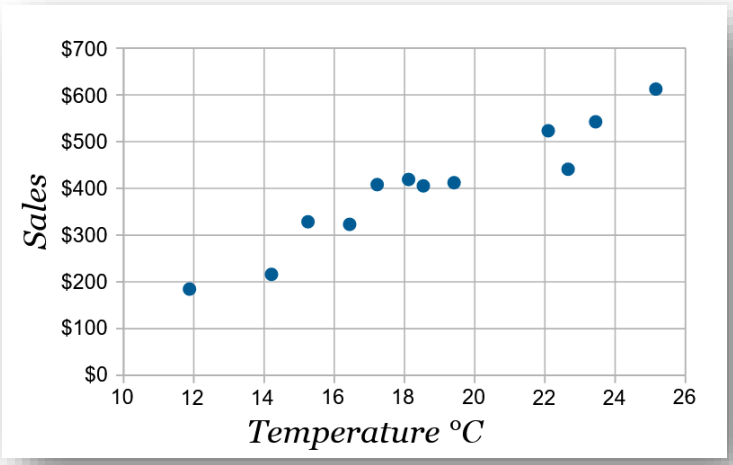
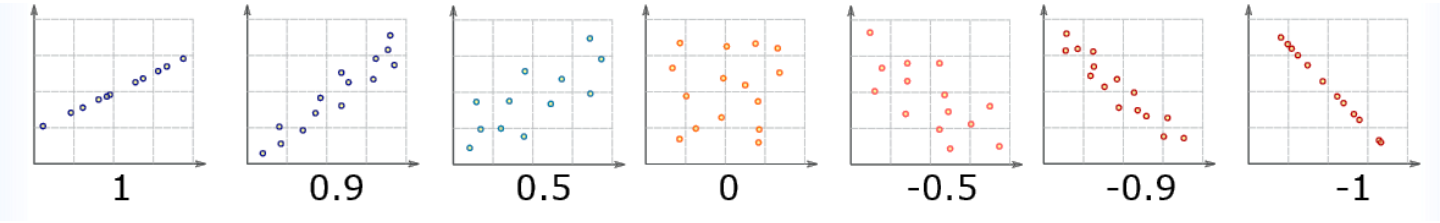
Scatter Diagrams:



- *Bivariate Data*: Each point has x and y values, for example from the diagram shown, at 12^0 C we get sales of $200
- An *Independent (or Explanatory) variable* : is one that is set independently of the other variable and is plotted on the x-axis
- A *dependent (or response) variable*: is one whose values are determined by the values of the independent variable, and is plotted on the y-axis
- *Correlation*: describes the nature of the linear relationship between two variables



Perfect positive correlation	Strong positive correlation	weak positive correlation	No correlation	weak negative correlation	Strong negative correlation	Perfect negative correlation

- *Causal Relationship*: when one variable causes the other

Example 1:

In the study of a city, the population density, in people/hectare, and the distance from the city centre, in km, was investigated by picking a number of sample areas with the following results.

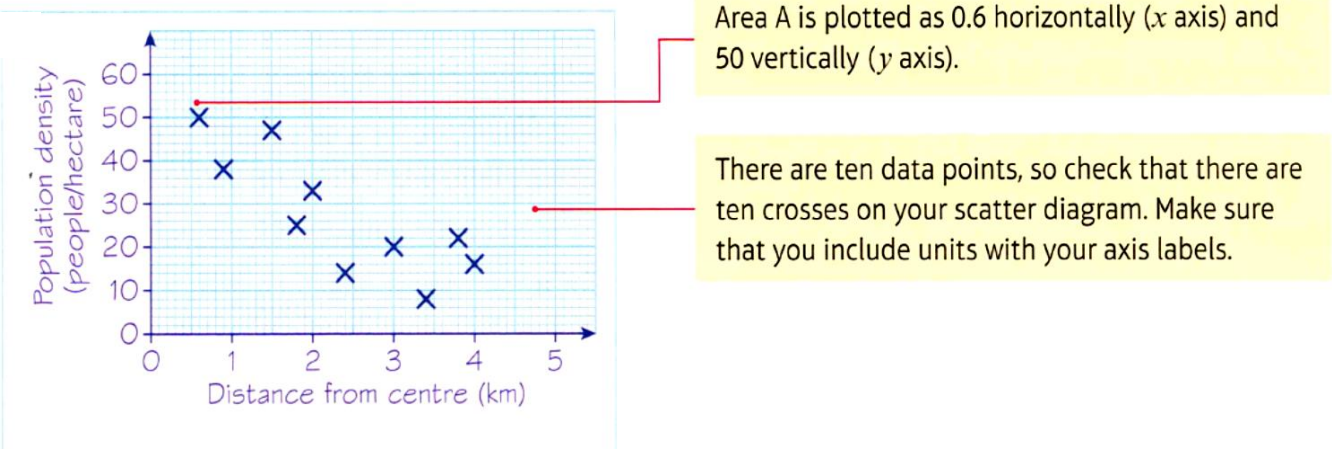
Area	A	B	C	D	E	F	G	H	I	J
Distance (km)	0.6	3.8	2.4	3.0	2.0	1.5	1.8	3.4	4.0	0.9
Population density (people/hectare)	50	22	14	20	33	47	25	8	16	38

a Draw a scatter diagram to represent these data.

b Describe the correlation between distance and population density.

c Interpret your answer to part **b**.

a.



b. A weak negative correlation

c. As the distance from the city centre increases, the population density decreases

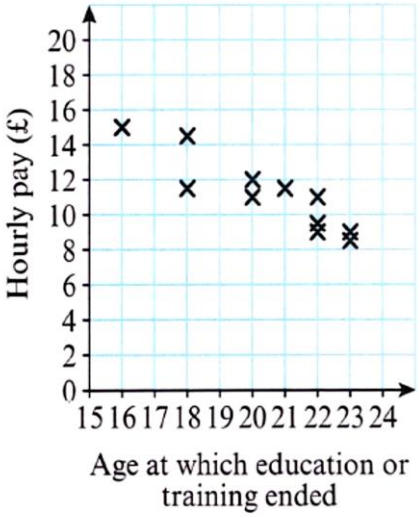
Example 2:

Hideko wanted to see if there was a relationship between what people earn and the age at which they left education or training. She asked 14 friends to fill in an anonymous questionnaire and recorded the results in a scatter diagram.

a Describe the type of correlation shown.

Hideko says that her data supports the conclusion that more education causes people to earn a lower hourly rate of pay.

b Give one reason why Hideko's conclusion might not be valid.

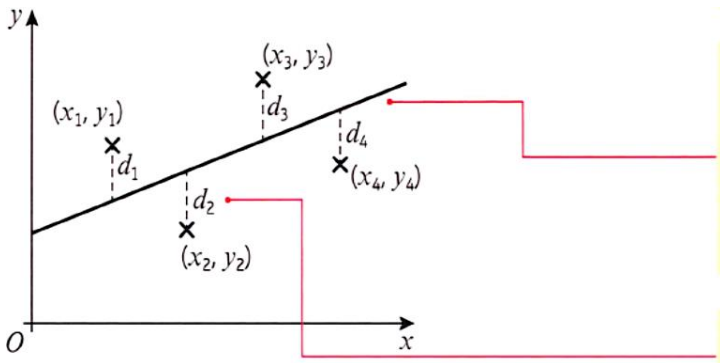


a. Weak negative correlation

b. She might have asked in a type of job that needs more hands' on experience, so those who left education early had more years of work experience and thus higher hourly rate of pay

Linear Regression

In a scatter diagram that shows a correlation, usually a line of *best fit* can be drawn. A famous line is the *least squares regression line (Simply called regression line)*, which minimizes the sum of the squares of the vertical distances between each point of data and the line



The regression line of y on x is the straight line that minimises the value of $d_1^2 + d_2^2 + d_3^2 + d_4^2$. In general, if each data point is a distance d_i from the line, the regression line minimises the value of Σd_i^2.

The point (x_2, y_2) is a vertical distance d_2 from the line.

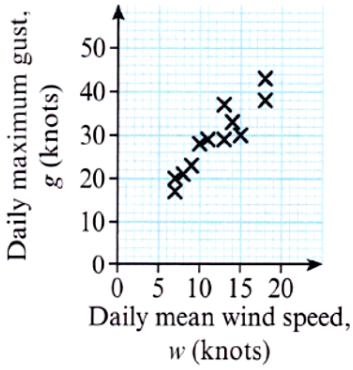
The regression line of y on x is $y = a + bx$

Example 3:

The daily mean wind speed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne, a small village near Cambridge, UK.

w	14	13	13	9	18	18	7	15	10	14	11	9	8	10	7
g	33	37	29	23	43	38	17	30	28	29	29	23	21	28	20

The data were plotted on a scatter diagram:



a Describe the correlation between daily mean wind speed and daily maximum gust.

The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$

b Give an interpretation of the value of the gradient of this regression line.

c Justify the use of a linear regression line in this instance.

a. Strong positive correlation
b. If the daily mean wind speed increases by 1 knot then the daily maximum gust increases 1.82 knots
c. The graph suggests a linear strong positive correlation and a best fit line

Interpolation vs Extrapolation:

Interpolation: Using the regression line to make predictions of the dependent variable *in* the given range

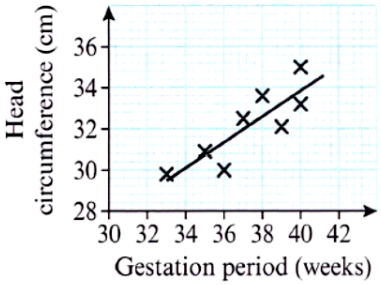
Extrapolation: Using the regression line to make predictions of the dependent variable *outside* the given range

Example 4:

The head circumference, y cm, and gestation period, x weeks, for a random sample of eight newborn babies at a clinic were recorded.

Gestation period, x (weeks)	36	40	33	37	40	39	35	38
Head circumference, y (cm)	30.0	35.0	29.8	32.5	33.2	32.1	30.9	33.6

The scatter diagram shows the results.



The equation of the regression line of y on x is $y = 8.91 + 0.624x$

The regression equation is used to estimate the head circumference of a baby born at 39 weeks and a baby born at 30 weeks.

a Comment on the reliability of these estimates.

A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6 cm.

b Explain why the regression equation given above is not suitable for this estimate.

a. A baby born at 39 weeks is within the range of the given data thus is quite reliable to estimate by interpolation. On the other hand a baby born at 30 weeks is outside the range which needs extrapolation, and thus is less reliable

b. In the scatter diagrams in general we should follow the given line. In our case y on x where gestation period is the independent variable. The nurse wants to use the head circumference as an independent variable, so she needs to use an x on y model for more reliability.

Calculating the least squares linear regression:

We know that $y = a + bx$ so let's find the *a* and the *b*

First, we have to introduce the ***summary statistics*** S_{xy}, S_{xx} and S_{yy}

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n} \qquad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \qquad S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

Now $b = \frac{S_{xy}}{S_{xx}}$, $a = \bar{y} - b\bar{x}$, then $y = a + bx$

Example 5:

The results from an experiment in which different masses were placed on a spring and the resulting length of the spring measured, are shown below.

Mass, *x* (kg)	20	40	60	80	100
Length, *y* (cm)	48	55.1	56.3	61.2	68

a Calculate S_{xx} and S_{xy}
(You may use $\sum x = 300$ $\sum x^2 = 22000$ $\bar{x} = 60$ $\sum xy = 18238$ $\sum y^2 = 16879.14$ $\sum y = 288.6$ $\bar{y} = 57.72$)

b Calculate the regression line of y on x.

c Use your equation to predict the length of the spring when the applied mass is:
i 58 kg
ii 130 kg

d Comment on the reliability of your predictions.

a. $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 22000 - \frac{300^2}{5} = 4000$

b. $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ = 18238 - $\frac{300 \times 288.6}{5}$ = 922

$b = \frac{S_{xy}}{S_{xx}} = \frac{922}{4000}$ = 0.2305

$a = \bar{y} - b\bar{x}$ = 57.72 – 0.2305(60) = 43.89

$y = a + bx$ y = 43.89 + 0.2305 x

c. i. Length = 43.89 + 0.2305(58) = 57.3 cm

ii. Length = 43.89 + 0.2305(130) = 73.9 cm

d. 58 cm is in the given range of the data so interpolation is quite reliable, while 130 is outside the range so extrapolation which is less reliable

Example 6:

A scientist working in agricultural research believes that there is a linear relationship between the amount of a food supplement given to hens and the hardness of the shells of the eggs they lay. As an experiment, controlled quantities of the supplement were added to the hens' normal diet for a period of two weeks and the hardness of the shells of the eggs laid at the end of this period was then measured on a scale from 1 to 10, with the following results:

Food supplement, f (g/day)	2	4	6	8	10	12	14
Hardness of shell, h	3.2	5.2	5.5	6.4	7.2	8.5	9.8

a Find the equation of the regression line of h on f.

(You may use $\sum f = 56$ $\sum h = 45.8$ $\bar{f} = 8$ $\bar{h} = 6.543$ $\sum f^2 = 560$ $\sum fh = 422.6$)

b Interpret what the values of a and b tell you.

a. $S_{ff} = \sum f^2 - \frac{(\sum f)^2}{n}$ =560 - $\frac{56^2}{7}$ = 112

$S_{fh} = \sum fh - \frac{\sum f \sum h}{n}$ = 422.6 - $\frac{56 \times 45.8}{7}$ = 56.2

$b = \frac{S_{fh}}{S_{ff}} = \frac{56.2}{112}$ = 0.5018 hardness units per g/day

$a = \bar{h} - b\bar{f}$ = 6.543 – 0.5018(8) = 2.5286 hardness units

$h = a + bf$ h = 2.5286 + 0.5018 f

b. *a* gives the hardness units when we have f = 0 which means at no given food supplements. It is quite reliable as f = 0 is very near to the given range

b is the rate at which the hardness units increase with increasing the food supplements. So for every g/day of food supplements we get an increase of 0.5018 hardness units

Example 7:

A repair workshop finds it is having a problem with a pressure gauge it uses. It decides to have the gauge checked by a specialist firm. The following data were obtained.

Gauge reading, x (bars)	1.0	1.4	1.8	2.2	2.6	3.0	3.4	3.8
Correct reading, y (bars)	0.96	1.33	1.75	2.14	2.58	2.97	3.38	3.75

(You may use $\sum x = 19.2$ $\sum x^2 = 52.8$ $\sum y = 18.86$ $\sum y^2 = 51.30$ $\sum xy = 52.04$)

a Show that $S_{xy} = 6.776$ and find S_{xx}

It is thought that a linear relationship of the form $y = a + bx$ could be used to describe these data.

b Use linear regression to find the values of a and b, giving your answers to 3 significant figures.

c Draw a scatter diagram to represent these data and draw the regression line on your diagram.

d The gauge shows a reading of 2 bars. Using the regression equation, work out what the correct reading should be.

a. $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ = 52.04 - $\frac{19.2 \times 18.86}{8}$ = 6.776

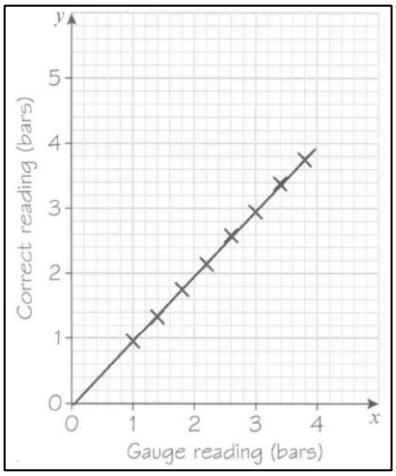
$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$ =52.8 - $\frac{19.2^2}{8}$ = 6.72

b. $b = \frac{S_{xy}}{S_{xx}} = \frac{6.776}{6.72}$ = 1.0083 = 1.01 (3 s.f.)

$a = \bar{y} - b\bar{x} = \frac{18.86}{8}$ – 1.0083($\frac{19.2}{8}$) = - 0.06242 = - 0.0624 (3 s.f.)

$y = a + bx$ y = - 0.0624 + 1.01 x

c. Look at the shown fig.



d. Correct reading = - 0.0624 + 1.01(2)

= 1.96 (3 s.f.)

Using Coding:

Example 8:

Eight samples of carbon steel were produced with different percentages, $c\%$, of carbon in them.

Each sample was heated in a furnace until it melted and the temperature, m in °C, at which it melted was recorded.

The results were coded such that $x = 10c$ and $y = \frac{m - 700}{5}$

The coded results are shown in the table.

Percentage of carbon, x	1	2	3	4	5	6	7	8
Melting point, y	35	28	24	16	15	12	8	6

a Calculate S_{xy} and S_{xx}
(You may use $\sum x^2 = 204$ and $\sum xy = 478$)

b Find the regression line of y on x.

c Estimate the melting point of carbon steel which contains 0.25% carbon.

a. $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ = 478 - $\frac{36 \times 144}{8}$ = - 170

$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$ =204 - $\frac{36^2}{8}$ = 42

b. $b = \frac{S_{xy}}{S_{xx}} = \frac{-170}{42}$ = - 4.0476 = - 4.05 (3 s.f)

$a = \bar{y} - b\bar{x} = \frac{144}{8}$ + 4.0476 ($\frac{36}{8}$) = 36.2142 = 36.2 (3 s.f)

$y = a + bx$ $\quad$ y = 36.2 - 4.05 x

c. Coding the 0.25% we get x = 10 x 0.25 = 2.5

y = 36.2 - 4.05(2.5) = 26.075

Melting point = 26.075 x 5 + 700 = 830 ^{0} C (3 s.f.)

The product moment correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$-1 \leq r \leq 1$

Where 1 is perfect positive correlation

-1 is perfect negative correlation

0 is no correlation

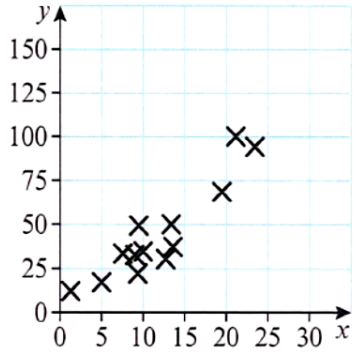
Example 9:

The number of vehicles, x millions, and the number of accidents, y thousands, were recorded in 15 different countries. The following summary statistics were calculated and a scatter diagram of the data is given to the right:

$\sum x = 176.9 \quad \sum y = 679 \quad \sum x^2 = 2576.47 \quad \sum y^2 = 39\,771 \quad \sum xy = 9915.3$

a Calculate the product moment correlation coefficient between x and y.

b With reference to your answer to part **a** and the scatter diagram, comment on the suitability of a linear regression model for these data.



a. $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$ =2576.47 - $\frac{176.9^2}{15}$ = 490.23

$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$ = 39771 - $\frac{679^2}{15}$ = 9034.93

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$ = 9915.3 - $\frac{176.9 \times 679}{15}$ = 1907.63

$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{1907.63}{\sqrt{490.23 \times 9034.93}} = 0.906$

b. From the answer to a we get a strong positive PMCC and the scatter diagram shows the possibility of a linear correlation, so a linear regression model is suitable

Example 10: (Important)

Data are collected on the amount of a dietary supplement, d grams, given to a sample of 8 cows and their milk yield, m litres. The data were coded using $x = \frac{d}{2} - 6$ and $y = \frac{m}{20}$. The following summary statistics were obtained:

$\sum d^2 = 4592 \quad S_{dm} = 90.6 \quad \sum x = 44 \quad S_{yy} = 0.05915$

a Use the formula for S_{yy} to show that $S_{mm} = 23.66$

b Find the value of the product moment correlation coefficient between d and m.

N.B: Before solving, there is one rule that we need to know

$\sum ax = a\sum x$ which is the same as $\sum \frac{x}{a} = \frac{1}{a}\sum x$

a. $S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} \quad S_{yy} = \sum \left(\frac{m}{20}\right)^2 - \frac{\left(\sum \frac{m}{20}\right)^2}{n}$

$0.05915 = \sum \frac{m^2}{400} - \frac{\frac{(\sum m)^2}{400}}{n} \quad 0.05915 = \sum \frac{m^2}{400} - \frac{(\sum m)^2}{400n}$

$0.05915 = \frac{1}{400}\left(\sum m^2 - \frac{(\sum m)^2}{n}\right) \quad 23.66 = \left(\sum m^2 - \frac{(\sum m)^2}{n}\right)$

$S_{mm} = 23.66$

b. $\sum x = 44 \quad \sum\left(\frac{d}{2} - 6\right) = 44 \quad \frac{1}{2}\sum d - \sum_1^8 6 = 44$

$\frac{1}{2}\sum d - 6 \times 8 = 44 \quad \Sigma d = 184$

$S_{dd} = \sum d^2 - \frac{(\sum d)^2}{n} = 4592 - \frac{184^2}{8} = 360$

$r = \frac{S_{dm}}{\sqrt{S_{dd}S_{mm}}} = \frac{90.6}{\sqrt{360 \times 23.66}} = 0.982$ (3 s.f.)