**3 CORRELATION AND**

**REGRESSION**

## Objectives

#### After studying this chapter you should

* be able to investigate the strength and direction of a

relationship between two variables by collecting measurements and using suitable statistical analysis;

* be able to evaluate and interpret the product moment

correlation coefficient and Spearman's correlation coefficient;

* be able to find the equations of regression lines and use them where appropriate.
	1. **Introduction**

**Is a child's height at two years old related to her later adult height? Is it true that people aged over twenty have slower**

**reaction times than those under twenty? Does a connection exist between a person's weight and the size of his feet?**

In this chapter you will see how to quantify answers to questions of the type above, based on observed data.

* 1. **Ideas for data collection**

Undertake at least one of the three activities below. You will need your data for further analysis later in this chapter.

**Covariance**

An attempt to quantify the tendency to go from bottom left to top right is to evaluate the expression

*i* 1

*xi*  *x* *yi*  *y* 

*n*

∑

1

*sxy*  *n*

which is known as the **covariance** and denoted by cov *X*, *Y*  or

#### *sxy* . For shorthand it is normally written as

1  *x*  *x* *y*  *y* 

##### n

where the summation over *i* is assumed.

The points in the top right have *x* and *y* values greater than *x*

and *y* respectively, so *x*  *x* and *y*  *y* are both positive and so is the product *x*  *x* *y*  *y* .

#### Those in the bottom left have values less than *x* and *y* , so *x*  *x* and *y*  *y* are both negative and again the product *x*  *x* *y*  *y*  is positive.

Points in the other two areas have one of *x*  *x* and *y*  *y*

positive and the other negative, so *x*  *x* *y*  *y*  is negative.

The 1

##### n

factor accounts for the fact that the number of points will

affect the value of the covariance.

In the example above, most of the points give positive values of

*x*  *x* *y*  *y* .

#### There is another form of the expression for covariance which is easier to use in calculations.

1 *x*  *x* *y*  *y*   1 *xy*  *xy*  *xy*  *xy* 

##### n n

 1  *xy*  *xy*  *xy*  *xy* 

##### n

 1  *xy*  *x*  *y*  *y*  *x*  *n xy* 

##### n

 1  *xy*  *xny*  *ynx*  *nxy* 

since *y*   *y* , *x*   *x*

#### Thus

##### n n n

 1  *xy*  *nxy*  .

##### n

*n*

*n*

1  *x*  *x* *y*  *y*   1  *xy*  *xy*

The right hand side is quicker to evaluate. For the example on page 216, this form of the expression is usually used when

calculating covariance.

*sxy*

 1  *xy*  17  30

10

 1  5313  510

10

 21.3

(  *xy* is a function available on calculators with LR mode.)

The fact that *sxy*  0 indicates that the points follow a trend with a positive slope. The size of the number, however, conveys little as it can easily be altered by a change of scale.

The following examples show this.

**Example**

Find the covariance for the following data.

(a) Height (m) *x* 1.60 1.64 1.71

Weight (kg) *y* 53 57 60

(b)

Height (cm) *x* 160 164 171

Weight (kg) *y* 53 57 60

**Solution**

(a) *s*  1  280.88  170  4.95

*xy* 3 3 3

 0.126˙

(b) *s*

 1  28088  170  495

*xy* 3 3 3

 12. 6˙

You can, of course, get quite different values by measuring in pounds and inches or kg and feet, etc. They will all be positive but their sizes will not convey useful information.

* 1. **Pearson's product moment correlation coefficient**

Dividing *x*  *x*  by the standard deviation *sx* gives the distance of each *x* value above or below the mean as so many standard

deviations. For the example on height and weight above, the standard deviations in m and cm are related, with the second being one hundred times the first, so

*x*  *x sx*

#### will give the same answer regardless of the units or scale involved. The quantity

1  *x*  *x*   *y*  *y* 

*n*  

*sx*  

*sy* 

#### can therefore be relied on to produce a value with more meaning than the covariance.

Since



1  *x*  *x*   *y*  *y* 



1 *xy*  *xy*

*n*  

*sx*  

*n*

*sy* 

*sxsy*

and the latter is easier to evaluate, **Pearson's product moment correlation coefficient** is often given as

1 *xy*  *xy r*  *n*

*sxsy*

where *sx*

 1 *x*2  *x* 2 and *s*  .

*n y*

1 *y*2  *y* 2

*n*

#### (Note that *r* is a function given on calculators with LR mode.) Returning to the example in Section 12.2:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
| 20 | 23 | 8 | 29 | 14 | 11 | 11 | 20 | 17 | 17 |
| 30 | 35 | 21 | 33 | 33 | 26 | 22 | 31 | 33 | 36 |

###### Pupil

Maths mark (out of 30)

*x*

Physics mark (out of 40)

*y*

 1  5313  17  30

*r*  10

*sx*  *sy*

*s*  1  3250  172  36  6

*x* 10

#### *s*  1  9250  302  25  5

*y* 10

 *r*  531.3  510

6  5

 0. 71

The value of *r* gives a measure of how close the points are to lying on a straight line. It is always true that

1  *r*  1

and *r*  1 indicates that all the points lie exactly on a straight line with positive gradient, while *r*  1 gives the same information with a line having negative gradient, and *r*  0 tells us that there is no connection at all between the two sets of data.

The sketches opposite indicate these and in between cases.

(Note that *sxy* is not a calculator key, but its value may be checked by *r*  *sx*  *sy* which are all available.)

**The significance of** *r*

With only **two** pairs of values it is unlikely that they will lie on the same horizontal or vertical line, giving a correlation

coefficient of zero but any other arrangement will produce a

value of *r* equal to plus or minus one, depending on whether the line through them has a positive or negative gradient. With **six** points, however, the fact that they lie on, or close to, a straight line becomes much more significant.

The following table, showing critical values at 5% significance level, gives some indication of how likely some values of the correlation coefficient are. For example, for *n*  5 , *r*  0.878

means that there is only a 5% chance of getting a result of 0.878 or greater if there is **no** correlation between the variables. Such a value, therefore, indicates the likely existence of a relationship between the variables.

|  |  |
| --- | --- |
| **(no.of pairs)***n* | *r* |
| 3 | 0.997 |
| 4 | 0.950 |
| 5 | 0.878 |
| 6 | 0.811 |
| 7 | 0.755 |
| 8 | 0.707 |
| 9 | 0.666 |
| 10 | 0.632 |

More detailed tables of critical values are available for a range of significant levels and values of *n*. Their calculation relies on the data being drawn from joint normal distributions, so using

them in other circumstances cannot provide an accurate assessment of significance.

**Example**

A group of twelve children participated in a psychological study designed to assess the relationship, if any, between age, *x* years, and average total sleep time (ATST), *y* minutes. To obtain a

measure for ATST, recordings were taken on each child on five consecutive nights and then averaged. The results obtained are shown in the table.

|  |  |  |
| --- | --- | --- |
| **Child** | **Age (***x* years) | **ATST** (*y* minutes) |
| A | 4.4 | 586 |
| B | 6.7 | 565 |
| C | 10.5 | 515 |
| D | 9.6 | 532 |
| E | 12.4 | 478 |
| F | 5.5 | 560 |
| G | 11.1 | 493 |
| H | 8.6 | 533 |
| I | 14.0 | 575 |
| J | 10.1 | 490 |
| K | 7.2 | 530 |
| L | 7.9 | 515 |

 *x*  108  *y*  6372  *x*2  1060.1  *y*2  3396942  *xy*  56825. 4

Calculate the value of the product moment correlation coefficient between *x* and *y*. Assess the statistical significance of your value and interpret your results.

**Solution**

1. Use the formula

*sxy*

 1  *xy*  *xy n*

#### when *x*  108  9 and *y*  6372  531 .

12 12

Thus *sxy*

 1 56825. 4  9  531  43. 55

12

Also *sx* 

12

 1  1060.1  92

 2.7096

*s*  1  3396942  5312  33. 4290

*y* 12

#### Hence *r*    43.55   0.481

2.7096  33. 4290

This indicates weak negative correlation. But to apply a

significance test, the null and alternative hypotheses need to be defined:

*H*0 : *r*  0

*H*1: *r*  0

#### significance level : 5% (two tailed).

Using the table of critical values in the Appendix, for *n*  12 ,

*rcrit*   0.576

#### That is, the critical region where *H*0 is rejected is *r*   0. 576

and *r*  0. 576 .

Since *r*   0.481, there is insufficient evidence to reject the null hypothesis.

**Limitations of correlation**

You should note that

1. *r* is a measure of **linear** relationship only. There may be an exact connection between the two variables but if it is not a straight line *r* is no help. It is well worth studying the

scatter diagram carefully to see if a non-linear relationship may exist. Perhaps studying *x* and ln *y* may provide an

answer but this is only one possibility.

1. Correlation does not imply **causality**. A survey of pupils in a primary school may well show that there is a strong

correlation between those with the biggest left feet and those who are best at mental arithmetic. However it is

P

unlikely that a policy of 'left foot stretching' will lead to

improved scores. It is possible that the oldest children have the biggest left feet and are also best at mental arithmetic.

1. An unusual or freak result may have a strong effect on the value of *r*. What value of *r* would you expect if point P

were omitted in the scatter diagram opposite?

# Spearman's rank correlation coefficient

#### Two judges at a fete placed the ten entries for the 'best fruit cakes' competition in order as follows (1 denotes first, etc.)

###### Entry

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
| 2 | 9 | 1 | 3 | 10 | 4 | 6 | 8 | 5 | 7 |
| 6 | 9 | 2 | 1 | 8 | 4 | 3 | 10 | 7 | 5 |

Judge 1 (*x*)

Judge y

 NO actual marks like 73/100 have been awarded in this case where only ranks exist.

**Is there a linear relationship between the rankings produced by the two judges?**

*y*  5.5

*y*

#### Spearman's rank correlation coefficient answers this question by 10

#### simply using the ranks as data and in the product moment 8

#### coerrelation coefficient, *r*, and denoting it *rs* . Again a scatter

#### diagram may be drawn and the presence of the points plotted in, or 6

#### very near, the top right and bottom left areas indicates a positive 4

#### correlation. 2

*x*  5.5

#### Spearman's rank correlation coefficient,

00 2 4 6 8 10 *x*

*rs* 

 10

 1  *xy*  *xy sx sy*

#### where *x*  *y*  55  5. 5

10

*sx*  *sy*

 385  5. 52  8. 25

10

and *xy*  2  6  9  9  K  7  5  362

 *rs* 

 1  362  5.52

8.25 8.25

 10

 36. 2  30. 25

8. 25

 0. 721

(The significance tables for *r* should certainly not be used here as the ranks definitely do not come from normal distributions.)

It can be shown that, when there are no tied ranks,

1 *xy*  *xy* 2

 *n*  1  6*d*

#### and so

*sxsy*

*n**n*2  1

#### where *d*  *x*  *y* , is the difference in ranking.

6 *d*2

*n**n*2  1

*s*

*r*  1 

For the example just considered

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Entry** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** |
| Judge 1 | 2 | 9 | 1 | 3 | 10 | 4 | 6 | 8 | 5 | 7 |
| Judge 2 | 6 | 9 | 2 | 1 | 8 | 4 | 3 | 10 | 7 | 5 |
| *d* | 4 | 0 | 1 | 2 | 2 | 0 | 3 | 2 | 2 | 2 |
| *d* 2 | 16 | 0 | 1 | 4 | 4 | 0 | 9 | 4 | 4 | 4 |

So *d* 2  16  0  1  K  4  46

 *rs*

 1  6  46

10100  1

 1  6  46

10  99

 119

165

 0. 721 to 3 decimal places.

As with the product moment correlation coefficient, Spearman's correlation coefficient also obeys

1  *rs*  1

#### where *r*  1 corresponds to perfect positive correlation and

*r*  1 to perfect negative correlation.

The definition of the formula from the product moment

correlation coefficient will not be given here but you will see in the following Activity how it can be deduced.

**Significance of** *r*

*s*

#### If the tables of significance for *r* cannot be used here, you can still assess the importance of the value by noting that the formula

*r*  1  6*d* 2

*s n**n*2  1

#### contains the term *d* 2 . Tables giving the critical values of *rs* for various values of *n* are available.

So at 5% significance level, the hypotheses are defined by

*H*0 : *rs*  0

#### *H*1 : *rs*  0 (two tailed) and, with *n*  10 , the tables show that

Note: *rs*  0. 6485 means

*rs*   0. 6485 or *r*  0. 6485.

*p**rs*  0.6485  0.05

#### So for a two tailed test, you should reject *H*0 since in the example on page 226, *rs*  0.721, and accept *H*1 , the alternative

hypothesis, which says that there is significant correlation.

You can test for positive correlation, by using the hypothesis

*H*0 : *rs*  0

#### *H*1 : *rs*  0 (one tailed) At 5% level, and with *n*  10 as before,

*p**rs*  0.5636  0.05

#### and since 0.721 > 0.5636, again *H*0 is rejected. You accept the

alternative hypothesis that there is significant positive correlation.

* 1. **Linear regression**

In linear regression you start by looking at a set of points to see if there is a relationship between them and if there is you

proceed to establish it in such a way that further points may be deduced from it with the minimum possible error. That is, start with points, proceed to a line and regress to points again.

Here are some results for the elastic band experiment suggested in Activity 6.

Mass g *(x)* 50 100 150 200 250 300 350 400

Length mm ( *y)* 37 48 60 71 80 90 102 109

In the diagram opposite, the points lie very close to a straight line and the value of *r* is 0.999.

**Activity 9**

Find the value of

1. *r*, the product moment correlation coefficient.
2. *rs* , Spearman's rank correlation coefficient. Comment on their values.

*y*

100

80

60

40

20

00 100 200 300 400 *x*

#### Having decided that the points follow a straight line, with some small variations due to errors in measurement, changes in the

environment etc, the problem is to find the line which best fits the data.

It may seem natural to try to find the line so that the points' distances from it have as small a total as possible. However,

since the line will need to produce values of *y* for given values of *x* (or vice versa) it is more sensible to seek to produce a line

so that any distances in the *y* direction, and therefore any errors in predicting *y* given *x*, should be a minimum.

If the line is to be used to predict values of *y* based on known values of *x* it is called the '*y* on *x*' line and its equation is

determined by making *d*12  *d*22  K   *d* 2 as small as possible. The equation of this line can be shown to be

*x*4, *y*4

*sx* 2

*y*  *y*  *sxy* *x*  *x* 

#### and for this line *d* 2  *ns* 2 1  *r*2 . You will notice that when

*d*4

*x* 1, *y*1

*d*1

*d*3

*d*2

*x*2, *y*2

*x* , *y* 

3 3

*y*

*r*  1 (i.e. the points lie exactly on a straight line) then

*d* 2  0 as would be expected. The procedure used to obtain the equation is called the **method of least squares** and the '*d*'s are

#### often referred to as the **residuals**. The gradient is called the

**regression coefficient**.

For the elastic band example,

*x*  1800  225, *y*  597  74. 625

8 8

*sxy*

 156150  225  74. 625  2728.125

8

*s* 2  510000  2252  13125

*x* 8

 *y*  74. 625  2728.125 *x*  225

13125



*y*  0. 208*x*  27.857

The values of 0.208 and 27.857 represent the gradient of the line and its intercept on the *y*-axis and are available directly from a

calculator with LR mode. The gradient has units mm/g and tells us how much extension would be caused by the addition of 1

extra gram to the suspended mass. This line can now be used to find values of *y* given values of *x*.

**Example**

What length would you expect the elastic band to be if a weight of

(a) 375 g (b) 1 kg was suspended by it?

**Solution**

(a) *y*ˆ  0. 208  375  27.857

 105.9 mm

(The ^ above the *y* indicates that this is an estimate, however

accurate. Calculators with LR mode usually have a *y*ˆ function giving the answer directly.)

(b) *y*ˆ  0. 208  1000  27.857

 235. 9

The first of these answers is an example of **interpolating**, (that

#### is 'putting between' known values) and is quite trustworthy. The latter, though, is a case of **extrapolating** (that is 'putting beyond' known values) and may be wildly inaccurate. The elastic may

well break under the action of the 1 kg mass!

The mass *x* is known as the **independent** or **exploratory**

**variable** and is controlled by the experimenter. The length *y* is called the **dependent** or **response variable** and is less accurate.

#### For any fixed value of *x* used repeatedly the resulting readings for *y* will form a normal distribution.

It may be tempting to extrapolate in the example illustrated

opposite, and modern day planners have to do just that, but the Plague of 1665 and the Great Fire of 1666 would be guaranteed to sabotage any attempt in this case.

Any estimates outside the range of the data are dangerous and the further away they are the less trust can be placed in them.

Estimates of *x* based on given values of *y* may be obtained from the line but since it was constructed to minimise errors in the *y* direction it was not designed for this use, so answers are bound to be unreliable.

**Drawing the line**

Looking at the equation

population

of London ?

30 40 50 60 70

years 1600+

*y*  *y*  *sxy*

*sx* 2

*x*  *x* 

#### we can see that *x*  *x* , *y*  *y* satisfies it so *x*, *y*  will always be a point on the line. To find a couple more points to enable you to draw the line use the *y*ˆ values with the two *x* values at the ends of the given set of values.

So, for the elastic band example,

*x*  50 

*x*  400 

*y*ˆ  38.3

*y*ˆ  111.

**Other forms of the equation**

Since *y*  *y*  *sxy sx* 2

 *y*  *y*  *sxy*

*x*  *x* 

 *x*  *x* 

*sy sxsy*

*sx*

 

*sy*

 *y*  *y*  *r*  *x*  *x*



 *sx* 

*s* 1 *x*  *x* *y*  *y* 

#### Also

 *xy*  *n*

*sx* 2

1 *x*  *x* 2

##### n

 *x*  *x* *y*  *y* 

*x*  *x* 2

so *y*  *y*  ˆ*x*  *x* 

#### where

ˆ  *x*  *x* *y*  *y* 

*x*  *x* 2

 *n* *xy*   *x* *y n* *x*2   *x*2