Bivariate Data

1. Concept of least squares
2. Equation of regression lines
3. Estimation using equation of regression lines
4. Product moment correlation coefficient, r – Part 1
5. Product moment correlation coefficient, r – Part 2 (Further examples)
6. Relationships between the product moment correlation coefficient and the regression coefficients
7. Hypothesis test on product moment correlation coefficien
8. Discussion of exam-styled questions.

Learning outcomes

1. Concept of least squares

- To explain the concept of least squares
- To describe the meaning of regression lines
- To explain correlation in the context of a scatter diagram.

2. Equation of regression lines

- To calculate the equations of regression lines from raw data.
- To calculate the equations of regression lines from summarized data.
- To visualize and plot the calculated regression lines.

Learning outcomes

3. Estimation using equation of regression lines

- To distinguish and appreciate the distinction between the regression line of y on x and that of x on y.
- To select and use, in the contest of a problem, the appropriate regression line to estimate a value.
- To state the uncertainties associated with such estimation in (2).

Learning outcomes

4 & 5 Product moment correlation coefficient, r – Part 1 & part 2 (Further examples)

- To calculate the product moment correlation coefficient from raw data.
- To calculate the product moment correlation coefficient from summarized data
- To interpret the value of the product moment correlation coefficient and relate it to the appearance of the scatter diagram.
- To explain the interpretation of cases where the value of the product moment correlation coefficient is close to +1, -1 or 0.

Learning outcomes

6. Relationships between the product moment correlation coefficient and the regression coefficients

- To recall and use the facts that both the regression lines pass through the mean center.
- The product moment correlation coefficient and the regressions coefficients are related by $r^2 = b_1 b_2$.

7. Hypothesis test on product moment correlation coefficient

- To carry out the hypothesis test on the product moment correlation coefficient.
- To draw conclusion correctly.

Learning outcomes

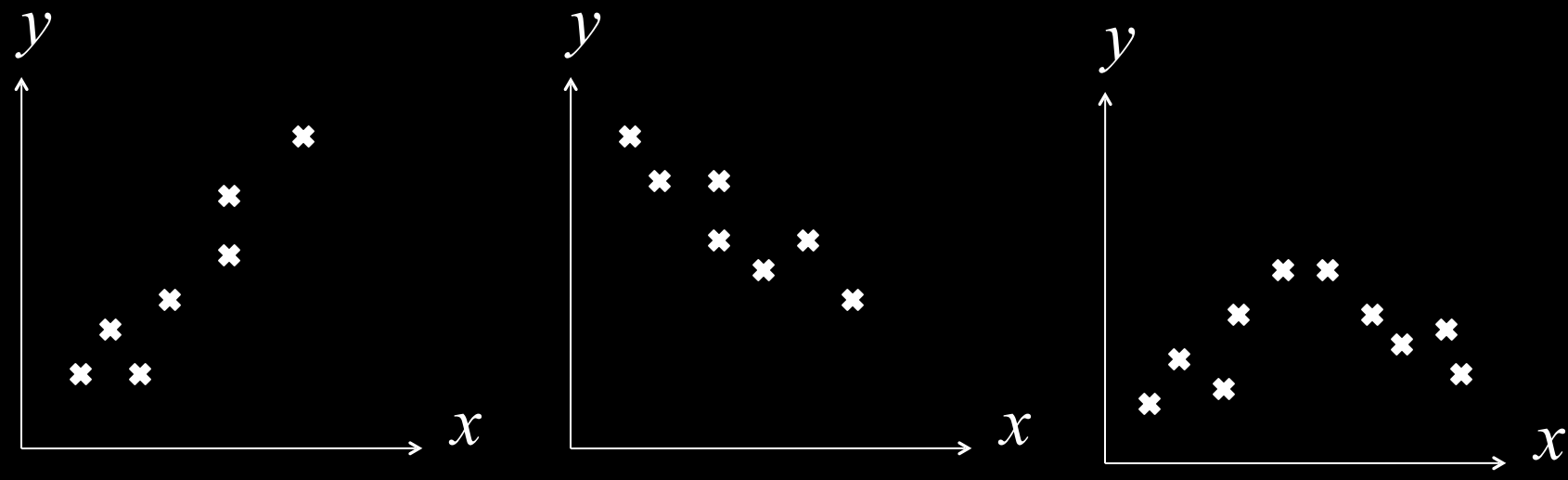
8. Discussion of exam-styled questions.

- Able to answer exam-styled questions consisting of mixed sub-topics.
- Able to relate the relationships between the sub-topics.

Bivariate Data

Scatter Diagram

- Sometimes we wish to investigate the results of a statistical enquiry or experiment by comparing two sets of data, x and y.
- Data connecting two variables are known as *bivariate data*.
- When pairs of values are plotted, a *scatter diagram* is produced.



Regression Function

- We then look for a relationship $y = f(x)$, where the function f is to be determined. This function is called the ***regression function***.

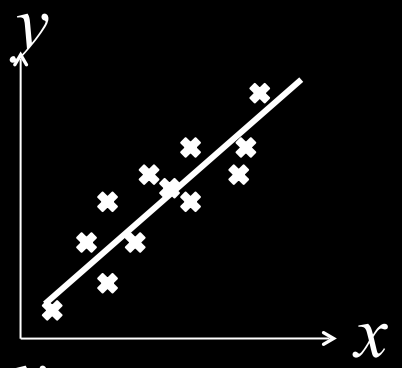
Linear Correlation and Regression lines

- We shall consider only the simplest type of function where $y = f(x)$ is a straight line. If all the points in the scatter diagram seem to lie near a straight line (regression line), we say that there is ***linear correlation*** between x and y.

Linear Correlation and Regression lines

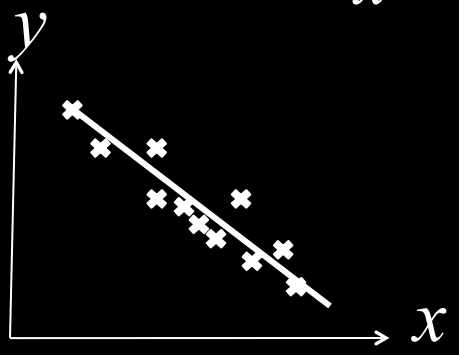
(a) If y tends to increase as x increases,

$\Rightarrow$ positive linear correlation



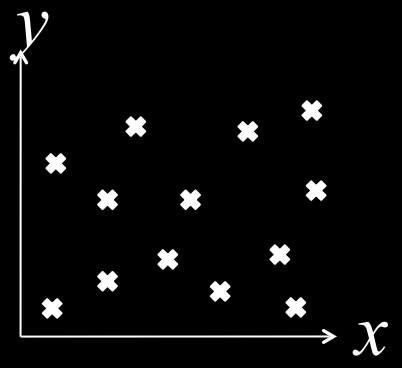
(b) If y tends to decrease as x increases,

$\Rightarrow$ negative linear correlation



(c) If there is no relationship between x and y,

$\Rightarrow$ no correlation



Least Squares Regression lines

(a) The least squares regression line y on x

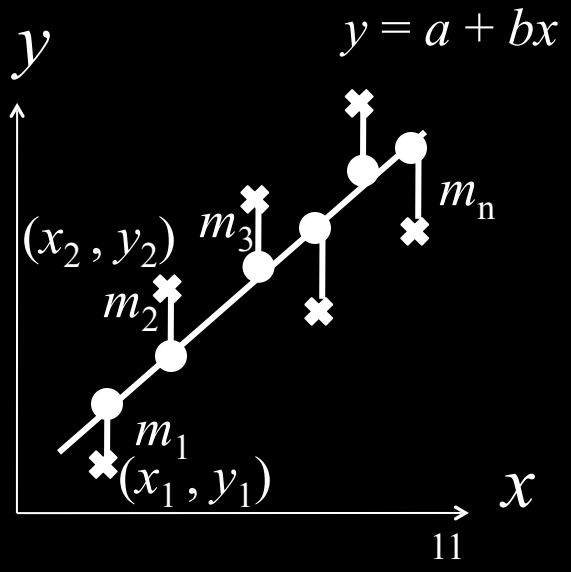
Let the equation of the line be $y = a + bx$ on the scatter diagram, shows the point (x_i , y_i) where i = 1, 2, … , n. The vertical distances $m_1 , m_2 , m_3 , \ldots, m_n$ drawn from each point to the regression line, are called *residuals*.

$$m_1 = a + bx_1 - y_1$$

$$\Rightarrow m_1^{\,2} = (a + bx_1 - y_1)^2$$

$$m_2^{\,2} = (a + bx_2 - y_2)^2$$

$$\sum m_i^{\,2} = \sum (a + bx_i - y_i)^2 \quad i = 1,2,\ldots,n$$



Least Squares Regression lines

(a) The least squares regression line y on x

$\sum m_i^2$ is the sum of the squares of the residuals.

It is possible to find value of a and b such that $\sum m_i^2$ is a minimum. With these values, the line $y = a + bx$ is known as *the least squares regression line y on x*.

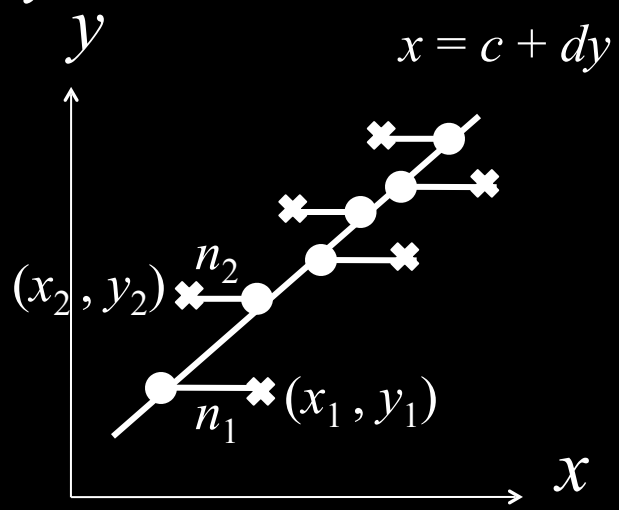
Least Squares Regression lines

(b) The least squares regression line x on y

Let the equation of the line be $x = c + dy$.

It is possible to find value of c and d such that $\sum n_i^2$ is a minimum.

With these values, the line $x = c + dy$ is known as *the least squares regression line x on y*.



Covariance, s_{xy}

For n pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the covariance, s_{xy}

$$s_{xy} = \frac{1}{n}\sum (x - \bar{x})(y - \bar{y})$$

If x and y are independent, , $s_{xy} = 0$

Note: $$s_{xx} = \frac{1}{n}\sum (x - \bar{x})(x - \bar{x}) = s_x^2 = \text{variance}$$

$$s_{yy} = \frac{1}{n}\sum (y - \bar{y})(y - \bar{y}) = s_y^2 = \text{variance}$$

Covariance, s_{xy}

Alternative forms of the formulae

$$s_{xy} = \frac{\sum xy}{n} - \bar{x}\bar{y} \qquad \text{covariance of } (x, y)$$

$$s_x^2 = \frac{\sum x^2}{n} - \bar{x}^2 \qquad \text{variance of } x$$

$$s_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 \qquad \text{variance of } y$$

Covariance, s_{xy}

Alternative forms of the formulae

Note : $$S_{xy} = \sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)$$

The connection between "big s" and "little s":

$$S_{xy} = ns_{xy}$$

$$S_{xx} = ns_x^{\,2}$$

$$S_{yy} = ns_y^{\,2}$$

Regression Coefficients

For the least squares regression line y on x, $y = a + bx$

$$\sum y = na + b\sum x \qquad \text{.....(1)}$$

$$\sum xy = a\sum x + b\sum x^2 \qquad \text{.....(2)}$$

$$(1) \times \sum x \Rightarrow \sum x \sum y = na\sum x + b\left(\sum x\right)^2 \qquad \text{.....(3)}$$

$$(2) \times n \Rightarrow n\sum xy = an\sum x + nb\sum x^2 \qquad \text{.....(4)}$$

$(4) - (3)$

$$n\sum xy - \sum x \sum y = nb\sum x^2 - b\left(\sum x\right)^2$$

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

Regression Coefficients

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

$$b = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} \qquad \Rightarrow b = \frac{s_{xy}}{s_x^{\,2}}$$

b is known as the coefficient of regression of y on x, where

$$b = \frac{s_{xy}}{s_x^{\,2}}$$

Regression Coefficients

For the least squares regression line x on y, $x = c + dy$,

d is known as the coefficient of regression of x on y, where

$$d = \frac{s_{xy}}{s_y^2}$$

Calculating the equation of least squares regression lines

The equation of the least squares regression of y on x, $y = a + bx$, has gradient b and passes through $(\bar{x}, \bar{y})$, so

The equation of the least squares regression line y on x is

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Similarly,

The equation of the least squares regression line x on y is

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

Calculating the equation of least squares regression lines

Note : using "big s"

The equation of the least squares regression lines are:

y on x : $$y - \bar{y} = \frac{S_{xy}}{S_x^{\,2}}\left(x - \bar{x}\right)$$

x on y : $$x - \bar{x} = \frac{S_{xy}}{S_y^{\,2}}\left(y - \bar{y}\right)$$

Example 1:

Calculate the equation of the least squares regression lines of

(a) y on x,

(b) x on y

for the following data.

x	1	2	4	6	7	8	10
y	10	14	12	13	15	12	13

Example 2:

For twelve consecutive months a factory manager recorded the number of items produced by the factory and the total cost of their production. The following table summarizes the manager's data.

Number of items (x) thousands	**18**	**36**	**45**	**22**	**69**	**72**	**13**	**33**	**59**	**79**	**10**	**53**
Production cost (y) £1000	***37***	***54***	***63***	***42***	***84***	***91***	***33***	***49***	***79***	***98***	***32***	***71***

Example 2:

(a) Draw a scatter diagram for the data.

(b) Give a reason to support the use of the regression line

$$(y - \bar{y}) = b(x - \bar{x})$$

as a suitable model for the data.

(c) Giving the values of $\bar{x}, \bar{y}$ and b to 3 decimal places, obtain the regression equation for y on x in the above form.

(You may use $\sum x^2 = 27963$, $\sum xy = 37249$.)

Example 2:

(d) Rewrite the equation in the form

$$y = a + bx$$

giving a to 3 significant figures.

(e) Give a practical interpretation of the values of a and b.

(f) The selling price of each item produced is £1.60. Find the level of output at which total income and estimates total costs are equal. Give a brief interpretation of this value.

Example 3:

An electric fire was switched on in a cold room and the temperature of the room was noted at five-minute intervals.

Time, minutes, from switching on	**0**	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
Temperature, °	***0.4***	***1.5***	***3.4***	***5.5***	***7.7***	***9.7***	***11.7***	***13.5***	***15.4***

You may assume that

$\sum x = 180$ $\sum y = 68.8$ $\sum xy = 1960$ $\sum x^2 = 5100$

Example 3:

(a) Plot the data on a scatter diagram.

(b) Calculate the regression line $y = a + bx$ and draw it on your scatter diagram.

(c) Predict the temperature 60 minutes from switching on the fire. Why should this prediction be treated with caution?

(d) Starting from the equation of the regression line $y = a + bx$, derive the equation of the regression line of

(i) y on t where y is temperature in $^{o}\mathbf{C}$ (as above) and t is time in hours.

Example 3:

(d) Starting from the equation of the regression line $y = a + bx$, derive the equation of the regression line of

(ii) z on x where z is temperature in **°K** and x is time in minutes (as above).

(A temperature in **°C** is converted to **°K** by adding 273, e.g. 10 **°C** → 283 **°K**)

(e) Explain why, in (b), the line $y = a + bx$ was calculated rather than $x = a' + b' y$. If, instead of the temperature being measured at five-minute intervals, the time for the room to reach predetermined temperatures (e.g. 1 , 4, 7, 10, 13 **°C**) had been observed, what would the appropriate calculation have been? Explain your answer.

The product-moment correlation coefficient, *r*

The product-moment correlation coefficient, *r*, is a numerical value which indicates the degree of scatter.

$$-1 \leq r \leq 1$$

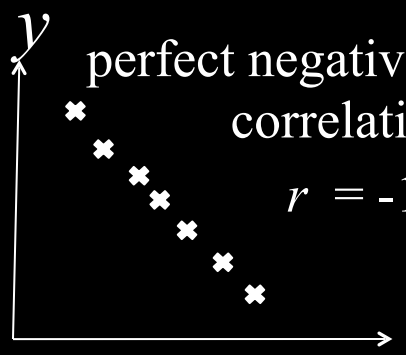
r is a very useful measure because it is independent of units of scale of the variables.

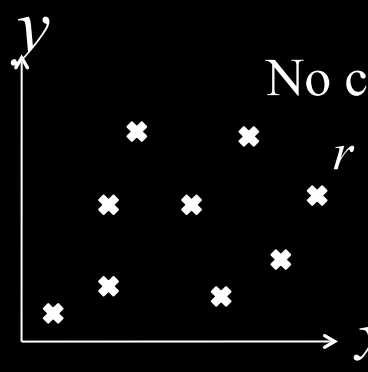
The product-moment correlation coefficient, r

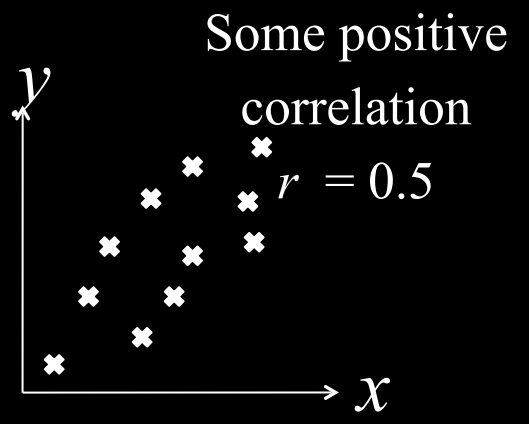
$r = 1 \Rightarrow$ perfect positive linear correlation

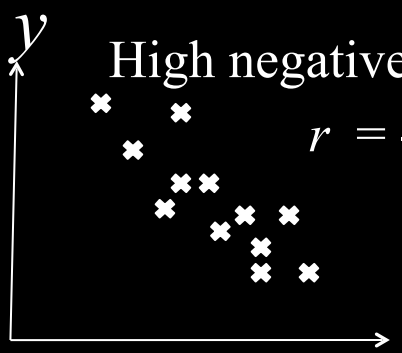
$r = -1 \Rightarrow$ perfect negative linear correlation

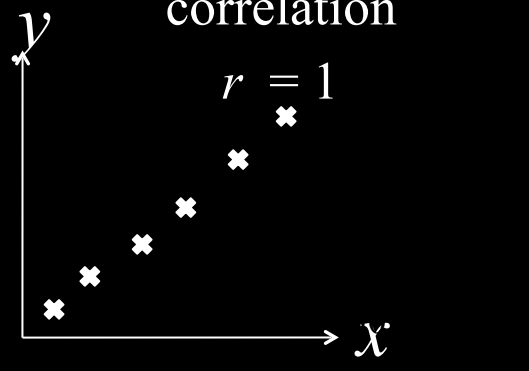
$r = 0 \Rightarrow$ no correlation











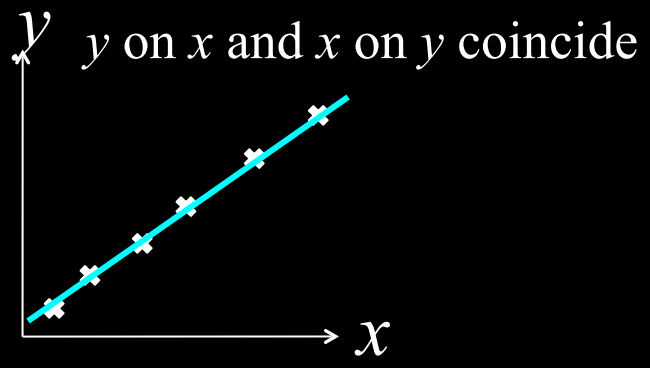
The product-moment correlation coefficient, *r*

The product-moment correlation coefficient, *r*, is given by

$$r = \frac{S_{xy}}{S_x S_y}$$

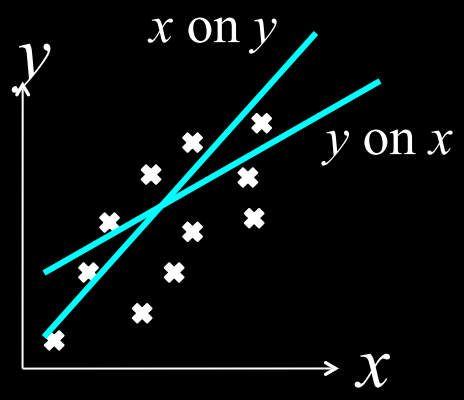
Diagrammatic representation of the value of *r*

Two regression lines, *y* on *x* and *x* on *y*, on a scatter diagram can give a good idea of the value of *r*. The closer the two lines are together, the nearer *r* is to 1 or -1.



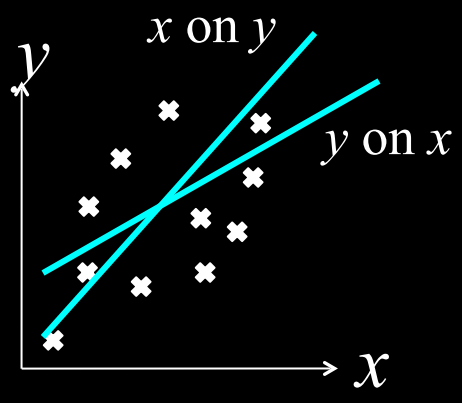
Perfect positive Correlation

$r = 1$



High positive correlation

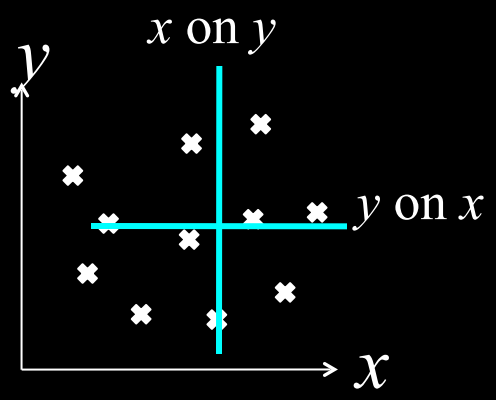
$r = 0.8$



Some positive correlation

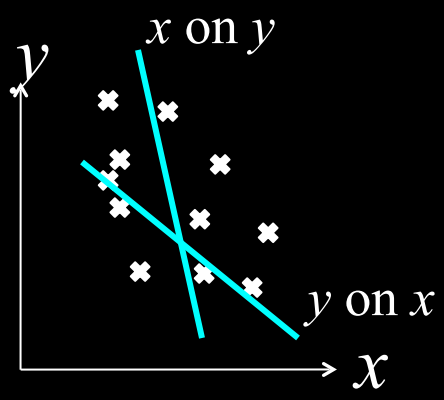
$r = 0.5$

Diagrammatic representation of the value of *r*



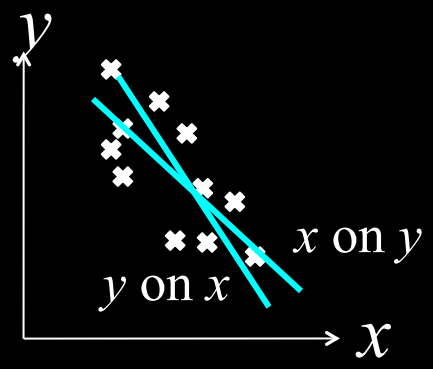
No correlation

$r = 0$



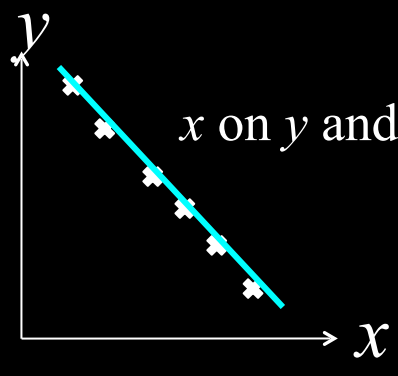
Some negative correlation

$r = -0.4$



Some negative correlation

$r = -0.9$



Perfect negative Correlation

$r = -1$

Example 4:

The following table shows the marks of ten candidates in Physics and Mathematics. Find the product-moment correlation coefficient and comment on your value.

Marks in Physics (x)	**18**	**20**	**30**	**40**	**46**	**54**	**60**	**80**	**88**	**92**
Marks in Mathematics (y)	***42***	***54***	***60***	***54***	***62***	***68***	***80***	***66***	***80***	***109***

Relationship between regression coefficient and *r*

For the regression lines :

y on $x \Rightarrow y = a + bx$ where $b = \dfrac{s_{xy}}{s_x s_x}$

x on $y \Rightarrow x = c + dy$ where $d = \dfrac{s_{xy}}{s_y s_y}$

b and d are the regression coefficients.

$$bd = \frac{s_{xy}}{s_{xx}} \times \frac{s_{xy}}{s_{yy}} = \frac{s_{xy}}{s_x s_x} \times \frac{s_{xy}}{s_y s_y} = \left(\frac{s_{xy}}{s_x s_y} \right)^2 = r^2$$

Relationship between regression coefficient and *r*

Either *b* and *d* are both positive or *b* and *d* are both negative,

$$r^2 = bd \quad \text{and}$$

$$r = +\sqrt{bd} \quad \text{if } b, d \text{ are positive}$$

$$r = -\sqrt{bd} \quad \text{if } b, d \text{ are negative}$$

Example 5:

In Example 1, we found that the least squares regression line y on x is $y = 11.7 + 0.186x$ and the least squares regression line x on y is $x = -4.34 + 0.769y$. Using this information find r, the product-moment correlation coefficient.

x	1	2	4	6	7	8	10
y	10	14	12	13	15	12	13

Example 6:

Show that if $r = \pm 1$, the regression lines of y on x and x on y are identical.

Example 7:

If $r = 0$, show that the two regression lines are at right angles.

Example 8:

For each set of bivariate data, find the product-moment correlation coefficient, draw a scatter diagram and then comment on your value of r.

(a)

x	-2	-1	0	1	2
y	4	1	0	1	4

(b)

x	1	1	1	2	2	2	3	3	3	9
y	1	2	3	1	2	3	1	2	3	8

$r = 0$ implies either

(a) There is no correlation between the variables and they are independent

OR

(b) The variables are related in a non-linear way.

Example 9:

Draw a diagram to illustrate the lengths whose sum of squares is minimized in the least squares method for finding the regression line of y on x.

State which is the independent and which is the dependent variable.

State, giving your reason, whether or not the equation of this line can be used to estimate the value of x for a given value of y.

Example 9:

The length (L mm) and width (W mm) of each of 20 individuals of a single species of fossil are measured. A summary of the results is :

$$\sum L = 400.20,\ \sum W = 176.00,\ \sum LW = 3700.20$$
$$\sum L^2 = 8151.32,\ \sum W^2 = 1780.52.$$

(a) Obtain the product-moment correlation coefficient between the length and the width of these fossils. Without performing a significance test interpret your result.

Example 9:

(b) Obtain an equation of the line of regression from which it is possible to estimate the length of a fossil of the same species whose width is known, giving the values of the coefficients to 2 decimal places.

(c) From your equation find the average increase or decrease in length per 1 mm increase in width of these fossils.

Using a method of coding

For data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ suppose we use the coding

$$X = \frac{x - a}{b} \quad \text{and} \quad Y = \frac{y - c}{d}$$

b and d are the *scaling constants*.

$$\Rightarrow x_i = a + bX_i \qquad y_i = c + dY_i \qquad \text{for i} = 1, 2, \ldots, \text{n}$$

$$\bar{x} = a + b\bar{X} \qquad \bar{y} = c + d\bar{Y}$$

$$s_x = bs_X \qquad s_y = ds_Y$$

Using a method of coding

For covariance

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$= \frac{1}{n} \sum \left[a + bX_i - (a + b\bar{X}) \right] \times \left[c + dY_i - (c + d\bar{Y}) \right]$$

$$= \frac{1}{n} \sum b\left[X_i - \bar{X}) \right] \times d\left[Y_i - \bar{Y} \right]$$

$$s_{xy} = bds_{XY}$$

Using a method of coding

For the product-moment correlation coefficient

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{1}{bd} s_{xy}}{\frac{1}{b} s_x \times \frac{1}{d} s_y} = \frac{s_{xy}}{s_x s_y} = r_{xy}$$

$$r_{XY} = r_{xy}$$

The product-moment correlation coefficient remain *unchanged*. This is because *r* is a measure of the degree of scatter and this is unchanged by a change of origin and scaling.

Example 10:

For the following data, use a method of coding to find

(a) The covariance,

(b) The product-moment correlation coefficient,

(c) The least squares regression line y on x and x on y.

x	1000	1012	1009	1007	1010	1015	1010	1011
y	235	240	245	250	255	260	265	270

Significance tests for *r*, the product-moment correlation coefficient

Assume that the two variables *x* and *y* are jointly normally distributed with the population correlation coefficient ρ,

$H_o : \rho = 0$

(H_o always ρ = 0, no correlation between the variables)

$H_1 : \rho > 0$ (there is a positive correlation between the variables)

OR

$H_1 : \rho < 0$ (there is a negative correlation between the variables)

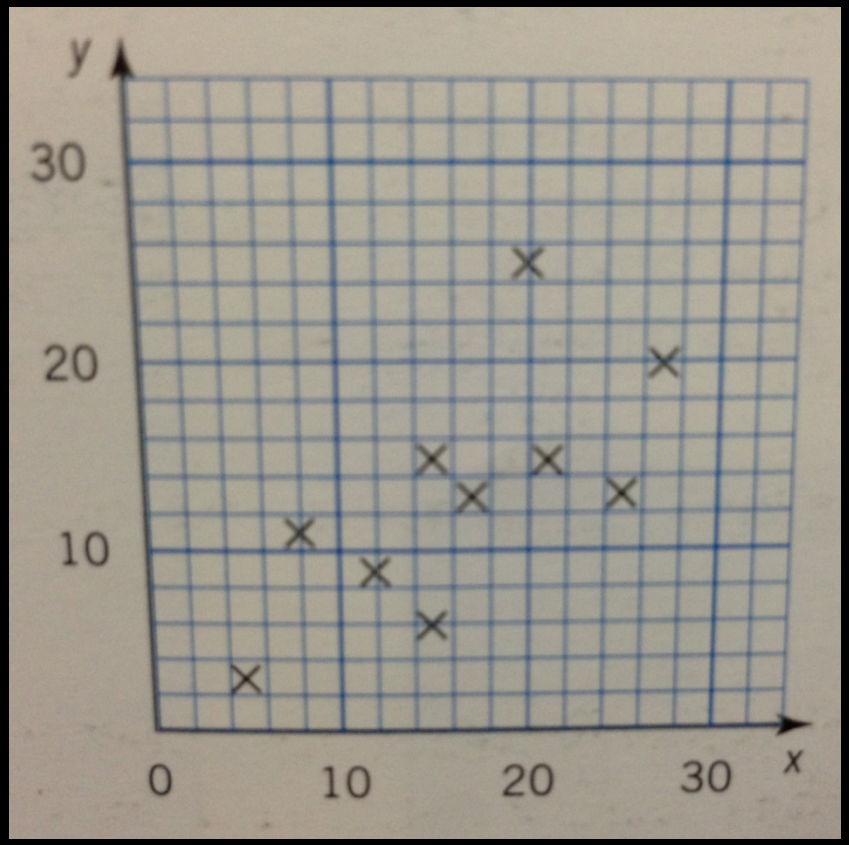
OR

$H_1 : \rho \neq 0$ (there is some correlation between the variables)

The calculated value of *r*, is compared with the critical value from tables.

Example 11:

The scatter diagram illustrating ten pairs of values (x , y) is shown below.



(a) Comment on the diagram.

Example 11:

The scatter diagram illustrating ten pairs of values (x , y) is shown below.

(b) Calculate the product-moment correlation coefficient for the pairs of data shown in the diagram.

(c) Assuming that X and Y are jointly normal distributed with correlation coefficient ρ, and the data constitutes a random sample, test, at the 5% level, whether there is a positive correlation between X and Y.

(d) Would your conclusion be the same at the 1% level?