

Correlation and Regression

Correlation and regression are techniques which are used to see whether a relationship exists between two or more different sets of data

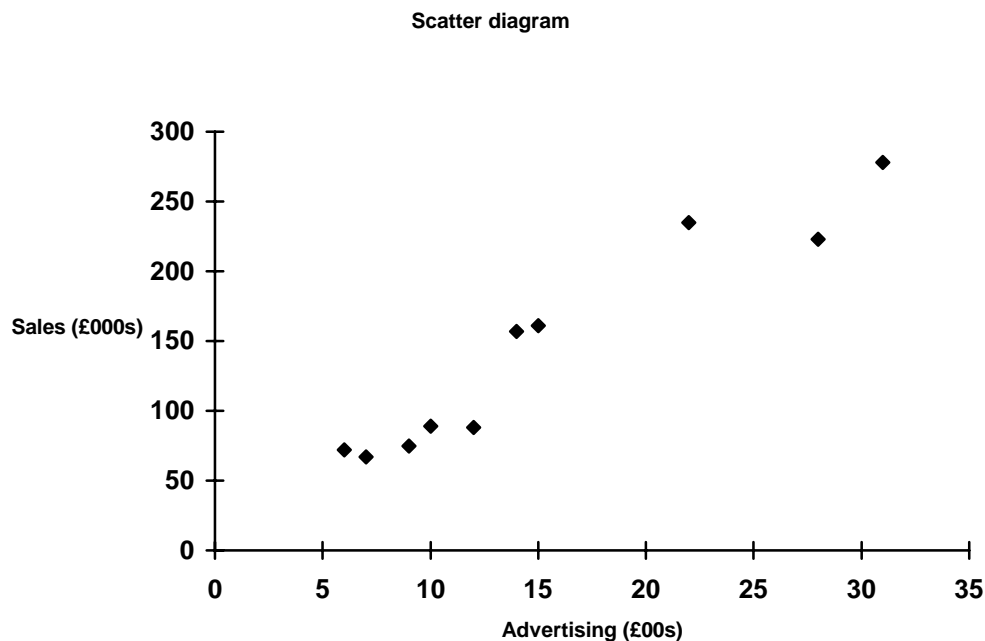
Learning Objectives:

- To identify, by diagram, whether a possible relationship exists between two variables;
- To quantify the strength of association between variables using the correlation coefficient;
- To show how a relationship can be expressed as an equation;
- To identify linear equations when written and when graphed;
- To examine regression, a widely used linear model, and to consider its uses and limitations.

Correlation and Regression

Scatter Diagrams

A graph known as a scatter diagram is used to identify the possibility and type of relationship.



y is defined as the variable which it is believed is being influenced (dependent)

x is defined as the variable which is doing the influencing (independent).

Correlation and Regression

Correlation

The strength of a relationship between two sets of data is measured by Pearson's correlation coefficient (r). It is found by the following formula:

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

EXCEL: =CORREL(Y DATA, X DATA)

Correlation and Regression

Example 4.1 - Calculation of r

Sales (y)	Expenditure (x)	x^2	y^2	xy
25	8	$8*8=64$	$25*25=625$	$8*25=200$
35	12	etc. 144	etc. =1225	etc. = 420
29	11	121	841	319
24	5	25	576	120
38	14	196	1444	532
12	3	9	144	36
18	6	36	324	108
27	8	64	729	216
17	4	16	289	68
30	9	81	900	270
$\Sigma y =$ 255	$\Sigma x =$ 80	$\Sigma x^2 =$ 756	$\Sigma y^2 = 7097$	$\Sigma xy =$ 2289

and $n = 10$

Correlation and Regression

The summary values are substituted into the correlation coefficient formula and worked through:

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

$$r = \frac{(10 \cdot 2289) - (80 \cdot 255)}{\sqrt{[(10 \cdot (756) - 80^2)][(10 \cdot 7097) - 255^2]}}$$

$$r = \frac{(22890 - 20400)}{\sqrt{(7560 - 6400)(70970 - 65025)}}$$

$$r = \frac{2490}{\sqrt{6896200}} = \frac{2490}{2626.0617}$$

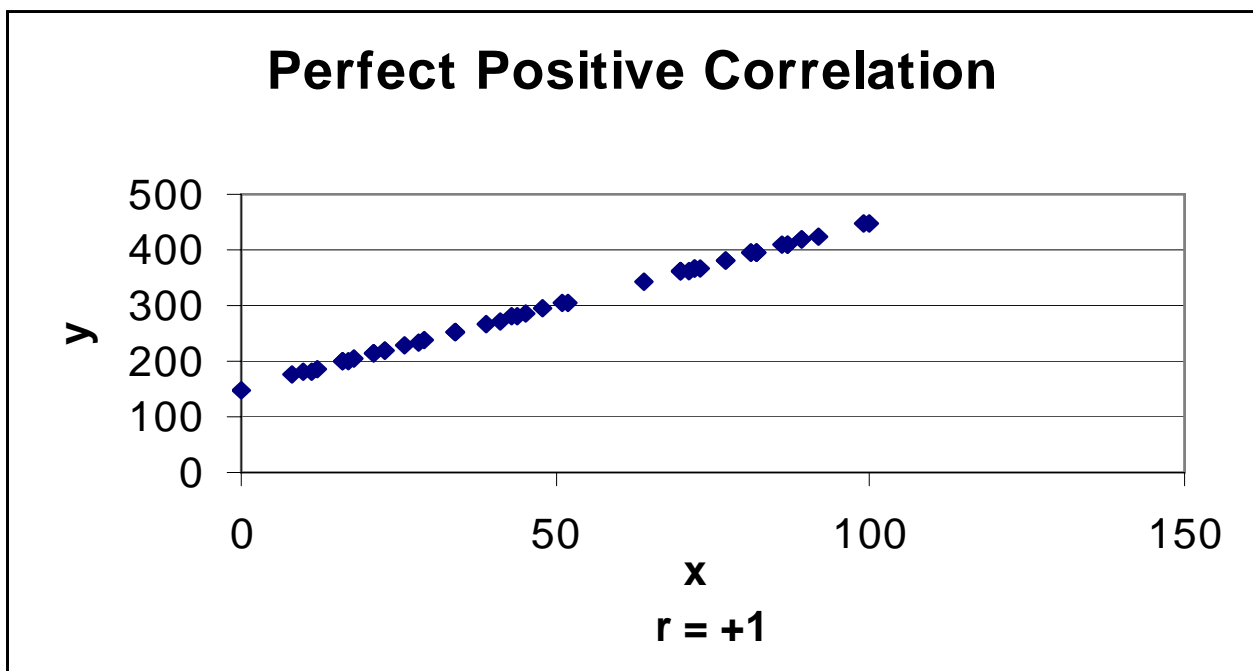
so $r = 0.948$ (to 3 d.p.)

Correlation and Regression

Interpretation of r

The value of r can only take a value of -1 to $+1$ inclusive:

+1 Perfect positive correlation exists between the data. If x is known y can be predicted **exactly**.



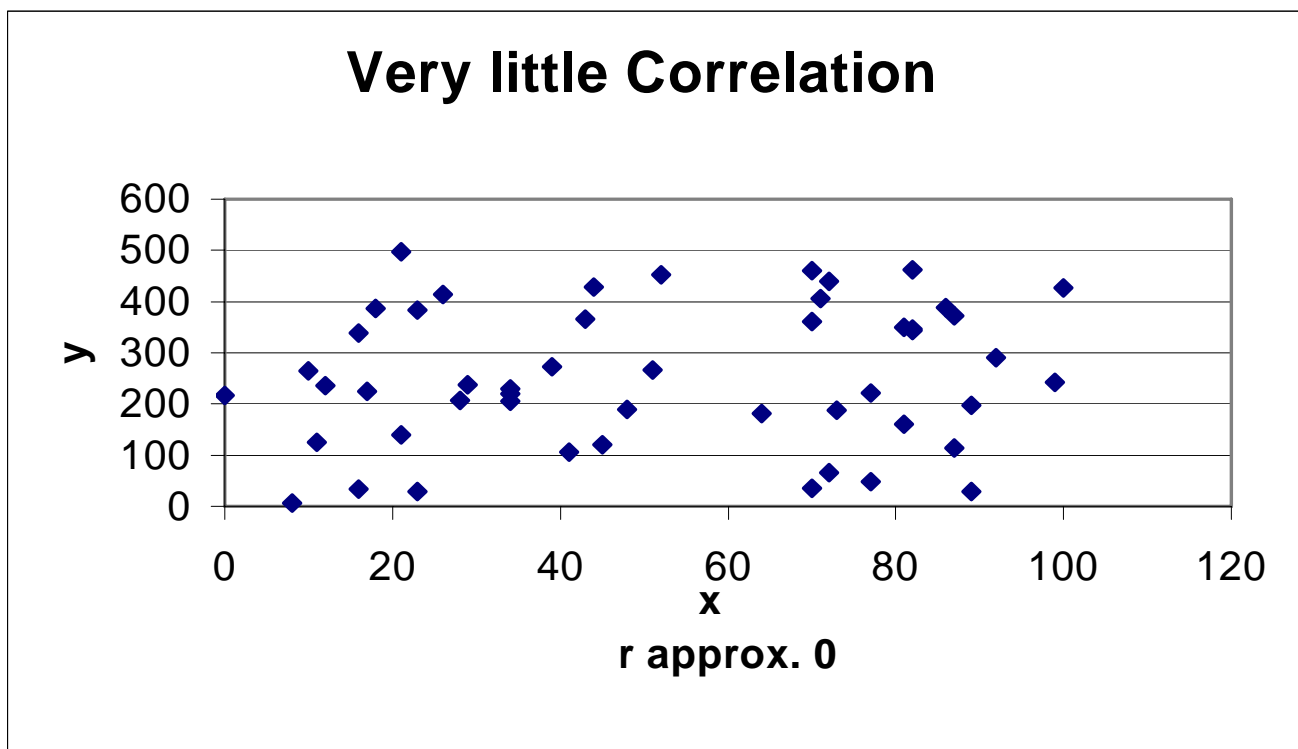
+0.8 < +1 Strong positive correlation exists between the data. As x increases y increases.

Correlation and Regression

Interpretation of r

$+0.4 < +0.8$ Moderate positive correlation exists between the data. As x increases y increases

$-0.4 < +0.4$ Very little correlation exists between the data

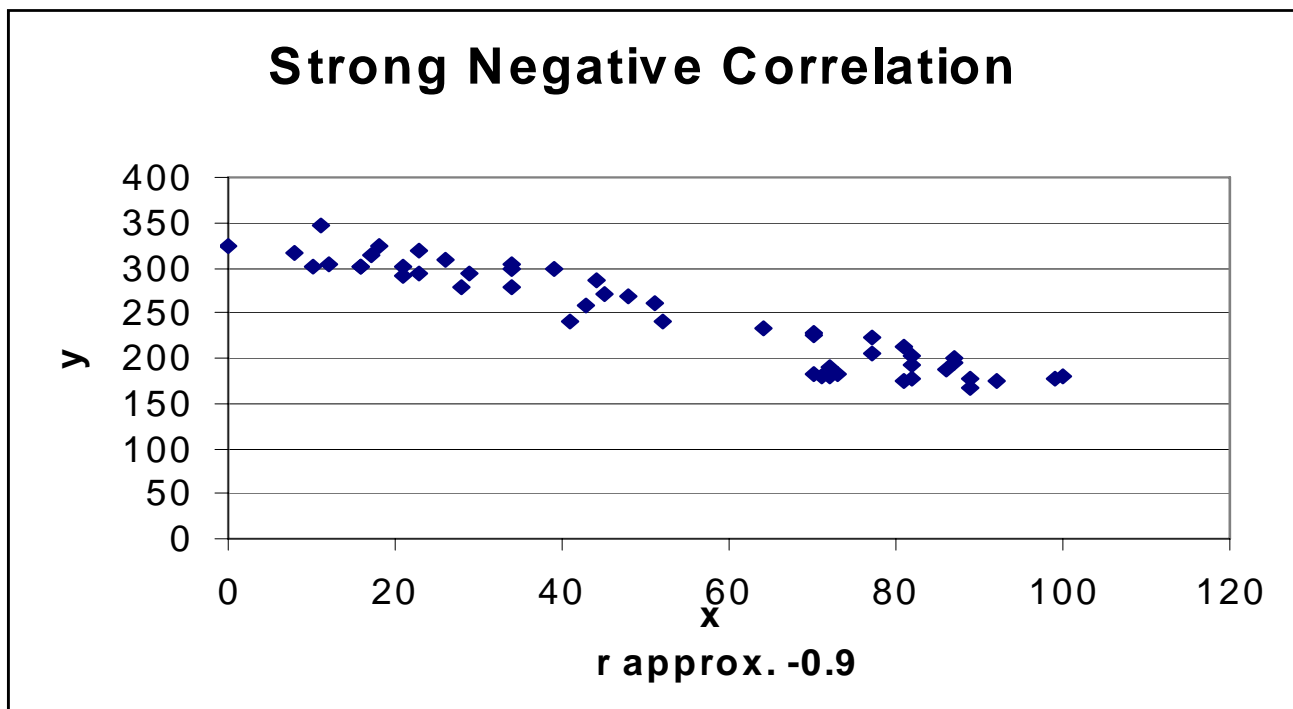


$-0.4 < -0.8$ Moderate negative correlation exists between the data. As x increases y decreases.

Correlation and Regression

Interpretation of r

$-0.8 < -1$ Strong negative correlation exists between the data. As x increases y decreases.



-1 Perfect negative correlation exists between the data. If x is known y can be predicted exactly.

Correlation and Regression

Regression

Regression is a technique which builds a straight line relationship between two sets of data.

This relationship is of the form

$$y = a + bx$$

where a and b are found by the following formulae

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

EXCEL: =SLOPE(Y DATA, X DATA)

$$a = \frac{\sum y}{n} - b \frac{\sum x}{n}$$

EXCEL: =INTERCEPT(Y DATA, X DATA)

Correlation and Regression

Example 4.5 - Calculation of a and b

To calculate use Summary values from Correlation Calculation: i.e.

Σy	Σx	Σx^2	Σy^2	Σxy	n
255	80	756	7097	2289	10

SLOPE:

$$b = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(10*2289) - (80*255)}{(10*756) - (80)^2}$$

$$b = \frac{22890 - 20400}{7560 - 6400} = \frac{2490}{1160}$$

$$b = 2.1465517$$

INTERCEPT:

$$a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n} = \frac{255}{10} - 2.1465517 * \frac{80}{10}$$

$$a = 25.5 - 17.172413 = 8.327587$$

Correlation and Regression

Example 4.5 - Calculation of a and b

The final answers (rounded to three decimal places) are:

$$a = 8.328 \quad b = 2.147$$

(note that 3 decimal places were chosen as the data supplied were in thousands and hundreds)

These give the linear regression equation

$$y = 8.328 + 2.147x$$

or, if preferred,

$$\text{sales} = 8.328 + 2.147 * \text{advertising expenditure}$$

Correlation and Regression

Forecasts

Forecasts may be made using the resulting model.

If the x (independent) value used falls within the original data set then this forecast is known as **interpolation**.

e.g. Advertising expenditure = £700 (inside original range) i.e. $x = 7$ giving

$$y = 8.328 + 2.147 * 7 = 23.357$$

i.e. 23,357 sales are forecast

If the x value falls outside the bounds of the original data then this forecast is known as **extrapolation** and care must be taken in its use.

Expenditure = £1800, so $x = 18$

$$y = 8.328 + 2.147 * 18 = 46.974$$

i.e., 46,974 sales are forecast