KING EDWARD VI
HANDSWORTH GRAMMAR
SCHOOL FOR BOYS

KING EDWARD VI
ACADEMY TRUST
BIRMINGHAM

Year 12
Statistics
4 Correlation

HGS Maths

Dr Frost Course

Name: _____

Class: _____

Contents

Prior knowledge check

1 The table shows the scores out of 10 on a maths test and on a physics test for 7 students.

Maths	6	7	7	8	9	9	10
Physics	9	7	6	7	5	4	5

Show this information on a scatter diagram.

← GCSE Mathematics

2 A straight line has equation $y = 0.34 - 3.21x$.
Write down

a the gradient of the line

b the y-intercept of the line

← GCSE Mathematics

4.1 Correlation

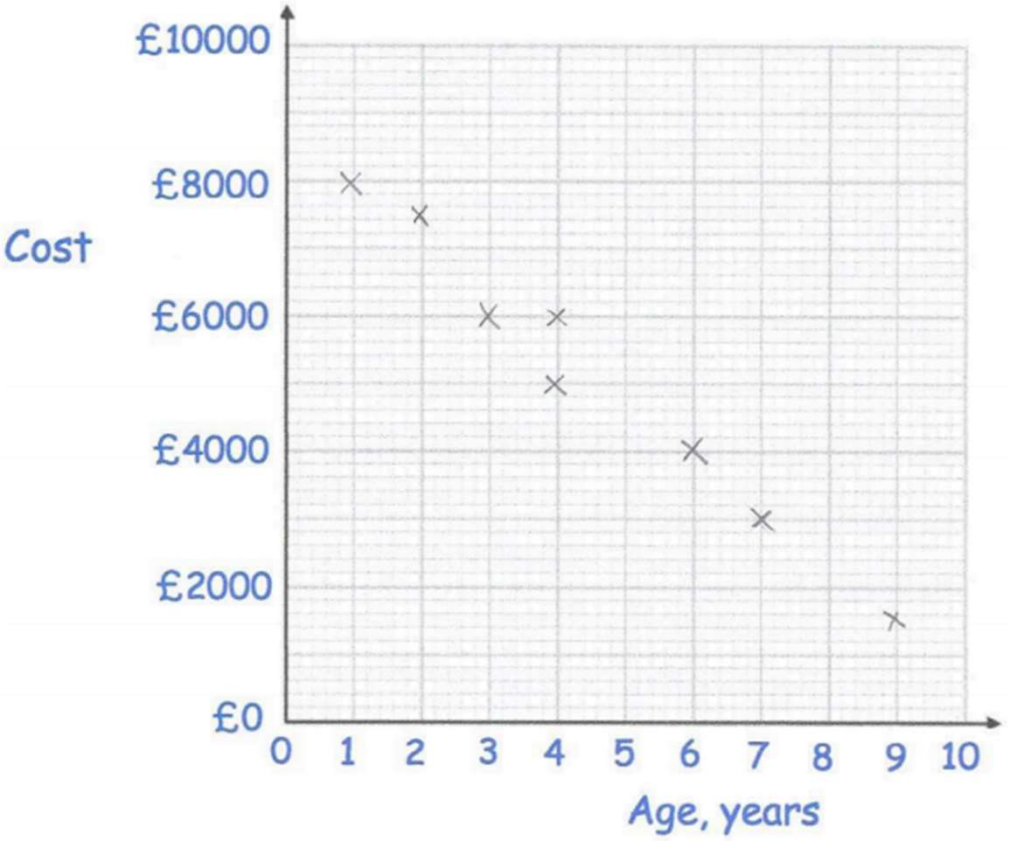
Notes

Worked Example

Use the scatter diagram to:

a) Describe the correlation

b) Interpret the correlation



Worked Example

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 this data.

b Describe the correlation between distance and population density.

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

Worked Example

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.

Exam-Style Question 1

The table shows the annual salary, s and journey time, t, in minutes to travel to work of seven office workers in a city.

t (minutes)	25	50	30	45	18	29	33
s (£1000s)	35	22	17	49	25	37	43

a Draw a scatter diagram to represent the data. **(3)**

b Describe the correlation between annual salary and journey time to work. **(1)**

c Do you think there is a causal relationship between the annual salary and journey time to work? Explain your reasoning. **(1)**

T.61: Ex 4A all Qs, P.26: all Qs

4.2 Linear Regression

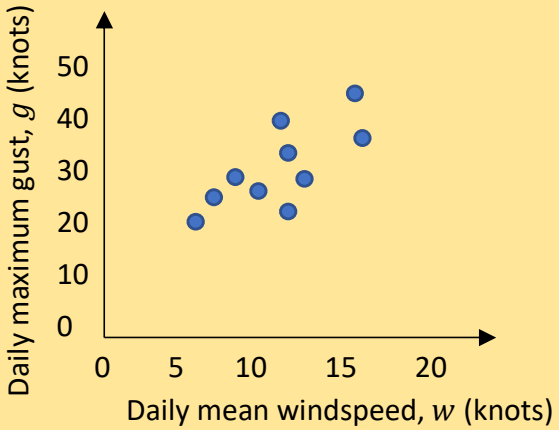
Notes

Example

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.



(a) Describe the correlation between daily mean windspeed 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 There is a strong positive correlation between daily mean windspeed and daily maximum gust.

b 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.

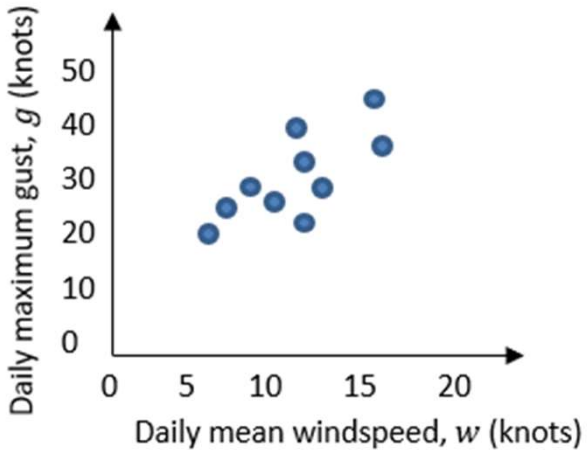
The stronger the (linear) correlation, the more suitable a linear regression line is.

Worked Example

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 $g = 7.23 + 1.82w$

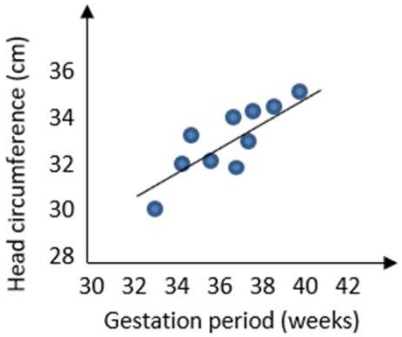
a) Draw the regression line on your diagram.

b) Give an interpretation of the value of the gradient.

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

Worked Example

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.

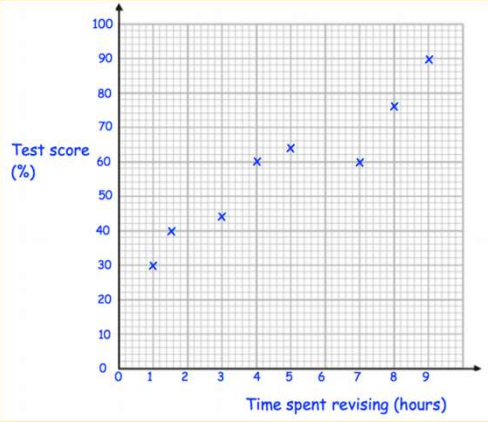
A nurse wants to estimate the gestation period for a baby born with a head circumference of 31.6cm.

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

T.65: Ex 4B all Qs, P.28: all Qs

Your Turn

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.



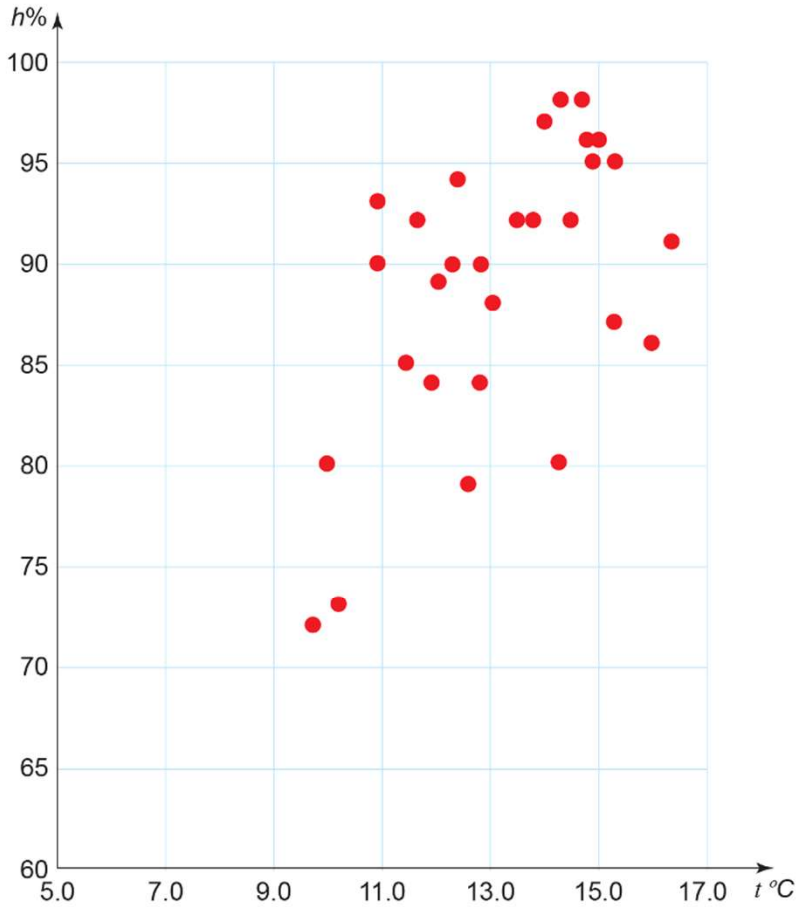
The equation of the regression line of y on x is $y = 27.9 + 6.25x$.
The regression equation is used to estimate the test score of a student who revised for 5.5 hours and a student who revised for 15.5 hours.

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.

Exam-Style Question 2



The scatter graph shows the relationship between temperature and humidity, recorded at noon, in a city in February.

Comment on the suitability of the graph.

Exam-Style Question 3

The table shows the daily total rainfall, r mm, and daily total hours of sunshine, s, in Tenby for a random sample of 9 days in February 1998.

r	17.4	3.8	0	3.1	0.7	2.3	1.3	2.7	3.8
s	5.2	1.6	7.9	6.2	6.6	5.8	3.1	5.4	3.2

The median and quartiles for the rainfall data are $Q_1 = 1$, $Q_2 = 2.7$, $Q_3 = 3.8$
An outlier is defined as a value which lies either 1.5 × the interquartile range above the upper quartile or 1.5 × the interquartile range below the lower quartile.

a Show that $r = 17.4$ is an outlier. **(3)**

b Exclude this day's readings and draw a scatter diagram to represent the data for the remaining 8 days. **(3)**

c Describe the correlation between rainfall and hours of sunshine. **(1)**

Exam-Style Question 4

The table shows the daily total rainfall, r mm, in Tenby for a random sample of 11 days in April 1998.

r	15	14	13	12	12	28	13	10	6	0	2

An outlier is defined as a value which lies either 1.5 × the interquartile range above the upper quartile or 1.5 × the interquartile range below the lower quartile.

Arthur claims $r = 28$ is an outlier.

Give a reason why you might

a include this day's reading

b exclude this day's reading.

Past Paper Questions

AS 2019

Correlation

1. A sixth form college has 84 students in Year 12 and 56 students in Year 13

The head teacher selects a stratified sample of 40 students, stratified by year group.

(a) Describe how this sample could be taken.

(3)

The head teacher is investigating the relationship between the amount of sleep, s hours, that each student had the night before they took an aptitude test and their performance in the test, p marks.
For the sample of 40 students, he finds the equation of the regression line of p on s to be

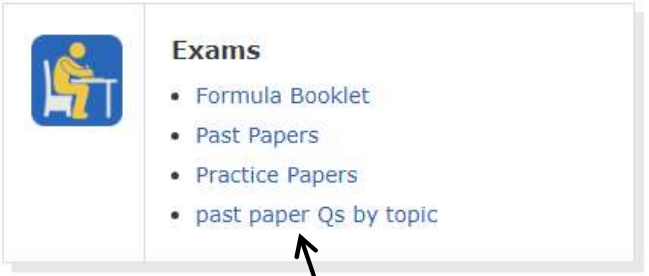
$$p = 26.1 + 5.60s$$

(b) With reference to this equation, describe the effect that an extra 0.5 hours of sleep may have, on average, on a student's performance in the aptitude test.

(1)

(c) Describe one limitation of this regression model.

(1)



Past paper practice by topic. Both new and old specification can be found via this link on hgsmaths.com

Part	Working or answer as expected	Mark	Notes
(a)	Label Year 12 as 1–84 and Year 13 as 85–140	B1	This mark is given for a suitable labelled list for each year group
	Use random number to select a …	B1	This mark is given for the use of random number to select students
	…simple random sample of 24 Year 12s and 16 Year 13s	B1	This mark is given for 24 Year 12s and 16 Year 13s
(b)	$5.60 \times 0.5 = 2.8$	B1	This mark is given for the value for the regression coefficient
	Increase by 2.8 marks		
(c)	For example: The model suggests that the longer students sleep, the better they will perform in the test The best performance is predicted for the students who never wake up The model is only valid between 0 and 24 hours (the range of the data)	B1	This mark is given for a valid limitation of the model

Summary of Key Points

1 **Bivariate data** is data which has pairs of values for two variables.

2 **Correlation** describes the nature of the linear relationship between two variables.

3 When two variables are correlated, you need to consider the context of the question and use your common sense to determine whether they have a causal relationship.

4 The **regression line** of y on x is written in the form $y = a + bx$.

5 The coefficient b tells you the change in y for each unit change in x.

- If the data is positively correlated, b will be positive.
- If the data is negatively correlated, b will be negative.

6 You should only use the regression line to make predictions for values of the dependent variable that are within the range of the given data.

T.66: mixed ex. P.29: BSG

Large Data Set

Large data set

You will need access to the large data set and spreadsheet software to answer these questions.

1 Investigate the relationship between daily mean windspeed, w, and daily maximum gust, g, in Leeming in 2015.

a Draw a scatter diagram of w against g for the entire data set for Leeming in 2015.

b Describe the correlation shown.

c Comment on whether there is likely to be a causal relationship between mean windspeed and maximum gust.

The equation of the regression line of g on w is given by $g = 4.97 + 2.15w$.

d Use the equation of the regression line to predict the maximum gust on a day when the mean windspeed is:

i 0.5 knots **ii** 5 knots **iii** 12 knots **iv** 40 knots.

e Comment on the accuracy of each prediction in part **d**.

f Calculate the equation of the regression line of w on g, and use it to predict the mean windspeed on a day when the maximum gust was 30 knots.

2 Use a similar approach to investigate the daily total sunshine and daily mean total cloud cover in Heathrow in 1987.

a Use a regression model to suggest values for the missing total sunshine data in the first half of May.

b Do you think there is a causal relationship between these two variables? Give a reason for your answer.

Hint You can use the SLOPE and INTERCEPT functions in some spreadsheets to find the values of a and b in a regression equation.