Y2S1 XMQs and MS

(Total: 51 marks)

1. P3_2018 Q2 . 7 marks - Y2S1 Regression, correlation and hypothesis testing
2. P3_Sample Q2 . 6 marks - Y1S4 Correlation
3. P3_Specimen Q2 . 7 marks - Y1S4 Correlation
4. P31_2019 Q3 . 9 marks - Y2S1 Regression, correlation and hypothesis testing
5. P31_2020 Q2 . 7 marks - Y2S1 Regression, correlation and hypothesis testing
6. P31_2021 Q2 . 6 marks - Y2S1 Regression, correlation and hypothesis testing
7. P31_2022 Q6 . 9 marks - Y2S1 Regression, correlation and hypothesis testing

2. Tessa owns a small clothes shop in a seaside town. She records the weekly sales figures, £w, and the average weekly temperature, t °C, for 8 weeks during the summer.
The product moment correlation coefficient for these data is −0.915

(a) Stating your hypotheses clearly and using a 5% level of significance, test whether or not the correlation between sales figures and average weekly temperature is negative.

(3)

(b) Suggest a possible reason for this correlation.

(1)

Tessa suggests that a linear regression model could be used to model these data.

(c) State, giving a reason, whether or not the correlation coefficient is consistent with Tessa's suggestion.

(1)

(d) State, giving a reason, which variable would be the explanatory variable.

(1)

Tessa calculated the linear regression equation as $w = 10\,755 - 171t$

(e) Give an interpretation of the gradient of this regression equation.

(1)

Qu 2	Scheme	Marks	AO
(a)	$H_0 : \rho = 0$ $H_1 : \rho < 0$	B1	2.5
	Critical value: -0.6215 (Allow any cv in range $0.5 < \|cv\| < 0.75$)	M1	1.1a
	$r < -0.6215$ so significant result and there is evidence of a negative correlation between *w* and *t*	A1	2.2b
		(3)	
(b)	e.g. As temperature increases people spend more time on the beach and less time shopping (o.e.)	B1	2.4
		(1)	
(c)	Since *r* is close to -1 , it **is** consistent with the suggestion	B1	2.4
		(1)	
(d)	*t* will be the explanatory variable since sales are likely to depend on the temperature	B1	2.4
		(1)	
(e)	Every degree rise in temperature leads to a drop in weekly earnings of £171	B1	3.4
		(1)	
		(7 marks)	

	Notes
(a)	B1 for both hypotheses in terms of ρ M1 for the critical value: sight of $\pm$ 0.6215 or any cv such that $0.5 < \|cv\| < 0.75$ A1 must reject H_0 on basis of comparing -0.915 with -0.6215 (if $-0.915 < 0.6215$ is seen then A0 but may use $\|r\|$ o.e. which is fine) and mention "negative", "correlation/relationship" and at least "*w*" and "*t*"
(b)	B1 for a suitable reason to explain negative correlation using the context given. e.g. "As temperature drops people are more likely to go shopping (than to the beach)" e.g. "As temperature increases people will be outside rather than in shops" A mere description in context of negative correlation is B0 SO e.g. "As temperature increases people don't want to go shopping/buy clothes" is B0 e.g. "Less clothes needed as temp increases" is B0
(c)	B1 for a suitable reason e.g. "strong"/"significant"/"near perfect" "correlation", $\|r\|$ close to 1 and saying it is consistent with the suggestion. Allow "yes" followed by the reason.
(d)	B1 For identifying *t* and giving a suitable reason. Need idea that "*w* depends on *t*" or "*w* responds to *t*" or "*t* affects *w*" (o.e.) Allow *t* (temperature) affects the other variable etc Just saying " *t* is the independent variable" or "*t* explains change in *w*" is B0 N. B. Suggesting causation is B0 e.g. "*t* causes *w* to decrease"
(e)	B1 for a description that conveys the idea of rate per degree Celsius. Must have 171, condone missing "£" sign.

2. A meteorologist believes that there is a relationship between the daily mean windspeed, w kn, and the daily mean temperature, t °C. A random sample of 9 consecutive days is taken from past records from a town in the UK in July and the relevant data is given in the table below.

t	13.3	16.2	15.7	16.6	16.3	16.4	19.3	17.1	13.2
w	7	11	8	11	13	8	15	10	11

The meteorologist calculated the product moment correlation coefficient for the 9 days and obtained $r = 0.609$

(a) Explain why a linear regression model based on these data is unreliable on a day when the mean temperature is 24 °C

(1)

(b) State what is measured by the product moment correlation coefficient.

(1)

(c) Stating your hypotheses clearly test, at the 5% significance level, whether or not the product moment correlation coefficient for the population is greater than zero.

(3)

Using the same 9 days a location from the large data set gave $\overline{t} = 27.2$ and $\overline{w} = 3.5$

(d) Using your knowledge of the large data set, suggest, giving your reason, the location that gave rise to these statistics.

(1)

Question	Scheme	Marks	AOs
2(a)	e.g. It requires extrapolation so will be unreliable (o.e.)	B1	1.2
		(1)	
(b)	e.g. Linear association between w and t	B1	1.2
		(1)	
(c)	H_0: $\rho = 0$ H_1: $\rho > 0$	B1	2.5
	Critical value 0.5822	M1	1.1a
	Reject H_0		
	There is evidence that the product moment correlation coefficient is greater than 0	A1	2.2b
		(3)	
(d)	Higher $\bar{t}$ suggests overseas and not Perth…lower wind speed so perhaps not close to the sea so suggest **Beijing**	B1	2.4
		(1)	
			(6 marks)

Notes:

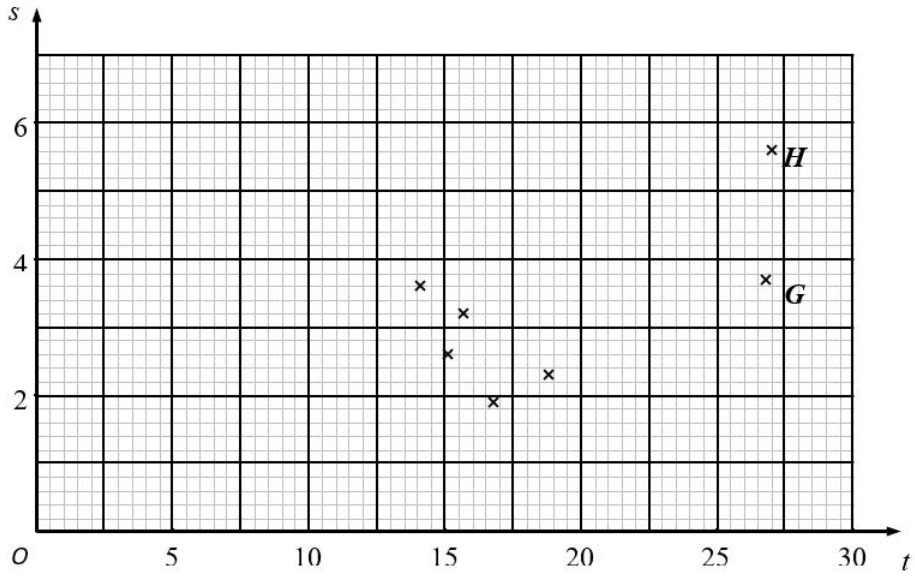
(a)
B1: for a correct statement (unreliable) with a suitable reason

(b)
B1: for a correct statement

(c)
B1: for both hypotheses in terms of ρ
M1: for selecting a suitable 5% critical value compatible with their H_1
A1: for a correct conclusion stated

(d)
B1: for suggesting Beijing with some supporting reason based on t or w
Allow Jacksonville with a reason based just on higher $\bar{t}$

2. A researcher believes that there is a linear relationship between daily mean temperature and daily total rainfall. The 7 places in the northern hemisphere from the large data set are used. The mean of the daily mean temperatures, t °C, and the mean of the daily total rainfall, s mm, for the month of July in 2015 are shown on the scatter diagram below.



(a) With reference to the scatter diagram, explain why a linear regression model may not be suitable for the relationship between t and s.

(1)

The researcher calculated the product moment correlation coefficient for the 7 places and obtained $r = 0.658$

(b) Stating your hypotheses clearly, test at the 10% level of significance, whether or not the product moment correlation coefficient for the population is greater than zero.

(3)

(c) Using your knowledge of the large data set, suggest the names of the 2 places labelled G and H.

(1)

(d) Using your knowledge from the large data set, and with reference to the locations of the 2 places labelled G and H, give a reason why these places have the highest temperatures in July.

(1)

(e) Suggest how you could make better use of the large data set to investigate the relationship between daily mean temperature and daily total rainfall.

(1)

Question	Scheme	Marks	AOs
2(a)	Not suitable with a correct reason eg the points do not lie close to a straight line. there appear to be two populations if *G* and *H* were removed it appears to be a negative correlation	B1	1.2
		(1)	
(b)	$H_0 : \rho = 0 \quad H_1 : \rho > 0$	B1	2.5
	Critical value 0.5509	M1	1.1a
	Reject H_0		
	There is evidence that pmcc is greater than zero	A1	2.2b
		(3)	
(c)	Beijing and Jacksonville	B1	2.2a
		(1)	
(d)	Beijing and Jacksonville are the closest to the equator	B1	2.4
		(1)	
(e)	Use data from one place.	B1	2.4
		(1)	
			(7 marks)

Notes:

(a) B1: for a correct statement using the data in the table

(b) B1: for both hypotheses in terms of ρ

M1: for selecting a suitable critical value compatible with their H_1

A1: for a correct conclusion stated

(c) B1: both Beijing and Jacksonville – they do not need to be attached to G and H correctly.

(d) B1: for the idea they are near the equator dependent only Beijing or Jacksonville being given in part(c)

3. Barbara is investigating the relationship between average income (GDP per capita), x US dollars, and average annual carbon dioxide (CO_2) emissions, y tonnes, for different countries.

She takes a random sample of 24 countries and finds the product moment correlation coefficient between average annual CO_2 emissions and average income to be 0.446

(a) Stating your hypotheses clearly, test, at the 5% level of significance, whether or not the product moment correlation coefficient for all countries is greater than zero.

(3)

Barbara believes that a non-linear model would be a better fit to the data.
She codes the data using the coding $m = \log_{10} x$ and $c = \log_{10} y$ and obtains the model $c = -1.82 + 0.89m$

The product moment correlation coefficient between c and m is found to be 0.882

(b) Explain how this value supports Barbara's belief.

(1)

(c) Show that the relationship between y and x can be written in the form $y = ax^n$ where a and n are constants to be found.

(5)

Question	Scheme		Marks	AOs
3(a)	$H_0 : \rho = 0$ $H_1 : \rho > 0$		B1	2.5
	Critical value 0.3438		M1	1.1a
	(0.446 > 0.3438) so there is evidence that the product moment correlation coefficient (pmcc) is greater than 0/there is positive correlation		A1	2.2b
			(3)	
(b)	The value is close(r) to 1 **or** there is strong(er) (positive) correlation		B1	2.4
			(1)	
(c)	$\log_{10} y = -1.82 + 0.89(\log_{10} x)$	$y = ax^n \rightarrow$ $\log_{10} y = \log_{10}(ax^n)$	M1	1.1b
	$y = 10^{-1.82+0.89(\log_{10} x)}$	$\log_{10} y = \log_{10} a + \log_{10} x^n$	M1	2.1
	$y = 10^{-1.82} \times 10^{0.89(\log_{10} x)}$ $[= 10^{-1.82} \times 10^{(\log_{10} x)^{0.89}}]$	$\log_{10} y = \log_{10} a + n\log_{10} x$ $[\log_{10} a = -1.82, n = 0.89]$	M1	1.1b
	$y = 0.015x^{0.89}$	$y = 0.015x^{0.89}$	A1A1	1.1b 1.1b
			(5)	
				(9 marks)

Notes

(a)

B1: for both hypotheses correct in terms of ρ

M1: for the critical value: sight of 0.3438 or any cv such that $0.25 < |cv| < 0.45$

A1: a comment suggesting a significant result/ H_0 is rejected on the basis of seeing +0.3438 **and** which mentions "pmcc/correlation/relationship" and "greater than 0/positive" (not just $\rho > 0$)

or an answer in context e.g. 'as "income"(o.e.) increases, "CO_2/emissions"(o.e.) increases'

A **contradictory** statement scores A0 e.g. 'Accept H_0, therefore positive correlation'

(b)

B1: for suitable reason e.g. *r* is close(r) to 1 **or** "strong(er)"/"near perfect" "correlation"

Do not allow 'association'

(c)

For both methods, once an M0 is scored, no further marks can be awarded and condone missing base 10 throughout

Method 1: (working to the model)

M1: Correct substitution for both *c* and *m* (may be implied by 2nd M1 mark)

M1: Making *y* the subject to give an equation in the form $y = 10^{a+b(\log_{10} x)}$ (may be implied by 3rd M1 mark)

M1: Correct multiplication to give an equation in the form $y = 10^a \times 10^{b(\log_{10} x)}$ (this line implies M1M1M1 provided no previous incorrect working seen)

Method 2: (working from the model)

M1: Taking the log of both sides (may be implied by 2nd M1 mark)

M1: Correct use of addition rule (may be implied by 3rd M1 mark)

M1: Correct multiplication of power (this line implies M1M1M1 provided no previous incorrect working seen)

A1: $n = 0.89$ **or** $a =$ awrt 0.015 **or** $y = ax^{0.89}$ **or** $y =$ awrt$0.015x^n$ (**dep on M3**)

A1: $n = 0.89$ **and** $a =$ awrt 0.015 / $y =$ awrt$0.015x^{0.89}$ (**dep on M3**)

do not award the final A1 if answer is given in an incorrect form e.g $y = 0.015 + x^{0.89}$

2. A random sample of 15 days is taken from the large data set for Perth in June and July 1987.

The scatter diagram in Figure 1 displays the values of two of the variables for these 15 days.

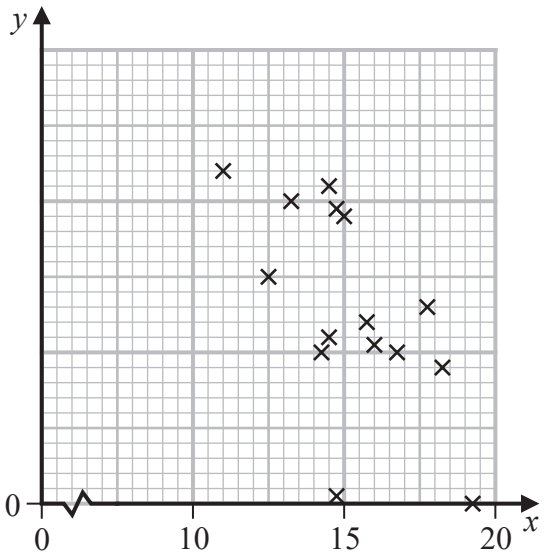


Figure 1

(a) Describe the correlation.

(1)

The variable on the x-axis is Daily Mean Temperature measured in °C.

(b) Using your knowledge of the large data set,

(i) suggest which variable is on the y-axis,

(ii) state the units that are used in the large data set for this variable.

(2)

Stav believes that there is a correlation between Daily Total Sunshine and Daily Maximum Relative Humidity at Heathrow.

He calculates the product moment correlation coefficient between these two variables for a random sample of 30 days and obtains $r = -0.377$

(c) Carry out a suitable test to investigate Stav's belief at a 5% level of significance. State clearly

- your hypotheses
- your critical value

(3)

On a random day at Heathrow the Daily Maximum Relative Humidity was 97%

(d) Comment on the number of hours of sunshine you would expect on that day, giving a reason for your answer.

(1)

Qu 2	Scheme	Marks	AO
(a)	Negative	B1 **(1)**	1.2
(b)(i) **(ii)**	Rainfall mm or Pressure hPa or Pascals or hectopascals or mb or millibars	B1 B1ft **(2)**	2.2b 1.1b
(c)	$H_0 : \rho = 0$ $H_1 : \rho \neq 0$ Critical value: $-0.361(0)$ $r < -0.3610$ so significant result and there is evidence of a correlation between Daily Total Sunshine and Daily Maximum Relative Humidity	B1 M1 A1 **(3)**	2.5 1.1b 2.2b
(d)	Humidity is high and there is evidence of correlation and $r < 0$ So expect amount of sunshine to be lower than the average for Heathrow(oe)	B1 **(1)**	2.2b
		(7 marks)	

	Notes
(a)	B1 for stating negative. "Negative skew" is B0 though
(b)(i)	B1 for mentioning "rainfall" (allow "rain" or "precipitation") or "pressure" (if more than 1 answer both must be correct) NB the other quantitative variable for Perth is: Daily Mean Wind Speed and scores B0 [Not allowed "wind speed" since $r = +0.15$ and in winter might expect wind to raise temp]
(ii)	B1ft for giving the correct units. If Daily Mean Wind Speed (kn) or knots "Wind speed" and "knots" would score B0B1 but any other variable scores B0B0
(c)	B1 for both hypotheses correct in terms of ρ M1 for the correct critical value compatible with their H_1: allow $\pm 0.361(0)$ If the hypotheses are 1-tail then allow cv of ± 0.3061 e.g. Alternative hypothesis with $r < \pm 0.377$ implies a one-tail test or H_0 and H_1 in words saying "H_0: there is no correlation, H_1: there is correlation" is two-tail If there are no hypotheses (or they are nonsensical) assume 2-tail so M1 for $\pm 0.361(0)$ A1 for a correct conclusion in context based on comparing -0.377 with their cv. Condone incorrect inequality e.g. $-0.3610 < -0.377$ as long as they reject H_0 Do not accept contradictory statements such as "accept H_0 so there is evidence of …" Can say "support for Stav's belief"(o.e.e.g. "claim") or "evidence of a correlation between sunshine and humidity" condone "negative correlation" or comments such as "if humidity is high amount of sunshine will be low"
(d)	B1 for stating low amount of sunshine (o. e.) **and** some reference to $r < 0$ or fog Check for the following 2 features: (i) **low** sunshine: allow $\leqslant 5$ hrs (LDS mean for 2015 is 5.3, humidity 97% is 4.1, $\geqslant 97\%$ is 3.1) (ii) **negative** correlation may be described in words e.g. "high humidity gives low sunshine" or **fog** (LDS says >95% humidity is foggy) so less sunshine

2. Marc took a random sample of 16 students from a school and for each student recorded

- the number of letters, x, in their last name
- the number of letters, y, in their first name

His results are shown in the scatter diagram on the next page.

(a) Describe the correlation between x and y.

(1)

Marc suggests that parents with long last names tend to give their children shorter first names.

(b) Using the scatter diagram comment on Marc's suggestion, giving a reason for your answer.

(1)

The results from Marc's random sample of 16 observations are given in the table below.

x	3	6	8	7	5	3	11	3	4	5	4	9	7	10	6	6
y	7	7	4	4	6	8	5	5	8	4	7	4	5	5	6	3

(c) Use your calculator to find the product moment correlation coefficient between x and y for these data.

(1)

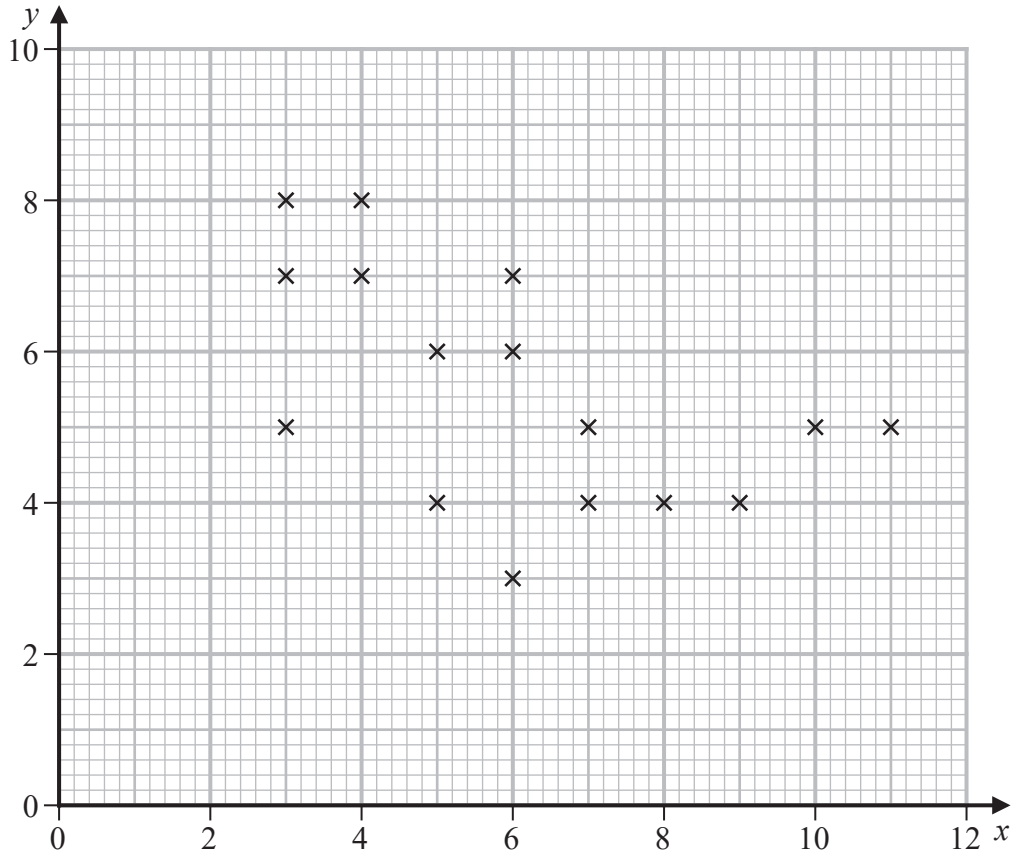
(d) Test whether or not there is evidence of a negative correlation between the number of letters in the last name and the number of letters in the first name.

You should

- state your hypotheses clearly
- use a 5% level of significance

(3)

Question 2 continued.



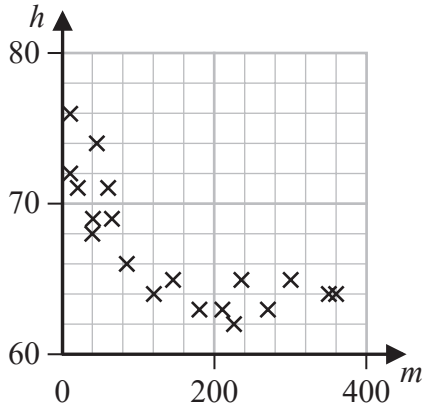
Qu 2	Scheme	Marks	AO
(a)	Negative	B1 **(1)**	1.2
(b)	Marc's suggestion <u>is compatible</u> because it's <u>negative correlation</u>	B1 **(1)**	2.4
(c)	$(r =)$ – 0.54458266… awrt **<u>– 0.545</u>**	B1 **(1)**	1.1b
(d)	$H_0 : \rho = 0$ $H_1 : \rho < 0$	B1	2.5
	[5% 1-tail cv =] (<u>+</u>) 0.4259	M1	1.1a
	(significant result / reject H_0) There <u>is</u> evidence of negative <u>correlation</u> between the <u>number of letters</u> in (or <u>length</u> of) a student's last <u>name</u> and their first <u>name</u>	A1 **(3)**	2.2b
		(6 marks)	

	Notes
(a)	B1 for "negative" Allow "slight" or "weak" etc Allow a description e.g. "as x increases y decreases" or in context e.g. "people with longer last names tend to have shorter first names" A comment of "negative skew" is B0 **Need to see distinct or separate responses for (a) and (b)**
(b)	B1 for a comment that suggests data is compatible with the suggestion **and** a suitable reason such as "there is negative correlation" <u>or</u> a description in x and y or in context <u>or</u> the points lie close to a line with <u>negative gradient</u> <u>or</u> draw line $y = x$ and state that <u>more points below the line</u> so <u>supports (or is compatible with)</u> his suggestion A reason based on just a **single point** is B0 e.g. " 11 letters in last name has only 5 in first name"
(c)	B1 for awrt – 0.545
(d)	B1 for both hypotheses correct in terms of ρ M1 for a critical value compatible with their H_1: 1-tail: awrt <u>+</u> 0.426 (condone <u>+</u> 0.425) **or** 2-tail (B0 scored for H_1) : awrt <u>+</u> 0.497 If hypotheses are in words and can deduce whether one or two-tail then use their words. If no hypotheses or their H_1 is not clearly one or two tail assume one-tail A1 for compatible signs between cv and r **and** a correct conclusion in context mentioning <u>correlation</u> and <u>number of letters</u> or <u>length</u> and <u>name</u> (ft their value from (c)) Do NOT award this A mark if contradictory comments or working seen e.g. "accept H_0" or comparison of 0.426 with significance level of 0.05 etc
NB	The M1A1 can be scored independently of the hypotheses

6. Anna is investigating the relationship between exercise and resting heart rate. She takes a random sample of 19 people in her year at school and records for each person

- their resting heart rate, h beats per minute
- the number of minutes, m, spent exercising each week

Her results are shown on the scatter diagram.



(a) Interpret the nature of the relationship between h and m

(1)

Anna codes the data using the formulae

$$x = \log_{10} m$$
$$y = \log_{10} h$$

The product moment correlation coefficient between x and y is -0.897

(b) Test whether or not there is significant evidence of a negative correlation between x and y
You should

- state your hypotheses clearly
- use a 5% level of significance
- state the critical value used

(3)

The equation of the line of best fit of y on x is

$$y = -0.05x + 1.92$$

(c) Use the equation of the line of best fit of y on x to find a model for h on m in the form

$$h = am^k$$

where a and k are constants to be found.

(5)

Question	Scheme		Marks	AOs
6(a)	eg As the number of minutes <u>exercise</u> (m) increases the resting <u>heart rate</u> (h) decreases **or** the gradient of the curve is becoming flatter with increasing m: diminishing effect of each <u>additional minute</u> of <u>exercise</u>		B1	2.4
			(1)	
(b)	$H_0 : \rho = 0$ $H_1 : \rho < 0$		B1	2.5
	Critical value – 0.3887 (Allow $\pm$)		M1	1.1b
	There is evidence that the product moment **<u>correlation</u>** is **<u>less than 0</u>/ <u>there is a negative correlation</u>**		A1	2.2b
			(3)	
(c)	$\log_{10} h = -0.05 \log_{10} m + 1.92$	$h = am^k \rightarrow \log_{10} h = \log_{10} am^k$	M1	1.1b
	$\log_{10} h = -\log_{10} m^{0.05} + 1.92$ or $\log_{10} h = \log_{10} m^{-0.05} + 1.92$ or $h = 10^{1.92 - 0.05 \log_{10} m}$ oe	$\log_{10} h = \log_{10} a + \log_{10} m^k$ or $\log_{10} a = 1.92$	M1	2.1
	$\log_{10} hm^{0.05} = 1.92$ or $\log_{10}\left(\frac{h}{m^{-0.05}}\right) = 1.92$ or $h = 10^{1.92} \times 10^{-0.05 \log_{10} m}$ oe	$\log_{10} h = \log_{10} a + k \log_{10} m$	M1	1.1b
	$hm^{0.05} = 10^{1.92}$ or $\frac{h}{m^{-0.05}} = 10^{1.92}$ or $h = 10^{1.92} \times 10^{\log_{10} m^{-0.05}}$	$\log_{10} a = 1.92$ and $k = -0.05$	M1	1.1b
	$h = 10^{1.92} m^{-0.05}$ or $h = 83.17...m^{-0.05}$ or a = awrt 83.17 **and** $k = -0.05$		A1	1.1b
			(5)	

Notes: **(9 marks)**

(a)	**B1**	eg Idea as one increases the other decreases (in context). Allow use of m and h eg As m increases h decreases. Do not allow negative correlation with no context or $\rho < 0$ Allow there is a negative correlation/association/relationship/exponential between minutes <u>exercise</u>(m) and resting <u>heart rate</u> (h) oe
(b)	**B1**	Both hypotheses correct in terms of ρ (allow p)
	M1	For the cv of -0.3887 or any cv such that $0.3 < \lvert cv \rvert < 0.5$
	A1	Independent of hypotheses. Correct conclusion that implies reject H_0 on basis of seeing – 0.3887 or if they give 0.3887 we must see the comparison $0.3887 < 0.897$ **and** which mentions "pmcc/correlation/relationship" and less than 0/ negative or ρ **< 0** A contradictory statement scores A0 eg Accept H_0 therefore negative correlation
(c)		**In this part once M0 is scored no more marks can be scored. Condone no base**
	M1	**May be implied by 2nd M1 mark** Method 1: Correct substitution for both x and y Method 2 : Taking the log of both sides
	M1	**May be implied by 3rd M1 mark** Method 1: Correct use of the power log rule **or** making h the subject Method 2 : Correct use of the addition/subtraction log rule
	M1	**This line implies M1M1M1** Method 1: Correct use of the addition/subtraction log rule or eq^n in the form $h = 10^{1.92} \times 10^{-0.05 \log m}$ Method 2: A second correct step for correct use of the power log rule
	M1	**This line implies M1M1M1M1** Method 1: Correct removal of logs or $h = 10^{1.92} \times 10^{\log m^{-0.05}}$ Method 2: Log a (or a) and k correct
	A1	Allow h = awrt $83.2m^{-0.05}$ NB award 5/5 for a = awrt 83.2 and $k = -0.05$ or h = awrt $83.2...m^{-0.05}$ or $h = 10^{1.92} m^{-0.05}$