4) Correlation

4.1) Correlation
4.2) Linear regression

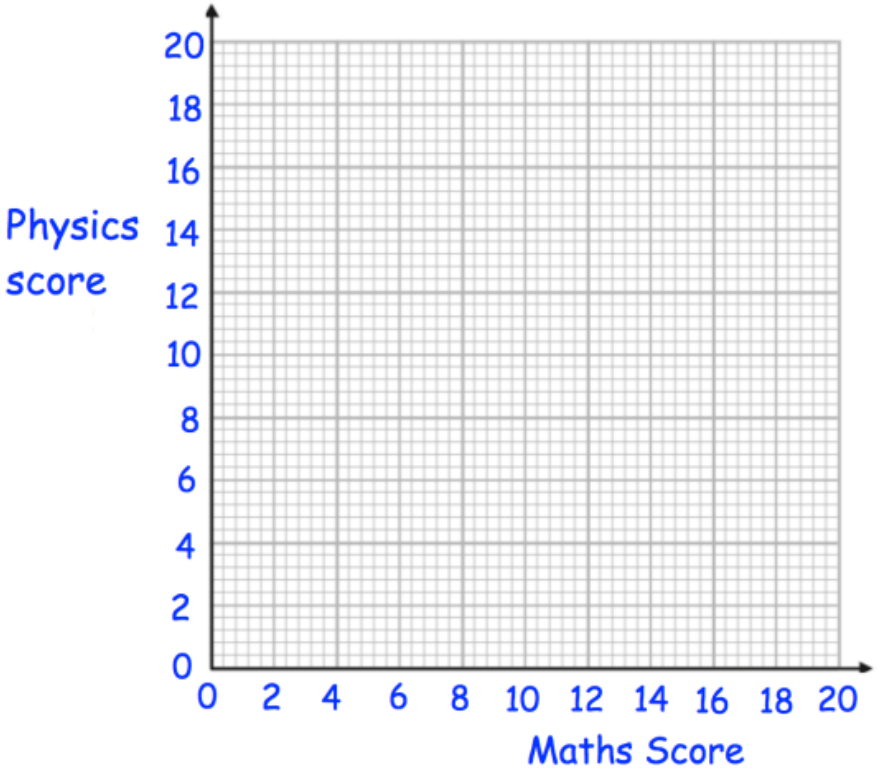
4.1) Correlation

Chapter CONTENTS

Worked example

Plot a scatter diagram to represent this data:

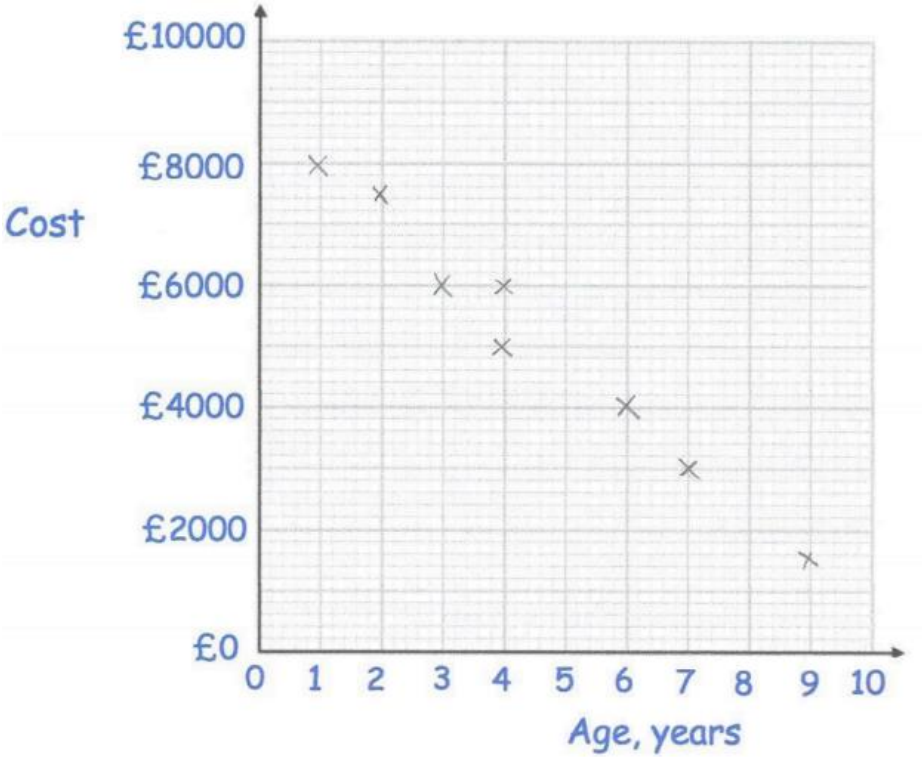
Maths score (x)	9	13	6	18	11	4	15	10
Physics score (y)	10	13	5	20	8	5	12	14



Your turn

Plot a scatter diagram to represent this data:

Age, years (x)	4	7	2	4	1	9	3	6
Cost, £ (y)	6000	3000	7500	5000	8000	1500	6000	4000

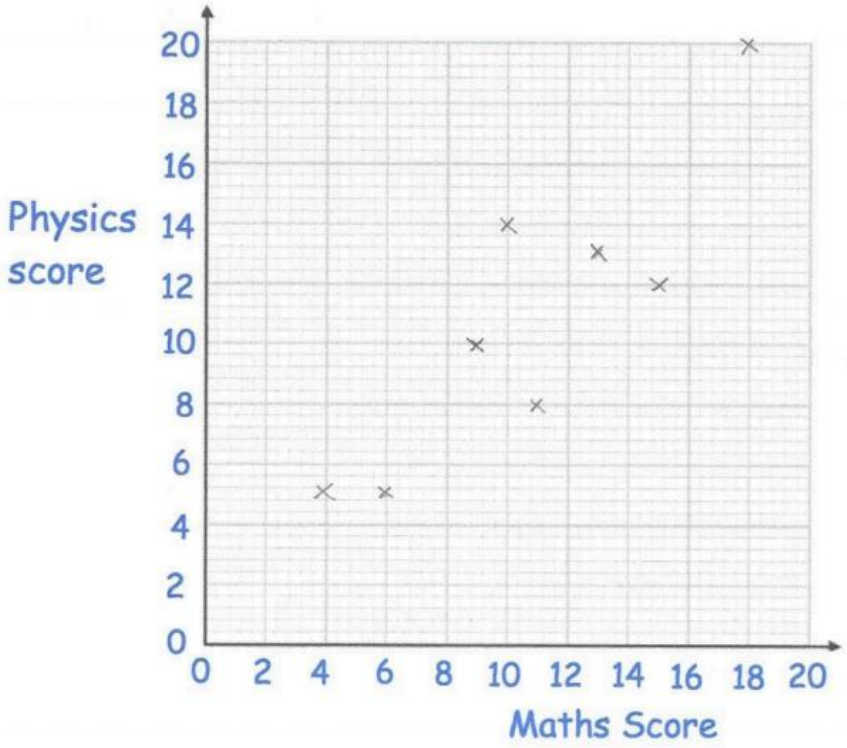


Worked example

Use the scatter diagram to:

a) Describe the correlation

b) Interpret the correlation

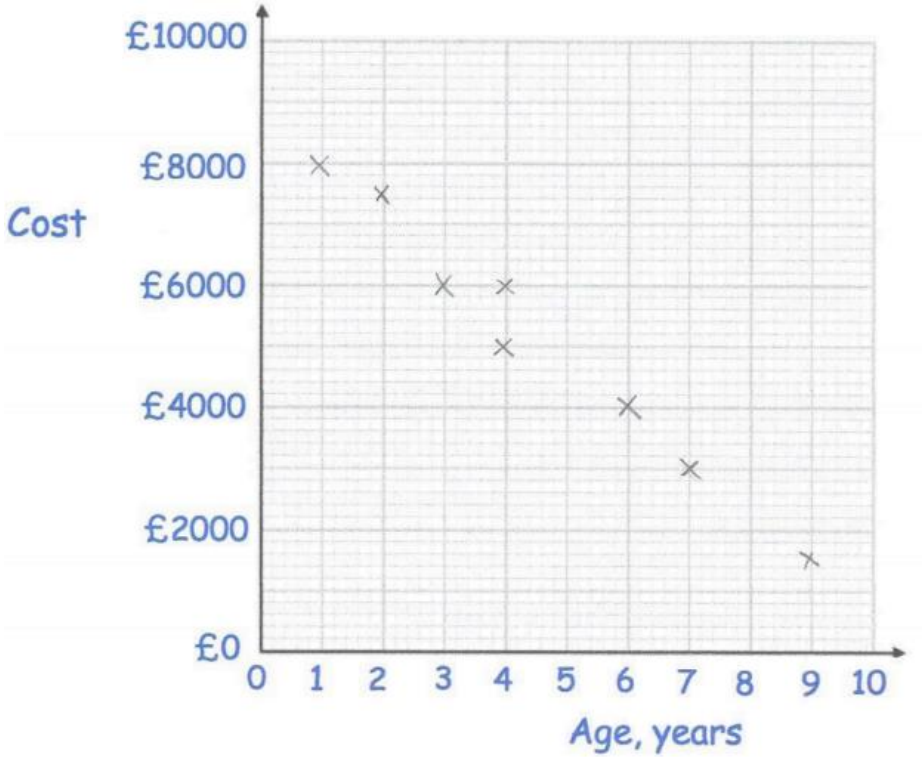


Your turn

Use the scatter diagram to:

a) Describe the correlation

b) Interpret the correlation



a) Strong negative correlation

b) As the age (of an object) increases, its cost decreases

Worked example

A student was interested to see if there as a relationship between the value of a house and the speed of its internet connection.
A scatter diagram was drawn with a weak negative correlation.
He says his data supports the conclusion that a slower internet connection reduces the value of a house.
Give one reason why his conclusion may not be valid.

Your turn

A student was interested to see if there was a relationship between what people earn and the age which they left education or training.
A scatter diagram was drawn with a weak negative correlation.
She says her data supports the conclusion that more education causes people to earn a lower hourly rate of pay.
Give one reason why her conclusion might not be valid.

- Respondents who left education later would have significantly less work experience than those who left education earlier. This could be the cause of the reduced income shown in her results.
- Small opportunistic sample used

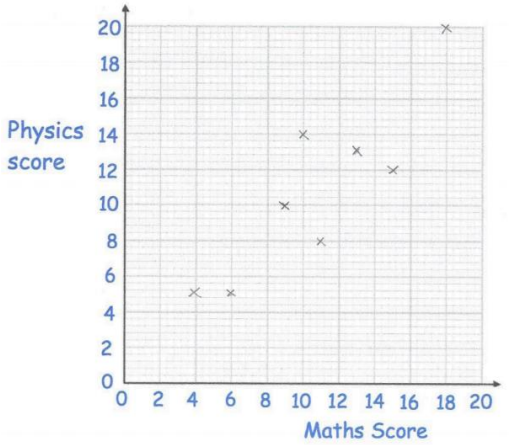
4.2) Linear regression

Chapter CONTENTS

Worked example

The following data was plotted on a scatter diagram.

Maths score (x)	9	13	6	18	11	4	15	10
Physics score (y)	10	13	5	20	8	5	12	14



The equation of the regression line of y on x for these 8 students is $y = 0.54 + 0.96x$

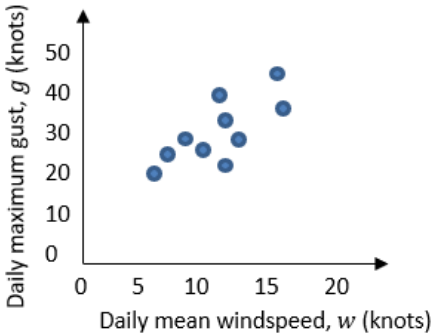
a) Draw the regression line on your diagram.
b) Give an interpretation of the value of the gradient.
c) Justify the use of a linear regression line in this instance.

Your turn

From the large data set, the daily mean windspeed, w knots, and the daily maximum gust, g knots, were recorded for the first 15 days in May in Camborne in 2015.

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 was plotted on a scatter diagram.



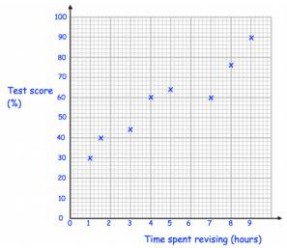
The equation of the regression line of g on w for these 15 days is $g = 7.23 + 1.82w$

a) Draw the regression line on your diagram.
b) Give an interpretation of the value of the gradient.
c) Justify the use of a linear regression line in this instance.

a) Drawn
b) Gradient $= 1.82$. If the daily mean windspeed increases by 10 knots the daily maximum gust increases by approximately 18 knots.
c) The correlation suggests that there is a linear relationship between g and w so a linear regression line is a suitable model.

Worked example

The test score, $y\%$, and time spent revising, x hours, for a random sample of eight new students are recorded.
The scatter graph shows the results.

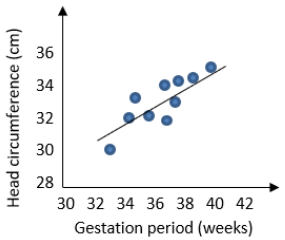


The equation of the regression line of y on x is $y = 27.9 + 6.25x$.
The regression equation is used to estimate the test score of a student who revised for 5.5 hours and a student who revised for 15.5 hours.
(a) Comment on the reliability of these estimates.

A teacher wants to estimate the time spent revising for a student who achieved a test score of 35%.
(b) Explain why the regression equation given above is not suitable for this estimate.

Your turn

The head circumference, y cm, and gestation period, x weeks, for a random sample of eight new born babies at a clinic are recorded.
The scatter graph 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.

The prediction for 39 weeks is within the range of the data (interpolation) so is more likely to be correct.
The prediction for 30 weeks is outside the range of the data (extrapolation) so is less likely to be accurate.

A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6cm.
(b) Explain why the regression equation given above is not suitable for this estimate.

The independent variable in this model is the gestation period, x. You should not use this model to predict a value of x for a given value of y.