting rank of taste minus rank of price, but it would be equally correct the other way round.

7.7 Which correlation coefficient to use

If the data have already been ranked, there is no option but to use the rank correlation coefficient (R). Where actual values of x and y are given, Pearson's coefficient (r) should generally be used since information is lost when values are converted into their ranks. In particular, Pearson's coefficient must be used if you intend to use regression for forecasting (see later). The only advantages in converting actual data into ranks and using Spearman's coefficient are:

1. that the arithmetic is easier, but this is a minor point given computers and scientific calculators;
2. that Spearman checks for a linear relationship between the ranks rather than the actual figures. If you simply want to confirm, say, that the variables increase together but have no concern about the linearity of the relationship, you might prefer to use the rank correlation coefficient.

7.8 Regression

The preceding sections give us a way of checking on whether it may be valid to assume that one variable, y , depends on another, x . We now proceed to consider how, after making such an assumption, y can be forecast from x .

For simplicity, we restrict our attention to instances of linear correlation. Thus we are interested in situations where the dependence is in the form of a straight line. As we saw in Chapter 2, this involves equations of the type

$$y = a + bx$$

where a and b are numbers. We are, therefore, initially concerned with determining suitable straight line(s) for the particular problem.

7.9 The least-squares criterion

The approach is illustrated through an example.

Example 7.D

A company has the following data on its sales during the last year in each of its regions and the corresponding number of salespersons employed during this time:

Region	Sales (units)	Salespersons
A	236	11
B	234	12
C	298	18
D	250	15
E	246	13
F	202	10

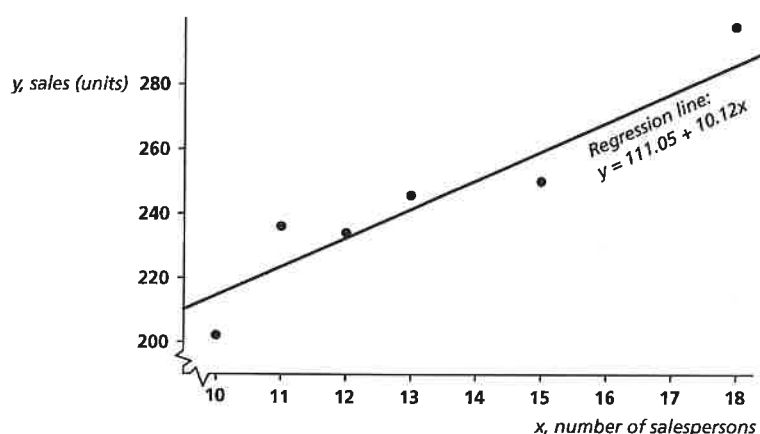
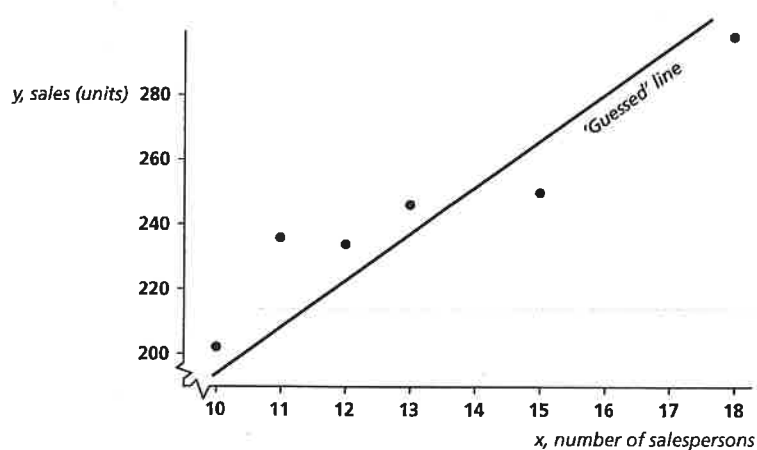
Develop a linear model for forecasting sales from the number of salespersons.

Solution

The linear correlation coefficient between these two variables can be shown to be 0.948. This high value encourages us to assume that sales, y , might depend on the number of salespersons, x , in a linear way.

The scatter diagram for the data is shown in Figure 7.5. For convenience of drawing, the scales on the axes do not start from zero. However, this has the effect of exaggerating the divergences from linearity. A truer impression would be obtained from a graph containing the origin, but this would not be so easy to draw.

Figure 7.5 'Guessed' line and regression line (Example 7.D)



In the upper part of the figure, a straight line has been gauged or 'guessed' by using a ruler to draw a line that appears to be 'close' to all five data points. We have deliberately fitted a very poor guessed line so that the errors are clear. If you do have to fit a line 'by eye', the aim is to follow the slope of the points and to draw the line as far as possible through the centre of the points with roughly equal numbers either side. The lower part of Figure 7.5 shows the best possible fitted line.

This approximate approach may well be accurate enough in many instances, but it is certainly arbitrary. A number of different, equally plausible, lines could be drawn in: the question is, how can you judge whether one line is 'better' than another? Indeed, which is the 'best'?

If we look at the 'guessed' line, it is clear that there are discrepancies between actual y -values and y -values obtained from the line. There are y -errors present, and the sizes of these enable us to judge one line against another. Examples of y -errors in this instance are:

$x = 13$: actual $y = 246$
 y from line ≈ 239 (approximately)
 y -error in line $= -7$

$x = 15$: actual $y = 250$
 y from line $= 266$
 y -error in line $= +16$

Some errors are positive and some negative. Simply adding the errors to judge the 'goodness' of the line, therefore, would not be a sensible idea, as positive errors would tend to be cancelled out by negative ones. To eliminate this effect, we *square* the errors, and deem one line 'better' than another if its sum of *squared* errors is lower. The 'best' line is thus the one with the least sum of squared errors: the so-called *least-squares* regression line of y on x . Without going through the theory, this can be shown to have equation

$$y = a + bx$$

where:

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

and

$$a = \bar{y} - b\bar{x} \quad (\bar{y}, \bar{x}: \text{means of } y \text{ and } x \text{ respectively})$$



These formulae are given in your exam. Note that, because $\bar{y} = a + b\bar{x}$, the regression line always passes through the point $(\bar{x}, \bar{y})$.

The calculation of a and b is set out in a familiar tabular form:

x	y	x^2	xy
11	236	121	2,596
12	234	144	2,808
18	298	324	5,364
15	250	225	3,750
13	246	169	3,198
10	202	100	2,020
<u>79</u>	<u>1,466</u>	<u>1,083</u>	<u>19,736</u>

$$b = \frac{(6 \times 19,736) - (79 \times 1,466)}{(6 \times 1,083) - 79^2} = \frac{2,602}{257} = 10.12$$

$$\bar{x} = \frac{79}{6} = 13.17$$

$$\bar{y} = \frac{1,466}{6} = 244.33$$

and so:

$$a = 244.33 - (10.12 \times 13.17) = 111.05$$

Thus the least-squares regression line in this case is $y = 111.05 + 10.12x$.

This line has been plotted on the lower scatter diagram of Figure 7.5 by calculating the coefficients of any two points on the line – for example:

- when $x = 10$, $y = 111.05 + 10.12 \times 10 = 212.25$;
- when $x = 17$, $y = 111.05 + 10.12 \times 17 = 283.09$.

These points should then be plotted on the graph and joined by a straight line.

7.10 Interpreting a and b

You may remember from Chapter 2 that a is the intercept on the y -axis and b is the gradient or slope of the line. Expressed slightly differently, that means that a is the value of y when $x = 0$, and b is the increase in y for each unit increase in x . The b -value of 10.12 tells us that each extra salesperson generates an extra 10.12 sales (on average), while the a -value of 111.05 means that 111.05 units will be sold if no salespeople are used. The latter conclusion may well be nonsensical because $x = 0$ is outside the range of the data, but we return to this later.

It should be noted that, unlike Pearson's correlation coefficient, these calculations do not use Σy^2 , and so no time has been wasted evaluating it. Also, it will be appreciated that calculations such as these can involve potentially large numbers, and so it might be worthwhile to use an available computer package or statistical function on a calculator.



Exercise 7.5

Using the data given in Exercise 7.1 on sales and floor space, find the least-squares regression lines for:

- sales of L against size;
- sales of M against size.

Interpret the values of a and b in your answers.

Solution

- Using the summations calculated in Exercise 7.3:

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(6 \times 15,971.2) - (559 \times 164.1)}{(6 \times 55,089) - 559^2} = \frac{4,095.3}{18,053} = 0.2268$$

$$a = \bar{y} - b\bar{x} = \frac{164.1}{6} - 0.2268 \times \frac{559}{6} = 6.2198$$

Rounding to two decimal places gives the regression line of sales of L against floor space:

$$y = 6.22 + 0.23x$$

(b) Similarly, the regression line of sales of M against floor space can be shown to be:

$$y = -39.05 + 0.87x$$

In part (a), $b = 0.23$ means that if the floor space were increased by one square metre, sales would increase by an average of 0.23 ('000 units), i.e. by 230 units. The a -value of 6.22 gives the absurd result (due to extrapolation) that an outlet with zero floor space would on average sell 6,220 units.

In part (b), the corresponding results are that sales can be expected to increase by 870 units for each extra square metre of floor space and the nonsensical sales of -39,050 units would occur if an outlet had zero size.

7.11 Forecasting

Once the equation of the regression line has been computed, it is a relatively straightforward process to obtain forecasts.

Example 7.E

In the situation of Example 7.D, forecast the number of sales that would be expected next year in regions that employed (a) 14 salespersons; and (b) 25 salespersons.

Solution

As we have the 'best' line representing the dependence of sales on the number of salespersons we shall use it for the forecasts. The values could be read off the line drawn on the scattergraph, but it is more accurate to use the equation of the line.

(a) The regression line is $y = 111.05 + 10.12x$ so, when $x = 14$:

$$y = 111.05 + 10.12 \times 14 = 252.73.$$

Rounding this to a whole number, we are forecasting that 253 units will be sold in a region employing 14 salespersons.

(b) Substituting $x = 25$ into the formula:

$$y = 111.05 + 10.12 \times 25 = 364.05$$

Hence the forecast is sales of 364 units in a region employing 25 salespersons.

Have you noticed that $x = 25$ is well outside the range of the data? What does this tell you about the reliability of the estimated 364? If you are in any doubt about its unreliability, look back to the absurd consequences of putting $x = 0$ in Exercise 7.5.

We give one more example to illustrate the complete process of forecasting from paired samples.

Example 7.F

A company has the following data on its profits and advertising expenditure over the last six years:

<i>Profits</i>	<i>Advertising expenditure</i>
<i>£m</i>	<i>£m</i>
11.3	0.52
12.1	0.61
14.1	0.63
14.6	0.70
15.1	0.70
15.2	0.75

Forecast the profits for next year if an advertising budget of £800,000 is allocated.

Solution

First of all, to justify our assumption that there is a relationship between the two variables, the correlation coefficient should be computed. It is left as an exercise for you to verify that its value is 0.936. This high correlation encourages us to proceed with the regression approach.

As we wish to forecast profits, we shall make this the dependent variable, y , and advertising expenditure the independent variable, x .

The next step is to evaluate the parameters a and b :

x	y	x^2	xy
0.52	11.3	0.2704	5.876
0.61	12.1	0.3721	7.381
0.63	14.1	0.3969	8.883
0.70	14.6	0.4900	10.220
0.70	15.1	0.4900	10.570
0.75	15.2	0.5625	11.400
<u>3.91</u>	<u>82.4</u>	<u>2.5819</u>	<u>54.330</u>

Thus:

$$b = \frac{(6 \times 54.33) - (3.91 \times 82.4)}{(6 \times 2.5819) - 3.91^2} = \frac{3.769}{0.2033} = 18.67$$

$$\bar{x} = \frac{3.91}{6} = 0.652$$

$$\bar{y} = \frac{82.4}{6} = 13.73$$

and so:

$$a = 13.73 - (18.67 \times 0.652) = 1.56$$

The least-squares regression line relating profits to advertising expenditure therefore has equation

$$y = 1.56 + 18.67x.$$

Hence each extra million pounds' advertising generates an extra £18.67 million profits. Also, profits would be £1.56 million without any advertising.

If advertising expenditure is to be £800,000 ($x=0.8$), then:

$$y = 1.56 + 18.67 \times 0.8 = 16.496$$

Rounding this value off to a sensible level of apparent accuracy, we are forecasting profits of £16.5 million next year, if advertising expenditure is £800,000.



Exercise 7.6

The company in Exercise 7.1 is considering opening two further outlets in the city: G, with 85 m² of floor space; and H, with 146 m² of floor space. Use the regression lines calculated in Exercise 7.5 to forecast the sales of L and M at the two possible new outlets.

Solution

Outlet G

$$\begin{aligned}\text{Forecast sales of L} &= 6.2198 + 0.2268 \times 85 = 25.4978 \text{ ('000 units)} \\ &= 25,500 \text{ units (to 3 significant figures)}\end{aligned}$$

$$\begin{aligned}\text{Forecast sales of M} &= -39.052 + 0.8664 \times 85 = 34.592 \text{ ('000 units)} \\ &= 34,600 \text{ units (to 3 s.f.)}\end{aligned}$$

Outlet H

Similarly, the forecast sales of L = 39,300 units and of M = 87,400 units.

7.12 Which variable to denote by y

When calculating the correlation coefficient it does not matter which variable you call x and which y , as the result will be the same either way. However, the regression equation and subsequent forecasts will be totally changed if you change the designation of the variables. This is because the regression line only minimises the sum of squares of the y -errors, and this is only equivalent to minimising x -errors in the case of perfect correlation, that is, when $r = \pm 1$. It is therefore essential that you stop and think about which variable to call y at the start of any exercise on regression.

Variable y is the dependent variable and sometimes it is very clear which that is. However, there are occasions when the dependency could perhaps work either way, in which case the following may be of assistance:

- if you wish to forecast a particular variable, that variable must be denoted by y ;
- if the question asks for the regression of a first variable on or against a second variable, the first variable is denoted by y . This is nothing to do with the order in which the variables are tabulated in the question, which could easily have x first.

For example:

- in an investigation of the downwards trend of sales over time, the independent variable x = time and the dependent variable y = sales;
- in an investigation of the fall in cinema takings as sales of videos increase, possibly the decline in the cinema mirrors the increased use of videos (in which case cinema sales = y), but perhaps it is the other way round. Perhaps cinemas were closed, in a property boom say, and people buy videos because there is no longer a convenient local cinema (in which case video sales = y). However, if the question referred to the regression of cinema sales against video sales or asked for an estimate of cinema sales for a known level of video sales then the doubt would be removed and cinema sales would have to be denoted by y .

CORRELATION AND REGRESSION 7.13 Judging the validity of forecasts

When we have made forecasts, obvious questions to be asked are 'how accurate are they?' and 'what validity do they have?'

Such queries can be addressed in a number of ways.

The importance of using the correlation coefficient as a check on the validity of the assumption of causality has already been stressed. In addition, you should bear in mind the caveats mentioned in earlier parts of the chapter. In particular, is there a hidden third variable in the problem? Thus, in Examples 7.D and 7.E, sales might not depend on the number of salespersons at all, but on the size of the region, as we mentioned when first discussing this problem. If this is the case, then simply increasing the number of salespersons within a region would not in itself increase sales. Even if this is not the case, have we got the causation the right way round? In Example 7.F it might be that, as profits increase, the company feels able to spend more on advertising, so that advertising expenditure depends on profits, contrary to the implicit assumption we made when forecasting profits. If this is the case, increasing the advertising would not necessarily increase profits.

Before leaving the correlation coefficient, we mention another, closely related, measure, the *coefficient of determination*, r^2 . The value of this measure, when expressed as a percentage, shows the percentage of variations in the variables that can be explained by the regression analysis. The remaining variation is due to factors omitted from the analysis.

Example 7.G

Evaluate the coefficients of determination for the situations in (a) Examples 7.D and 7.E; and (b) Example 7.F, and interpret their values.

Solution

- (a) We have seen that $r = 0.948$, so $r^2 = 0.948^2$, or 0.899.

Hence 89.9 per cent of the variations in sales from one region to the next can be explained by the corresponding differences in the number of salespersons. Only about 10 per cent of the differences in regional sales appear to be due to factors other than staffing levels.

- (b) From $r = 0.936$ we get $r^2 = 0.936^2$, or 0.876.

Thus, 87.6 per cent of the variations in profits from one year to the next can be explained by the corresponding variations in advertising expenditure, leaving a surprisingly low 12.4 per cent apparently due to other factors.



Exercise 7.7

Calculate and interpret the coefficients of determination for:

- (a) sales of L and outlet size;
- (b) sales of M and outlet size.

Solution

- (a) For L, $r = 0.984$ so the coefficient of determination = 0.968. This means that 96.8 per cent of the variations in sales of L from one outlet to the next can be explained by the corresponding variations in the amount of floor space.
- (b) For M, $r = 0.974$ so the coefficient of determination = 0.929. This means that 92.9 per cent of the variations in sales of M from one outlet to the next can be explained by the corresponding variations in the amount of floor space.

Consider now the two forecasts made in Example 7.E. The second one is distinctly different from the first, in that we have taken the regression line far beyond the upper data point ($x = 18$ salespersons) to 25 salespersons. The forecast is an *extrapolation* beyond the range of the data. This is an uncertain step to take, as the sales within a region at a certain time must have a ceiling: there must come a point where extra salespersons will generate no further sales. If this point has been passed with 25 salespersons, then our forecast will be an overestimate. The first case, by contrast, is an *interpolation* within the range of the data, and so can be considered more valid. In the same way, the profit forecast of Example 7.F is a slight extrapolation and so should be treated with some caution.

Extreme cases of extrapolation have already been seen when interpreting values of the coefficient a in earlier regression equations. In doing this, we are effectively extrapolating to the x -value of zero, and so we should not be surprised if the result seems implausible.

The approach we have adopted is, of course, a considerable simplification of reality. Profits, sales, and so on, depend on a number of factors, not all of them quantifiable, whereas we have assumed here that they depend on just one other quantitative variable. We have studied only *simple* regression.

There is an extension to the topic, known as *multiple* regression, that enables a variable to be forecast in terms of any number of other variables. This is beyond the scope of this text.

All the forecasts made in this chapter have been for 'next year', whereas the data comes, of course, from the past. There is, therefore, an implicit assumption that conditions that obtained in the past still obtain now and, more importantly, will continue to obtain during the period of the forecast. There is no mathematical way of checking that this is so, but the forecaster will have qualitative knowledge of the particular company and its market, and so will be able to form a judgement. If, for example, a new company was known to be making a big push in the market of the company in Example 7.E, you might doubt the forecast of next year's profit figures.

In conclusion, this section has looked at a number of considerations that should be borne in mind when judging the validity of a regression-based forecast. We shall summarise these in the next section.



Exercise 7.8

Comment on the likely reliability of your forecasts in Exercise 7.6.

Solution

The high values of the coefficients of variation suggest reliable forecasts, with the higher value for sales of L indicating the more reliable forecasts here. Forecasts for outlet G are interpolations and so are more reliable than those for outlet H, which are extrapolations. Finally, there is the possibility of other factors that might affect sales that are not covered in the analysis (advertising budgets at the various outlets, their relatively advantageous or disadvantageous locations, etc.).

7.14 Summary

Pearson's coefficient of linear correlation, r , is

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

where n is the sample size (number of data points). The linear relationship between x and y is strong if r is close to $+1$ or -1 and is weak if r is close to 0.

The *coefficient of determination* is given by r^2 and it gives the percentage of the variations observed in the y values that can be explained by the linear relationship and the corresponding variations in the x values.

When interpreting the value of correlation coefficients or coefficients of determination, care should be taken over:

- how representative (or otherwise) is the sample;
- the sample size;
- whether the correlation is spurious or accidental;
- whether a hidden third variable is present;
- that the interpretation is not generalised beyond the range of the data.

The *least-squares y-on-x regression line* is

$$y = a + bx$$

where:

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

and

$$a = \bar{y} - b\bar{x}$$

In using this line to forecast values of y , the following points/questions must be considered:

- is the correlation coefficient, r (or the coefficient of determination, r^2), large enough to support the assumption that y depends on x ?
- is there a hidden third variable?
- does y depend on x , or is it the case that x depends on y ?
- interpolated forecasts are more reliable than extrapolated ones;
- ensure that the y -on- x line is used only to forecast y ;
- are there any other variables that might affect y ?
- have there been, or are there likely to be, any changes in background circumstances that might invalidate the forecast?
- is the sample sufficiently representative and sufficiently large for reliable results?

7

Revision Questions

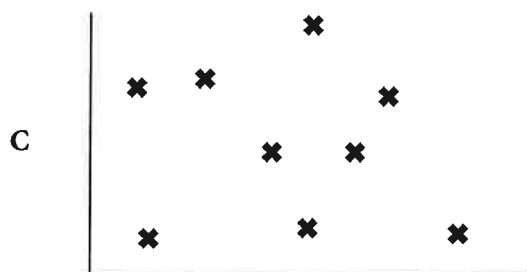
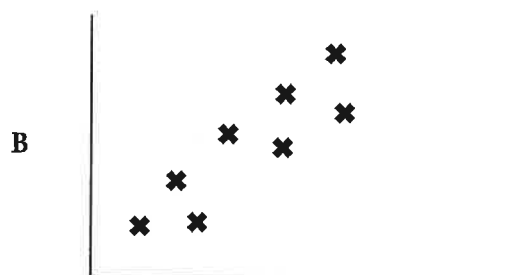
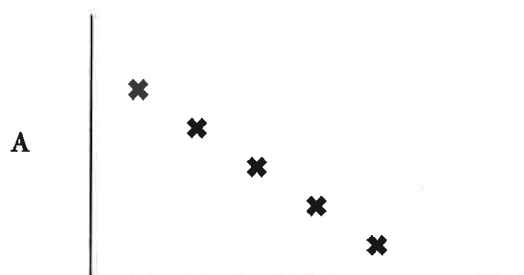
**Question 1** Objective testing selection

Questions 1.1 to 1.10 are standard multiple-choice questions with exactly one correct answer each. Thereafter, the style of question will vary.

- 1.1 The correlation between x and y is 0.85. This means that:
- A x is 85 per cent of y .
 - B y is 85 per cent of x .
 - C there is a strong relationship between x and y .
 - D there is a weak relationship between x and y .
- 1.2 If $\Sigma x = 440$, $\Sigma y = 330$, $\Sigma x^2 = 17,986$, $\Sigma y^2 = 10,366$, $\Sigma xy = 13,467$ and $n = 11$, then the value of r , the coefficient of correlation, to two decimal places, is:
- A 0.98
 - B 0.63
 - C 0.96
 - D 0.59
- 1.3 In a forecasting model based on $y = a + bx$, the intercept is £234. If the value of y is £491 and x is 20, then the value of the slope, to two decimal places, is:
- A -24.55
 - B -12.85
 - C 12.85
 - D 24.85
- 1.4 If the correlation coefficient is 0.8, what is the coefficient of determination?
- A 0.64
 - B 0.89
 - C -0.8
 - D 0.4

- 1.5 If the coefficient of determination is 0.49, which of the following is correct?
- A $y = 0.49x$
 - B $y = a + 0.49x$
 - C 49 per cent of the variation in y can be explained by the corresponding variation in x .
 - D 49 per cent of the variation in x can be explained by the corresponding variation in y .
- 1.6 Find the value of a in a regression equation if $b = 7$, $\Sigma x = 150$, $\Sigma y = 400$ and $n = 10$.
- A 145
 - B -65
 - C $y - 7x$
 - D -650
- 1.7 If the regression equation (in £'000) linking sales (y) and advertising expenditure (x) is given by $y = 5,000 + 10x$, forecast the sales when £100,000 is spent on advertising.
- A £1005,000
 - B £501,000
 - C £4m
 - D £6m
- 1.8 If the regression equation linking costs (£m) to number of units produced ('000s) is $y = 4.3 + 0.5x$, which of the following is correct?
- A For every extra unit produced, costs rise by £500,000.
 - B For every extra 1,000 units produced, costs rise by £500,000.
 - C For every extra 1,000 units produced, costs rise by £4.3m.
 - D For every extra unit produced, costs rise by £4,300.
- 1.9 The prices of the following items are to be ranked prior to the calculation of Spearman's rank correlation coefficient. What is the rank of item G?
- | Item | E | F | G | H | I | J | K | L |
|-------|----|----|----|----|----|----|----|----|
| Price | 18 | 24 | 23 | 23 | 19 | 23 | 19 | 25 |
- A 5
 - B 4
 - C 3
 - D 3.5
- 1.10 All of the following except one will adversely affect the reliability of regression forecasts. Which is the exception?
- A Small sample
 - B Low correlation
 - C Extrapolation
 - D Negative correlation

1.11 Associate each of the following scatter diagrams, A to C, with one of the correlation coefficients, P to R.



Correlation coefficients

P 0

Q -1

R +0.8

1.12 If $n=8$, $\Sigma x=10$, $\Sigma y=800$, $\Sigma xy=1,500$, $\Sigma x^2=20$ and $\Sigma y^2=120,000$, calculate the product moment correlation coefficient, giving your answer correct to three decimal places.

1.13 If the correlation between x and y is 0.9, which of the following is/are true?

- A There is a strong linear relationship between x and y .
- B y increases as x increases.
- C Ninety per cent of the changes in y can be explained by the corresponding changes in x .
- D The slope of the regression line of y on x is positive.

1.14 Over a period of twelve months, in which monthly advertising expenditure ranges from £20,000 to £50,000, the correlation between monthly advertising expenditure and monthly sales is 0.8. Which of the following is/are true on the basis of the information given?

- A Higher sales are caused by higher expenditure on advertising.
- B If advertising expenditure is increased to £100,000, sales will increase.
- C Sixty-four per cent of the changes in sales from one month to the next can be explained by corresponding changes in advertising expenditure.
- D A correlation coefficient derived from 24 months' data would be more reliable than that given above.

1.15 Two wine tasters ranked eight bottles of wine as follows:

Wine	A	B	C	D	E	F	G	H
Taster X	3	7	1	8	5	2	4	6
Taster Y	3	8	2	7	4	1	5	6

Find Spearman's rank correlation coefficient for this data, giving your answer to three decimal places.

1.16 Place the following values in rank order, with the smallest being ranked 1.

Values 3.21 3.49 3.99 4.05 3.49 4.49 4.99 4.05 3.49

1.17 If $n = 10$, $\Sigma x = 90$, $\Sigma y = 1,500$, $\Sigma x^2 = 1,000$ and $\Sigma xy = 20,000$, calculate the value of b in the regression line $y = a + bx$, giving your answer correct to three significant figures.

1.18 In the calculation of the regression equation $y = a + bx$ using ten pairs of x - and y -values, $\Sigma x = 80$, $\Sigma y = 500$ and $b = -1.59$. Calculate the value of a correct to three significant figures.

1.19 The regression equation $y = 50 - 2x$ has been obtained from fifteen pairs of x - and y -values, with the x -values ranging from 0 to 20. Which of the following is/are correct?

- A When $x = 0$, y is estimated to be 25.
- B y decreases by 2 whenever x increases by 1.
- C The equation cannot be relied upon for x -values bigger than 20.
- D The correlation between x and y must be negative.

1.20 The correlation coefficient for ten pairs of x - and y -values, with x ranging from £500 to £700, is calculated to be 0.79, and the regression equation is $y = 620 + 4.3x$. Which of the following is/are true?

- A When $x = £600$, the estimate of $y = 3,200$.
- B When $x = £550$, the estimate of y from the regression equation is likely to be reliable.
- C When $x = 0$, the estimate of y from the regression equation is likely to be reliable.
- D When x increases by £1, y increases by 0.79.



Question 2 Scatter diagram; line of best fit

An ice-cream supplier has recorded some sales data that he believes shows a relationship between temperature and sales. The results shown below are for ten sample days in the summer:

Temperature, °C	Cartons sold
X	y
13	10
16	11
17	14
19	15
20	16
21	19
23	24
26	25
27	26
28	27

- 2.1 Using the intermediate totals given below, calculate the coefficient of correlation giving your answer correct to 2 d.p.

$$\begin{array}{ll} \Sigma x = 210 & \Sigma y = 187 \\ \Sigma x^2 = 4,634 & \Sigma y^2 = 3,865 \\ \Sigma xy = 4,208 & \end{array}$$

Answer:

- 2.2 If the correlation coefficient was 0.95 calculate the coefficient of determination, giving your answer to the nearest whole number.

Answer:

- 2.3 If the correlation coefficient was 0.95, which of the following statements would be correct?

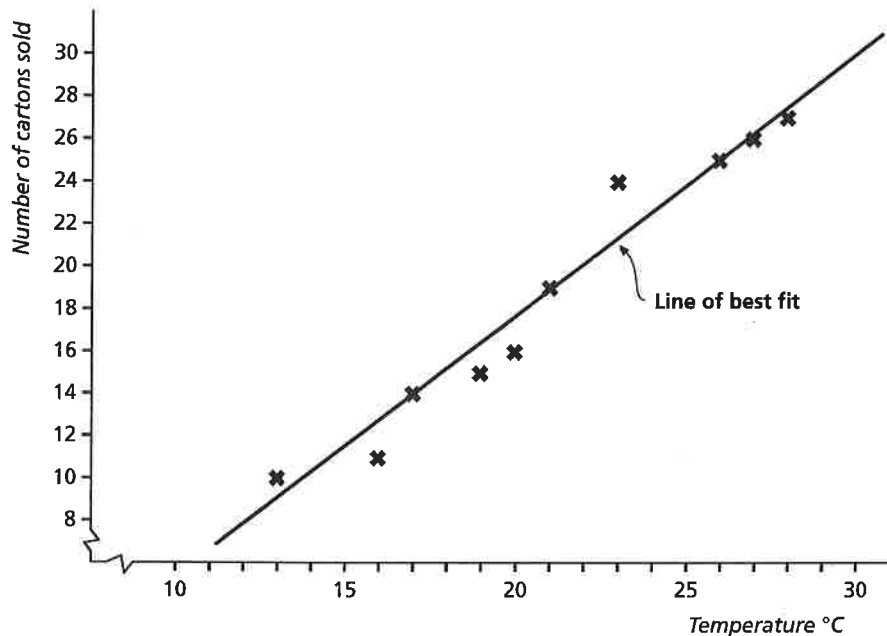
- A The positive sign tells us that there is a strong relationship between temperature and sales.
- B The positive sign tells us that as temperature rises, so do sales.
- C The value of the correlation coefficient tells us that there is a strong linear relationship between temperature and sales.
- D The value of the correlation coefficient tells us that for each increase of 1 degree in temperature, sales increase by 0.95 cartons.
- E The value of the correlation coefficient tells us that for each decrease of 1 degree in the temperature, sales decrease by 5%.
- F The value of correlation coefficient tells us that high temperatures cause high sales.

- 2.4 If the coefficient of determination was 85%, which of the following statements would be correct?

- G When temperature increases by 1 degree, sales increase by 85%.
- H When temperature increases by 1 degree, sales increase by 15%.

- I On 85% of days it is possible to accurately predict sales if an accurate prediction of temperature exists.
- J 85% of the changes in sales from one day to the next can be explained by corresponding changes in temperature.

2.5 The following graph displays the data. What type of graph is it?



- K Scattergram
- L Histogram
- M Pictogram
- N Ogive

2.6 A A freehand line of best fit has been fitted to the graph of the data. Estimate the likely sales when the temperature is 15°C, giving your answer to the nearest whole number.

Answer:

B Estimate the likely sales when the temperature is 30°C, giving your answer to the nearest whole number.

Answer:

2.7 Which of the following statements about the reliability of the estimates made in 2.6 are correct, assuming that the correlation is 0.95.

- O The estimate for a temperature of 15 degrees should be reliable because it involves interpolation.
- P Both estimates are less reliable than they otherwise would be because the sample is small.
- Q The estimate for 30 degrees should be reliable because it involves extrapolation.
- R Both estimates are more reliable than they would otherwise be because the correlation is high.



Question 3 Least-squares analysis

A travel agency has kept records of the number of holidays booked and the number of complaints received over the past ten years. The data is as follows:

Year	1	2	3	4	5	6	7	8	9	10
Number of holidays booked	246	192	221	385	416	279	343	582	610	674
Number of complaints received	94	80	106	183	225	162	191	252	291	310

The agency suspects there is a relationship between the number of bookings and the volume of complaints and wishes to have some method of estimating the number of complaints, given the volume of bookings.

- 3.1 Denoting number of holidays by X and number of complaints by Y , the following summations are given:

$$\Sigma X = 3,948, \Sigma Y = 1,894, \Sigma X^2 = 1,828,092, \Sigma Y^2 = 417,596, \Sigma XY = 869,790.$$

Calculate the value of the regression coefficient ' b ', giving your answer correct to 3 d.p.

Answer:

- 3.2 If the value of b is taken to be 0.4 calculate the value of the regression coefficient ' a ', giving your answer correct to 2 d.p.

Answer:

- 3.3 If the regression equation was $y = 31 + 0.4x$ forecast the likely number of complaints if 750 holidays are booked, giving your answer to the nearest whole number.

Answer:

- 3.4 Which of the following methods could be used to check whether there is in fact a linear relationship between the variables.

- A Scatter diagram
- B Times series analysis
- C Coefficient of variation
- D Regression analysis
- E Correlation coefficient

- 3.5 Which of the following comments about the likely reliability of the estimate of complaints arising from 750 holidays is/are correct?

- F The estimate is likely to be reliable because the value of ' a ' is positive.
- G The estimate is likely to be reliable because it lies outside the range of the data.
- H The estimate is likely to be unreliable because the sample is small.
- I The estimate is not likely to be reliable because the value of ' b ' is not close to 1.
- J The estimate is likely to be unreliable because it was obtained by extrapolation.



Question 4 Regression analysis

A company has a fleet of vehicles and is trying to predict the annual maintenance costs per vehicle. The following data has been supplied for a sample of vehicles:

Vehicle no	Age in years	Maintenance cost per annum £ × 10
	<i>x</i>	<i>y</i>
1	2	60
2	8	132
3	6	100
4	8	120
5	10	150
6	4	84
7	4	90
8	2	68
9	6	104
10	10	140

4.1 In the following tables, calculate the figures denoted by letters A to E.

<i>x</i>	<i>y</i>	<i>x</i> ²	<i>xy</i>
2	60	B	D
8	132	64	1,056
6	100	36	600
8	120	64	960
10	150	100	1,500
4	84	16	336
4	90	16	360
2	68	4	136
6	104	36	624
10	140	100	1,400
<u>A</u>	<u>1,048</u>	<u>C</u>	<u>E</u>

Answers:

- A
- B
- C
- D
- E

4.2 Calculate the regression coefficient denoted by 'b', giving your answer to 2 d.p.

Answer:

- 4.3 If the value of b was 10, calculate the regression coefficient denoted by ' a ', giving your answer to 1 d.p.

Answer:

- 4.4 If $b = 10$ and $a = 44$ state the regression equation where y is annual cost (in £10) and x is age in years.

Answer:

- 4.5 Estimate (to the nearest £) the annual maintenance cost of a 12-year-old vehicle.

Answer:

- 4.6 Does this estimate involve interpolation (F), extrapolation (G) or neither (H)?

Answer:

- 4.7 Does the answer in 4.6 mean that the estimate is (relatively) reliable (I), unreliable (J) or neither (K)?

Answer:

? Question 5

A company is building a model in order to forecast total costs based on the level of output. The following data is available for last year:

Month	Output '000 units [x]	Costs £000 [y]
January	16	170
February	20	240
March	23	260
April	25	300
May	25	280
June	19	230
July	16	200
August	12	160
September	19	240
October	25	290
November	28	350
December	12	200

- 5.1 If output is denoted by X and costs by Y , you are given that $\Sigma X = 240$, $\Sigma Y = 2,920$, $\Sigma XY = 61,500$, $\Sigma X^2 = 5,110$ and $\Sigma Y^2 = 745,200$. Calculate the correlation coefficient between output and costs, giving your answer to 3 d.p.

Answer:



- 5.2 If the correlation coefficient was 0.9, which of the following comments are correct?
- A The correlation coefficient shows that there is a strong linear relationship between output and costs.
 - B The correlation coefficient shows that high output causes high costs.
 - C The correlation coefficient shows that costs rise as output rises.
 - D Costs rise by 0.9 in £1000 for every extra 1000 units of output.
 - E The high value of the correlation coefficient means that estimates made using regression are likely to be reliable.
- 5.3. If the regression equation linking output and costs is $Y = 43 + 10X$, which of the following comments is/are correct?
- F For every extra unit produced, costs will rise by £43.
 - G Even with zero output there will be costs of £43,000.
 - H For every extra 1000 units produced, costs will rise on average by £10,000.
 - I For every extra unit produced, costs are likely to increase by £53.
 - J When 1000 units are produced, costs are likely to be £10,043.

7

Solutions to Revision Questions



Solution 1

1.1 Answer: C

Correlation coefficients measure the strength of the linear relationship between two variables. They vary numerically between 0 and 1, being weak when close to 0 and strong when close to 1.

1.2 Answer: B

$$r = \frac{(11 \times 13,467) - (440 \times 330)}{\sqrt{((11 \times 17,986) - 440^2)((11 \times 10,366) - 330^2)}} = \frac{2,397}{\sqrt{4,246 \times 5,126}} = 0.63$$

1.3 Answer: C

$$\begin{aligned} Y &= a + bX \\ 491 &= 234 + 20b \\ 257 &= 20b \\ 12.85 &= b \end{aligned}$$

1.4 Answer: A

The coefficient of determination is given by squaring the correlation coefficient. Often it is also multiplied by 100.

1.5 Answer: C

The coefficient of determination gives the percentage of the variation in y which can be explained by the regression relationship with x . Answers A and B are confusing this with the actual regression equation, while answer D has x and y the wrong way round.

1.6 Answer: B

$$a = \frac{\Sigma y}{n} - b \frac{\Sigma x}{n} = 40 - (7 \times 15) = -65$$

Solution A has added instead of subtracting and D has failed to divide by n . Solution C correctly states that, since $y = a + bx$, then $a = y - bx$. However, this is an equation satisfied by a and is not the value of a .

1.7 Answer: D

$y = 5,000 + 10x$, and $x = 100$ when advertising is £100,000. Hence $y = 5,000 + 10 \times 100 = 6,000$ (in £'000). Hence sales forecast is £6m.

A is wrong because $x = 100,000$ has been used and the units of y ignored.

In B, $5,000 + 10$ has been calculated before multiplication by 100 and in C the $10x$ has been wrongly subtracted.

1.8 Answer: B

In the equation $y = a + bx$, if x increases by one unit, y will increase by b units. In this case if x increases by 1 unit, y increases by 0.5 units, which translates into production increasing by 1,000 units and costs by £500,000.

All the other answers have either confused a and b or confused the units of x and y .

1.9 Answer: A

Item	E	F	G	H	I	J	K	L
Price	18	24	23	23	19	23	19	25
Rank	1	7	5	5	2.5	5	2.5	8

The two 19s occupy ranks 2 and 3 with an average of 2.5, and the three 23s occupy ranks 4, 5 and 6 with an average of 5. It is essential to count both the 19s and the fact that item G happens to be the first of the 23s listed doesn't give it a lower rank. Answer D seems to be misled by item G's position as third in the list but adjacent to another 23.

1.10 Answer: D

It is the strength of the correlation but not its sign that influences the reliability of regression forecasts. Correlation can be negative but still very strong so D is the exception. Small samples, low correlation and extrapolation all tend to give unreliable forecasts.

1.11 Correlation coefficient Scatter diagram

P (0)	C
Q (-1)	A
R (+0.8)	B

$$1.12 \quad r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)\}}}$$

$$= \frac{8 \times 1,500 - 10 \times 800}{\sqrt{\{(8 \times 20 - 10^2)(8 \times 120,000 - 800^2)\}}}$$

$$= \frac{4,000}{\sqrt{(60 \times 320,000)}} = 0.913$$

- 1.13 A True: the correlation is close to 1 in value.
 B True: the correlation is positive.
 C Untrue: the correct percentage would be 84.
 D True: the correlation is positive.

- 1.14 A Untrue on the basis of this information. Causation cannot be deduced from high correlation.
 B Untrue on the basis of this information. We cannot be sure that the positive correlation will continue for advertising greater than £50,000.
 C True.
 D True.

1.15	A	B	C	D	E	F	G	H
d	0	-1	-1	1	1	1	-1	0
d^2	0	1	1	1	1	1	1	0

$$\Sigma d^2 = 6; n = 8$$

$$\begin{aligned}\text{Rank correlation} &= 1 - \frac{6 \times \Sigma d^2}{n(n^2 - 1)} \\ &= 1 - \frac{6 \times 6}{8 \times 63} = 0.929\end{aligned}$$

1.16 Values	3.21	3.49	3.99	4.05	3.49	4.49	4.99	4.05	3c.49
Ranks	1	3	5	6.5	3	8	9	6.5	3

$$\begin{aligned}1.17 \quad b &= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \\ &= \frac{10 \times 20,000 - 90 \times 1,500}{10 \times 1,000 - 90^2} \\ &= \frac{65,000}{1,900} = 34.2\end{aligned}$$

$$\begin{aligned}1.18 \quad a &= \frac{\Sigma y}{n} - b \frac{\Sigma x}{n} \\ &= \frac{500}{10} - (-1.59) \times \frac{80}{10} = 50 + 12.72 = 62.7 \text{ to three s.f.}\end{aligned}$$

- 1.19 A Incorrect: when $x = 0$, $y = 50 - 0 = 50$.
 B Correct.
 C Correct.
 D Correct.
- 1.20 A Correct.
 B Correct.
 C Incorrect: $x = 0$ is outside the range of the data.
 D Incorrect: when x increases by 1, y increases by 4.3 from the regression equation.



Solution 2

2.1

$$\begin{aligned}r &= \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{((n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2))}} \\ &= \frac{(10 \times 4,208) - (210 \times 187)}{\sqrt{((10 \times 4,634) - (210)^2) \times ((10 \times 3,865) - (187)^2)}} \\ &= \frac{2,810}{\sqrt{(2,240 \times 3,681)}} = 0.98\end{aligned}$$

Answer: 0.98

- 2.2 Coefficient of determination = $(r^2) * 100 = (0.95^2) * 100 = 90$ per cent.

Answer: 90

- 2.3 The positive sign tells us that as temperature rises, so do sales but it tells us nothing about the strength of the relationship. So B is correct and A is incorrect.

The value of the correlation coefficient tells us that there is a strong linear relationship between temperature and sales (C) but it cannot prove cause and effect, so F is wrong. Equally it doesn't enable us to estimate likely changes in sales corresponding to known changes in temperature and hence D and E are both incorrect.

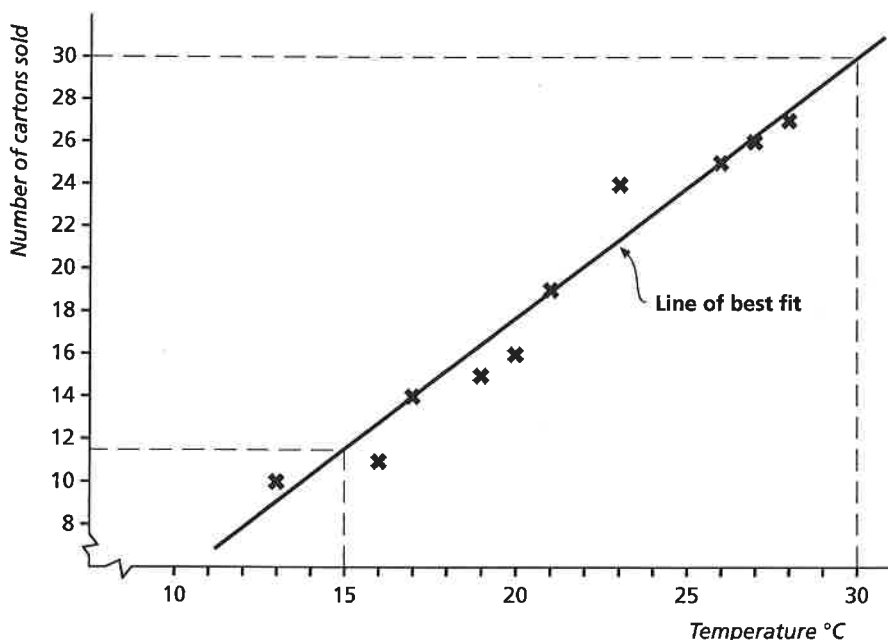
Answers: B, C

- 2.4 G, H and I are all incorrect because predictions cannot be made on the basis of the coefficient of determination.

Answer: J

85% of the changes in sales from one day to the next can be explained by corresponding changes in temperature.

- 2.5 *Cartons sold against daily temperature*



Answer: K Scattergram

- 2.6 A Answer: 11
B Answer: 30

- 2.7 Estimation within the range of the data is called interpolation and, all other things being equal, tends to give reliable estimates. Extrapolation – estimating outside the range of the data – is not reliable, although in this instance a temperature of 30°C is only just outside the range and would not therefore constitute so much of a problem. The high correlation coefficient will make both estimates more reliable, while the small sample of ten points will reduce the reliability of both.

The estimate for 30 degrees is not reliable because it involves extrapolation and the fact that correlation is high does not strictly speaking render it more reliable because it lies outside the range of the data for which the correlation has been calculated. The small sample reduces the reliability of all estimates.

Answers: O and P



Solution 3

$$3.1 \quad b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{(10 \times 869,790) - (3,948 \times 1,894)}{(10 \times 1,828,092) - 3,948^2} = \frac{1,220,388}{2,694,216} = 0.453$$

Answer: 0.453

$$3.2 \quad a = \sum Y/n - b\sum X/n = 1,894/10 - 0.4 \times 3,948/10 = 31.48$$

Answer: 31.48

3.3 The number of complaints for 750 holidays booked:

$$y = 31 + 0.4 \times 750 = 331$$

Answer: 331

3.4 Time series analysis and the coefficient of variation are not related to this question at all. Regression analysis is concerned with finding the best possible line to fit the data when it has been established that there is in fact an approximately linear relationship. Hence B, C and D are incorrect.

Answers: A and E

3.5 The sample has only 10 points and the estimate was obtained by extrapolation. For both these reasons it is not likely to be reliable.

Answers: H and J



Solution 4

4.1 A is the total of the x column; B is given by x^2 when $x = 2$, i.e. by $2^2 = 4$; C is then the total of the x^2 column; D is given by $x \times y$, i.e. by $2 \times 60 = 120$ and E is the total of the xy column.

Answers:

- A 60
- B 4
- C 440
- D 120
- E 7,092

$$4.2 \quad b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{10 \times 7,092 - 60 \times 1,048}{10 \times 440 - 60^2} = \frac{8,040}{800} = 10.05$$

Answer: 10.05

$$4.3 \quad a = \sum y/n - b\sum x/n = 1,048/10 - 10 \times 60/10 = 44.8$$

Answer: 44.8

$$4.4 \quad \text{Answer: } y = 44 + 10x$$

4.5 A twelve-year-old vehicle would have an estimated maintenance cost of:

$$44 + 10 \times 12 = 164 \text{ in } \pounds 10$$

Answer: 1,640

4.6 12 years is outside the range of the data, so extrapolation is being used. This means that the estimate will be relatively unreliable.

Answer: G

4.7 Answer: J



Solution 5

$$5.1 \quad r = \frac{12 \times 61,500 - 240 \times 2,920}{\sqrt{(12 \times 5,110 - 240^2)(12 \times 745,200 - 2,920^2)}} = \frac{37,200}{\sqrt{(3,720)(416,000)}}$$

$$= \frac{37,200}{39,338.53} = 0.946$$

Answer: 0.946

5.2 There is a strong linear relationship because the value of the correlation coefficient is close to 1 and the positive sign means that costs rise as output rises. The high value of the correlation coefficient does mean that estimates made using regression are likely to be reliable. However correlation cannot prove causation and it tells us nothing about the changes in costs resulting from particular levels of output. So B and D are incorrect.

Answer: A, C and E

5.3 It is important to remember that X is 1000 units of output and Y is £1000 of costs. When X = 0, Y = 43 so costs are £43,000 and G is correct. When X increases by 1 (i.e. 1,000 units), Y increases by 10 (i.e. £10,000 in costs) so H is also correct. All the other statements are incorrect.

Answers: G and H

