

QQQ – Statistics Year 1 – Chapter 4 – Correlation (Version 1)

Total Marks: 23

(23 = Platinum, 21 = Gold, 18 = Silver, 16 = Bronze)

1. Manny is studying the price and number of pages of a random sample of books.

He calculates the value of the product moment correlation coefficient between the price and number of pages in each book as 1.05

Which of the following best describes the value 1.05?

Tick (✓) **one** box.

[1 mark]

- definitely correct
- probably correct
- probably incorrect
- definitely incorrect

(Total 1 marks)

AQA June 2019 Paper 2 Q12

2. A company is introducing a job evaluation scheme. Points (x) will be awarded to each job based on the qualifications and skills needed and the level of responsibility. Pay (£ y) will then be allocated to each job according to the number of points awarded.

Before the scheme is introduced, a random sample of 8 employees was taken and the linear regression equation of pay on points was $y = 4.5x - 47$

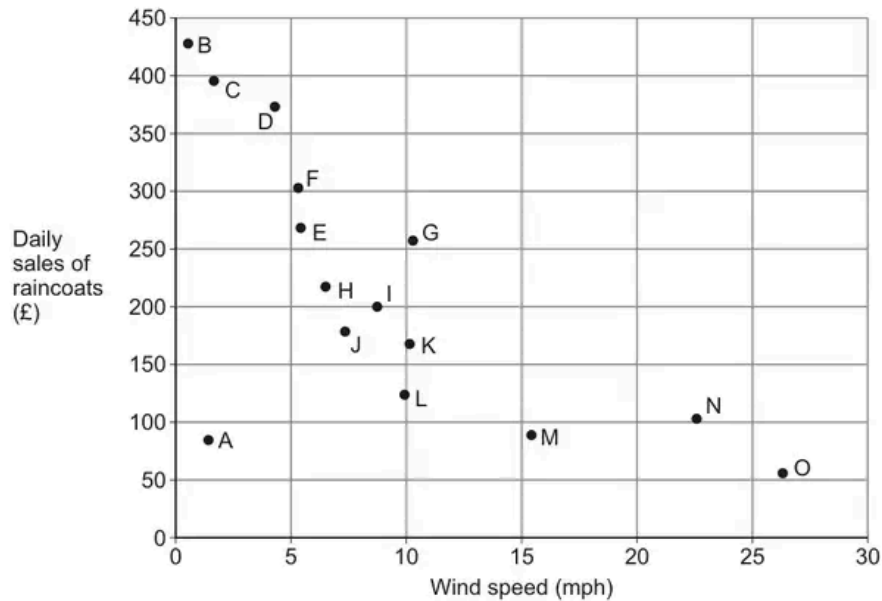
- (a) Describe the correlation between points and pay. **(1)**
- (b) Give an interpretation of the gradient of this regression line. **(1)**
- (c) Explain why this model might not be appropriate for all jobs in the company. **(1)**

(Total 3 marks)

Edexcel June 2018 Statistics Q1

3. The owner of a shop recorded the average daily wind speed in mph and the total sales of raincoats in pounds over a period of 6 months.

The data from a random selection of 15 days from the 6-month period is shown in the scatter diagram.



One of the points on the scatter diagram is an outlier.

- (a) State the letter from the diagram which corresponds to this outlier. **[1 mark]**
- (b) State the effect on the numerical value of the correlation coefficient if the outlier was removed from the data. **[1 mark]**
- (c) Describe, in context, the correlation indicated by the scatter diagram. **[1 mark]**
- (d) The owner of the shop claims that the scatter diagram proves that windier days cause lower sales of raincoats.

Comment on the validity of this claim.

[1 mark]

(Total 4 marks)

4. Fred and Nadine are investigating whether there is a linear relationship between Daily Mean Pressure, p hPa, and Daily Mean Air Temperature, t °C, in Beijing using the 2015 data from the large data set.

Fred randomly selects one month from the data set and draws the scatter diagram in Figure 1 using the data from that month.

The scale has been left off the horizontal axis.

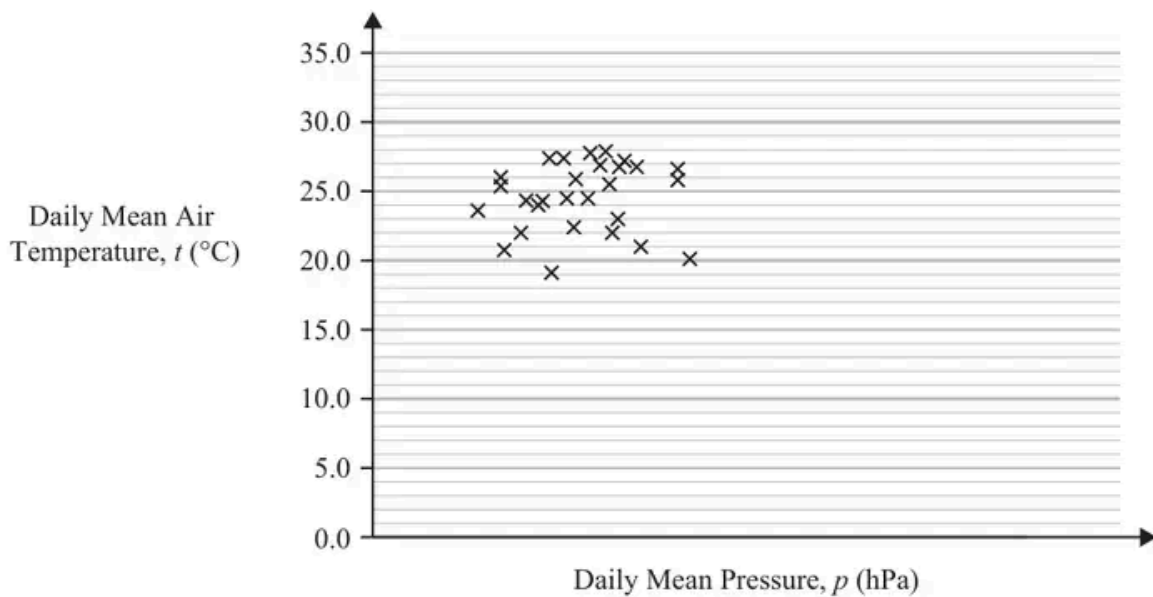


Figure 1

(a) Describe the correlation shown in Figure 1. (1)

Nadine chooses to use all of the data for Beijing from 2015 and draws the scatter diagram in Figure 2.

She uses the same scales as Fred.

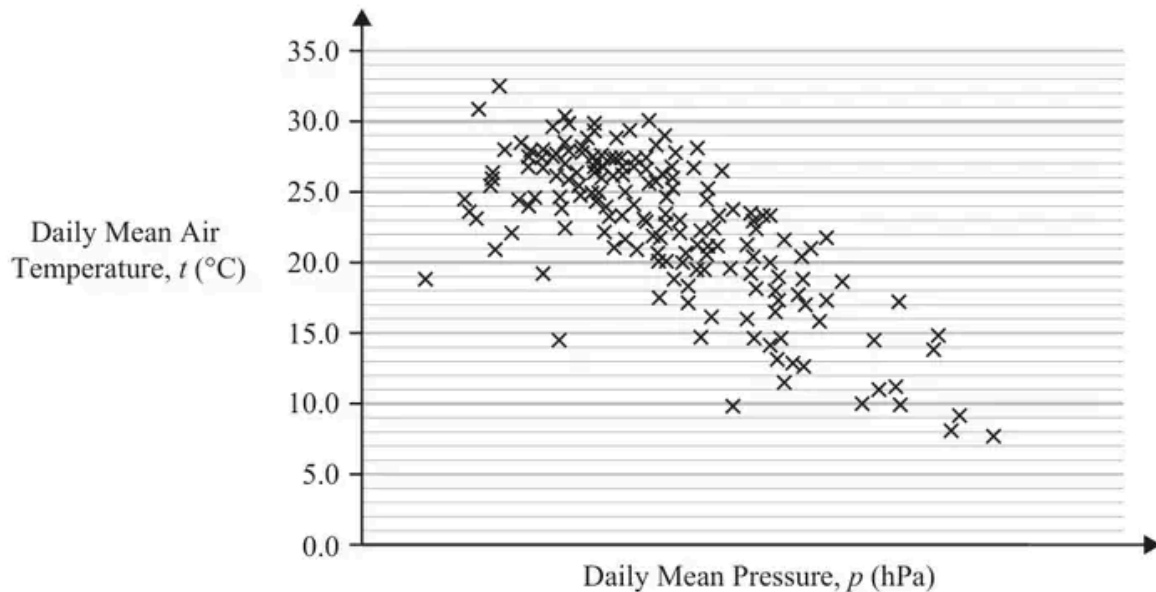


Figure 2

- (b) Explain, in context, what Nadine can infer about the relationship between p and t using the information shown in Figure 2. (1)
- (c) Using your knowledge of the large data set, state a value of p for which interpolation can be used with Figure 2 to predict a value of t . (1)
- (d) Using your knowledge of the large data set, explain why it is not meaningful to look for a linear relationship between Daily Mean Wind Speed (Beaufort Conversion) and Daily Mean Air Temperature in Beijing in 2015. (1)

(Total 4 marks)

5. The relationship between two variables p and t is modelled by the regression line with equation

$$p = 22 - 1.1 t$$

The model is based on observations of the independent variable, t , between 1 and 10

(a) Describe the correlation between p and t implied by this model. (1)

Given that p is measured in centimetres and t is measured in days,

(b) state the units of the gradient of the regression line. (1)

Using the model,

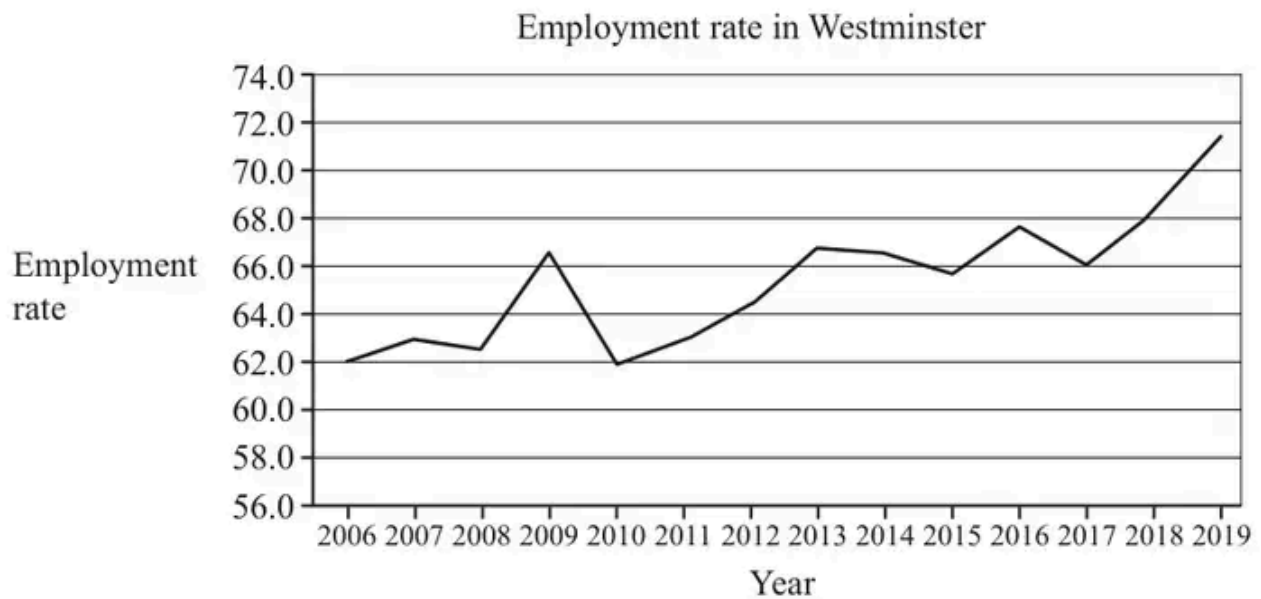
(c) calculate the change in p over a 3-day period. (2)

Tisam uses this model to estimate the value of p when $t = 19$

(d) Comment, giving a reason, on the reliability of this estimate. (1)

(Total 5 marks)

6. The pre-release material contains information about employment rates in London boroughs. The graph shows employment rates for Westminster between 2006 and 2019.



A local politician stated that the diagram shows that more than 60% of seventy-year-olds were in employment throughout the period from 2006 to 2019.

- (a) Use your knowledge of the pre-release material to explain whether there is any evidence to support this statement. [1]

In order to estimate the employment rate in 2020, two different models were proposed using the LINEST function in a spreadsheet.

Model 1 (using all the data from 2006 onwards)

$$Y = 0.549x - 1040,$$

Model 2 (using data from 2017 onwards)

$$Y = 2.65x - 5280,$$

where Y = employment rate and x = calendar year.

It was subsequently found that the employment rate in Westminster in 2020 was 68.4%.

- (b) Determine which of the two models provided the better estimate for the employment rate in Westminster in 2020. [3]
- (c) Use your knowledge of the pre-release material to explain whether it would be appropriate to use either model to estimate the employment rate in 2020 in other London boroughs. [1]
- (d) What does model 2 predict for employment rates in Westminster in the long term? [1]

(Total 6 marks)

OCR MEI June 2022 Paper 2 Q6

Solutions

1.

Scheme	Marks	AO
Ticks correct box	B1	3.2b
	(1 mark)	

Typical solution

definitely incorrect

2.

Mark scheme (a)

Scheme	Marks	AO
Positive (correlation)	B1	1.2
	(1)	

Notes

B1 for “positive”.

Allow an interpretation e.g. “as points increase pay increases” is B1

Read whole answer: contradictory comments such as “positive correlation, as points increase pay decreases” scores B0

Mark scheme (b)

Scheme	Marks	AO
Every extra point gives £4.5(0) more on pay (o.e.)	B1	3.4
	(1)	

Notes

B1 for any correct comment conveying idea of £s per point and including a correct value; must have idea of rate. Can condone missing £ sign. Accept 4.5

e.g. “every 10 points earns an extra (or increase) of £45” is B1

BUT “every point earns £4.5(0)” is B0 *doesn't have idea of rate*

Mark scheme (c)

Scheme	Marks	AO
e.g. For points < 11 it would give pay < 0 which is ridiculous	B1	2.4
	(1)	
	(3 marks)	

Notes

B1 for a suitable comment mentioning “points” or “pay” (o.e. e.g. “amount”) or commenting on “small sample” or “range of points” used to find line

The following examples would score B1

Can say that n points (for $n < 10.4$) would give negative pay so not suitable

Any comment suggesting that some jobs would end up with negative pay.

Don't know the range of points used to find the regression line

A small sample of size 8 may not be representative to cover all jobs

B0 for a focus on “qualifications” or “hours” worked only

The following examples would score B0

Some jobs require no (or low) skills or qualifications (*need negative pay*)

3.

Mark scheme (a)

Scheme	Marks	AO
States A	B1	2.2a
	(1)	

Typical solution

A

Mark scheme (b)

Scheme	Marks	AO
Explains that the value would become more negative or strong negative OE Accept smaller/decrease	E1	2.4
	(1)	

Typical solution

The value would become more negative

Mark scheme (c)

Scheme	Marks	AO
Describes the correlation in context. Require ' sale(s) ' and ' wind speed ' Do not accept 'number' /'amount' Accept negative correlation (with ' sale(s) ' and ' wind speed ') Do not accept 'number'/'amount'	E1	2.2b
	(1)	

Typical solution

The higher the wind speed the lower the daily sales of raincoats

Mark scheme (d)

Scheme	Marks	AO
States claim is not valid OE, with a comment; E.g. • Scatter diagram/correlation cannot prove causality • Could be another factor for drop in sales • 15 days is a small sample from a six-month period Accept it may be valid but the scatter diagram does not prove it.	E1	2.2b
	(1)	
	(4 marks)	

Typical solution

This claim is not valid. The scatter diagram cannot prove causality, there could be another factor which is causing the fall in sales, e.g. amount of rainfall

4.

Mark scheme (a)

Scheme	Marks	AO
--------	-------	----

No (correlation)/weak (correlation)	B1	1.1b
	(1)	

Notes

B1: correct description of correlation (oe) (ignore reference to positive/negative) condone neutral

Mark scheme (b)

Scheme	Marks	AO
(Negative correlation...) As p(ressure) increases, t(emperature) decreases.	B1	2.2b
	(1)	

Notes

B1: correct inference, allow equivalent statements.

Negative correlation on its own is B0. Inversely proportional on its own is B0.

Mark scheme (c)

Scheme	Marks	AO
990 to 1040 (hPa)	B1	3.4 LDS
	(1)	

Notes

B1: an answer in the range 990 to 1040 inclusive (ignore units)

Mark scheme (d)

Scheme	Marks	AO
Daily mean wind speed (Beaufort) is a qualitative variable.	B1	2.4 LDS
	(1)	
	(4 marks)	

Notes

B1: correct explanation that in the LDS, wind speed (Beaufort) is qualitative/not quantitative
 Allow e.g. 'categorical', e.g. 'given in words', e.g. 'wind speed is (always) light'
 Do not allow 'not continuous' on its own.

5.

Mark scheme (a)

Scheme	Marks	AO
Negative (since gradient of regression line is negative)	B1	1.2
	(1)	

Notes

Answers may be written within the question.

B1 for stating “negative”.

Allow a correct interpretation e.g. as t increases then p decreases (o.e.) [ignore any values]

B0 for contradictory statements e.g. “negative correlation since as t increases p increases”

Mark scheme (b)

Scheme	Marks	AO
cm/day (o.e. e.g. cm day^{-1})	B1	2.2a
	(1)	

Notes

B1 for a correct description of the units (allow fraction, /, or “per” and allow “d” for “day”)

Mark scheme (c)

Scheme	Marks	AO
$3 \times [\pm] 1.1$	M1	3.4
= decrease of 3.3 [cm]	A1	1.1b
	(2)	

Notes

M1 for attempt at a calculation (allow use of $t = x$ and $t = x + 3$ followed by subtraction that should lead to 3.3)

A1 for correct description must include word “decrease” (o.e.) and value “3.3”

Just seeing: $22 - 1.1 \times 3 = 18.7$ is M0A0 BUT going on to subtract 18.7 from 22 scores M1

Reaching 3.3 and stating “decrease” or “reduced” (o.e.) will score the A1 too

An answer of -3.3 without a word describing decrease (o.e.) will just score M1A0

Mark scheme (d)

Scheme	Marks	AO
19 is (well) outside the range [1, 10] <u>or</u> involves extrapolation (o.e.) so (possibly) unreliable/ inaccurate (o.e.)	B1	2.4
	(1)	
	(5 marks)	

Notes

B1 for stating “unreliable” (o.e.) **and** giving a suitable reason based on idea of extrapolation

Must have **both** statement about reliability **and** suitable reason e.g. $t = 19$ is too big or (Model is based on) t between 1 and 10 (only) [since this implies $t = 19$ is too big]

Allow e.g. (model) “may not work” because of “extrapolation”

Just saying “no” since “extrapolation” is B0 but “unreliable”(o.e.) since “extrapolation” is B1

6.

Mark scheme (a)

Scheme	Marks	AO
No evidence to support this statement as information in LDS is for 16 – 64 year olds	B1	2.4
	[1]	

Notes

B1: 0e This age group not in data set. Must refer to data set

LDS advantage

Mark scheme (b)

Scheme	Marks	AO
Model 1: 68.98%	*B1	3.4

Model 2: 73%	*B1	1.1
Model 1 gives the best approximation since 68.98 is closer to true value	DB1	3.5a
	[3]	

Notes

***B1 *B1:** Reverse calculations

Model 1 2019 (2018.9)

Model 2 2018 (2018.3)

DB1: Model 1 sufficient or unambiguous indication

Mark scheme (c)

Scheme	Marks	AO
Not appropriate since employment trends vary across the boroughs	B1	2.4
	[1]	

Notes

B1: Oe

Westminster employment trends differs to other boroughs
employment specific only to Westminster

LDS advantage

Mark scheme (d)

Scheme	Marks	AO
eg Employment rates will increase forever or eg Employment rates will exceed 100%	B1	3.5b
	[1]	

Notes

B1: Oe

not just increase
not exponential
Constant growth

Rise over time

Increase every year

No limit/maximum.