ADVANCED MATHEMATICS

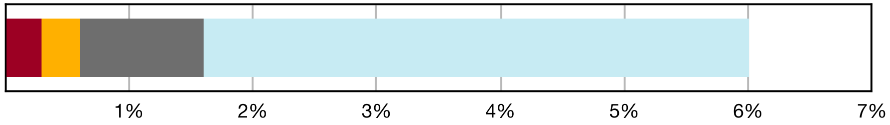
Statistics (Adv), S2 Interpretation and Bivariate Data (Adv)

Bivariate Data Analysis (Y12)

Teacher: Cathyanne Horvat

Exam Equivalent Time: 60 minutes *(based on allocation of 1.5 minutes per mark)*

S2 Interpretation and Bivariate Data



🟥 Bar Charts and Histograms
🟨 Other Chart Types
⬛ Summary Statistics - No Graph
🟦 Bivariate Data Analysis

***SmarterMaths analytics based on the average contribution to new syllabus Advanced Maths exams since 2020.**

HISTORICAL CONTRIBUTION

- *S2 Interpretation and Bivariate Data* has contributed an average of 6.0% per new syllabus Advanced exam since it began in 2020.
- We have split the area into six sub-topics for analysis purposes which are: *1-Classifying Data (0%), 2-Bar Charts and Histograms (0.3%), 3-Other Chart Types (0.3%), 4-Summary Statistics - Box Plots (0%), 5-Summary Statistics - No Graph (1.0%) and 6-Bivariate Data Analysis (4.4%).*
- This analysis looks at *Bivariate Data Analysis*.

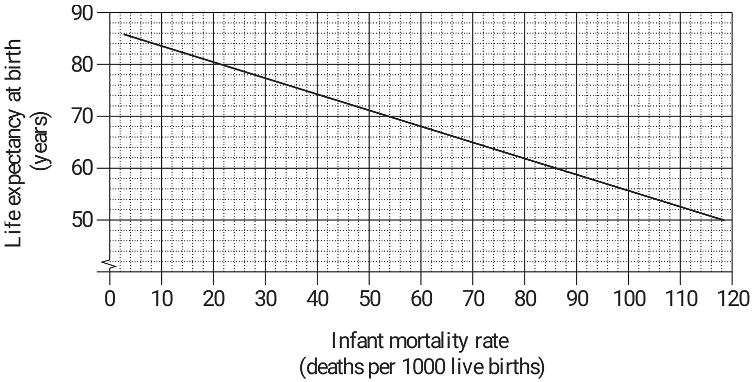
HSC ANALYSIS - What to expect and common pitfalls

- *Bivariate Data Analysis (4.4%)* investigates scatterplots, correlation, lines of best fit, least squares regression analysis etc ... and represents the key area where common questions appear in both the Adv and Std2 exams.
- Common Adv-Std2 exam questions have appeared in all new syllabus Adv exams, with significant allocations of between 4-5 marks on each occasion. This is a trend we expect to continue and any revision should reflect this.
- The word-heavy narrative style of the *2021 Q17* and *2020 Q27* represent critical revision questions that students must review carefully.
- Just as importantly, the *2022 Q24* example presented students with a novelly broad question to "describe and interpret the data and other information provided, with reference to the context given". *Revision here should include a critical focus on an efficient structure of any solution.*
- Calculating a least squares regression line from raw data is a key skill that students must be able to execute efficiently. *EQ-Bank Q1-3* are important revision examples covering this area.
- ***Pitfalls:*** Marker's comments have highlighted issues in finding equations of best fit, interpreting gradients and identifying limitations of an equation - all areas that are thoroughly covered in this database.

Questions

1. Algebra, STD2 A2 2017 HSC 3 MC

The graph shows the relationship between infant mortality rate (deaths per 1000 live births) and life expectancy at birth (in years) for different countries.



What is the life expectancy at birth in a country which has an infant mortality rate of 60?

A. 68 years

B. 69 years

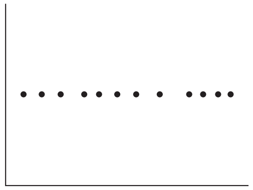
C. 86 years

D. 88 years

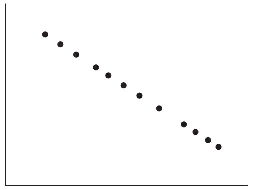
2. Statistics, STD2 S4 2017 HSC 12 MC

Which of the data sets graphed below has the largest positive correlation coefficient value?

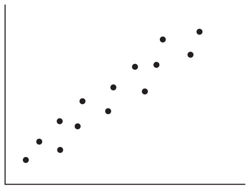
A.



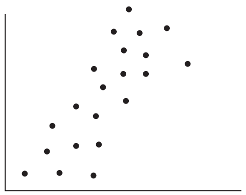
B.



C.

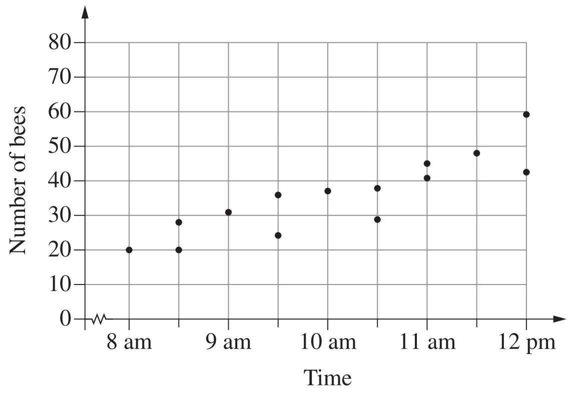


D.



3. Statistics, 2ADV S2 2023 HSC 1 MC

The number of bees leaving a hive was observed and recorded over 14 days at different times of the day.



Which Pearson's correlation coefficient best describes the observations?

A. – 0.8

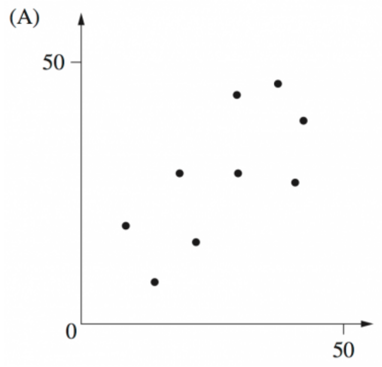
B. – 0.2

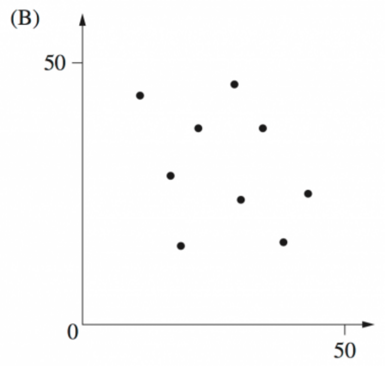
C. 0.2

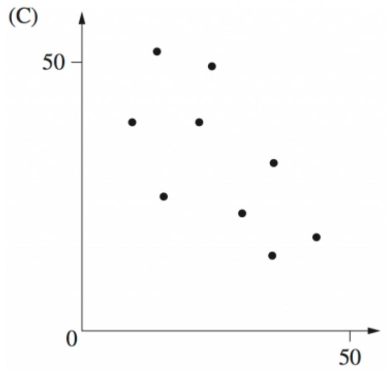
D. 0.8

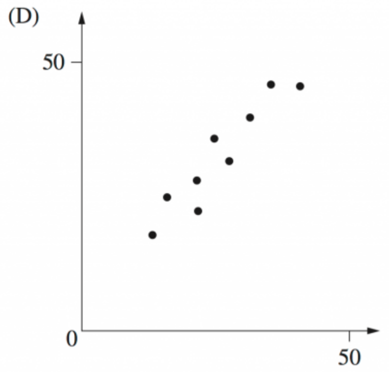
4. Statistics, STD2 S4 2013 HSC 2 MC

Which graph best shows data with a correlation closest to 0.3?









5. Statistics, STD2 S4 2012 HSC 11 MC

Which of the following relationships would most likely show a negative correlation?

A. The population of a town and the number of hospitals in that town.

B. The hours spent training for a race and the time taken to complete the race.

C. The price per litre of petrol and the number of people riding bicycles to work.

D. The number of pets per household and the number of computers per household.

6. Statistics, STD2 S4 SM-Bank 1

A student claimed that as time spent swimming training increases, the time to run a 1 kilometre time trial decreases.

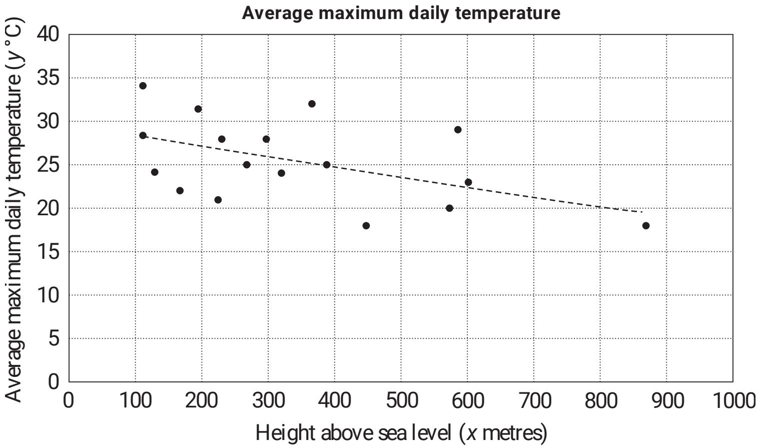
After collecting and analysing some data, the student found the correlation coefficient, r, to be – 0.73.

What does this correlation indicate about the relationship between the time a student spends swimming training and their 1 kilometre run time trial times. ***(1 mark)***

7. Statistics, 2ADV S2 2021 HSC 17

For a sample of 17 inland towns in Australia, the height above sea level, x (metres), and the average maximum daily temperature, y (°C), were recorded.

The graph shows the data as well as a regression line.



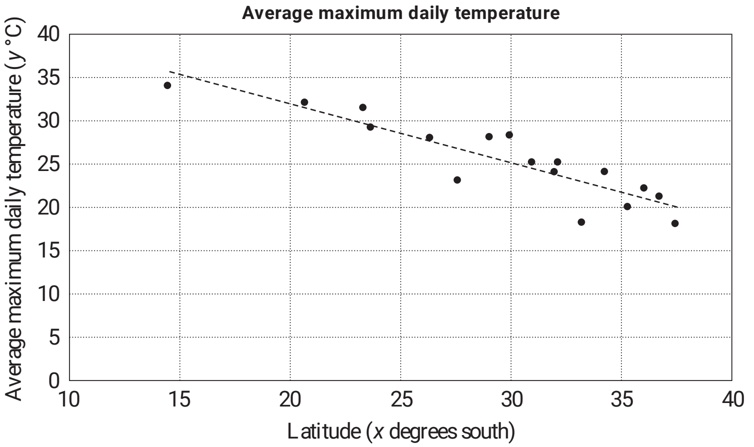
The equation of the regression line is $y = 29.2 - 0.011x$.

The correlation coefficient is $r = -0.494$.

a. **i.** By using the equation of the regression line, predict the average maximum daily temperature, in degrees Celsius, for a town that is 540 m above sea level. Give your answer correct to one decimal place. ***(1 mark)***

ii. The gradient of the regression line is −0.011. Interpret the value of this gradient in the given context. ***(2 marks)***

b. The graph below shows the relationship between the latitude, x (degrees south), and the average maximum daily temperature, y (°C), for the same 17 towns, as well as a regression line.



The equation of the regression line is $y = 45.6 - 0.683x$.

The correlation coefficient is $r = -0.897$.

Another inland town in Australia is 540 m above sea level. Its latitude is 28 degrees south.

Which measurement, height above sea level or latitude, would be better to use to predict this town's average maximum daily temperature? Give a reason for your answer. ***(1 mark)***

8. Statistics, 2ADV S2 EQ-Bank 2

The table below lists the average *body weight* (in kilograms) and average *brain weight* (in grams) of nine animal species.

species	body weight (kg)	brain weight (g)
baboon	10.55	179.5
cat	3.30	25.6
goat	27.70	115.0
guinea pig	1.04	5.5
rabbit	2.50	12.1
rat	0.28	1.9
red fox	4.24	50.4
rhesus monkey	6.80	179.0
sheep	55.50	175.0

A least squares regression line is fitted to the data using *body weight* as the independent variable.

i. Calculate the equation of the least squares regression line. ***(1 mark)***

ii. If dingos have an average body weight of 22.3 kilograms, calculate the predicted average brain weight of a dingo using your answer to part i. ***(1 mark)***

9. Statistics, 2ADV S2 EQ-Bank 3

The table below lists the average *life span* (in years) and average *sleeping time* (in hours/day) of 9 animal species.

species	**life span (years)**	**sleeping time (hours/day)**
baboon	27	10
cow	30	4
goat	20	4
guinea pig	8	8
horse	46	3
mouse	3	13
pig	27	8
rabbit	18	8
rat	5	13

i. Using *sleeping time* as the independent variable, calculate the least squares regression line. ***(1 mark)***

ii. A wallaby species sleeps for 4.5 hours, on average, each day.

Use your equation from **part i** to predict its expected *life span*, to the nearest year. ***(1 mark)***

10. Statistics, 2ADV S2 2023 HSC 18

A university uses gas to heat its buildings. Over a period of 10 weekdays during winter, the gas used each day was measured in megawatts (MW) and the average outside temperature each day was recorded in degrees Celsius (°C).

Using x as the average daily outside temperature and y as the total daily gas usage, the equation of the least-squares regression line was found.

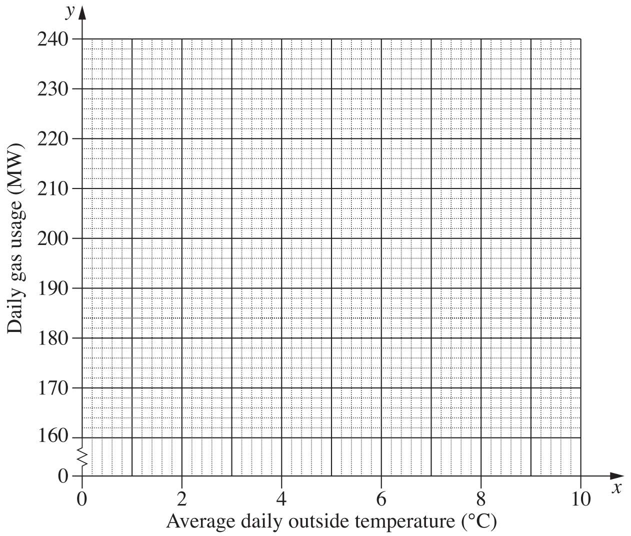
The equation of the regression line predicts that when the temperature is 0°C, the daily gas usage is 236 MW.

The ten temperatures measured were: 0°, 0°, 0°, 2°, 5°, 7°, 8°, 9°, 9°, 10°,

The total gas usage for the ten weekdays was 1840 MW.

In any bivariate dataset, the least-squares regression line passes through the point $(\bar{x}, \bar{y})$, where $\bar{x}$ is the sample mean of the x-values and $\bar{y}$ is the sample mean of the y-values.

a. Using the information provided, plot the point $(\bar{x}, \bar{y})$ and the y-intercept of the least-squares regression line on the grid. ***(3 marks)***



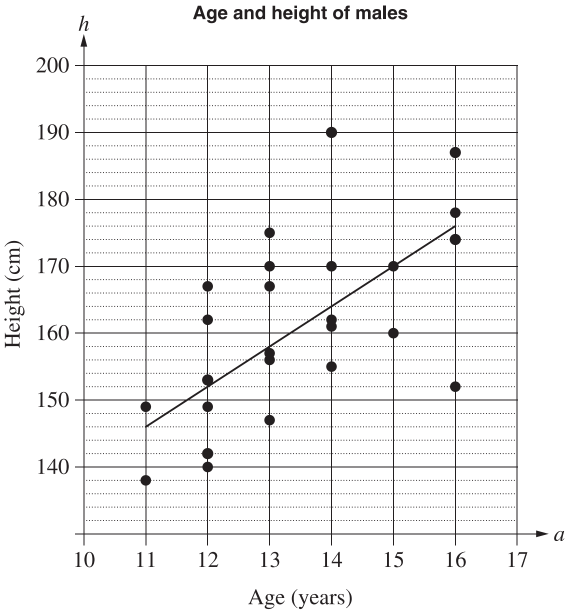
b. What is the equation of the regression line? ***(2 marks)***

c. In the context of the dataset, identify ONE problem with using the regression line to predict gas usage when the average outside temperature is 23°C. ***(1 mark)***

11. Statistics, STD2 S4 2013 HSC 28b

Ahmed collected data on the age (a) and height (h) of males aged 11 to 16 years.

He created a scatterplot of the data and constructed a line of best fit to model the relationship between the age and height of males.



i. Determine the gradient of the line of best fit shown on the graph. ***(1 mark)***

ii. Explain the meaning of the gradient in the context of the data. ***(1 mark)***

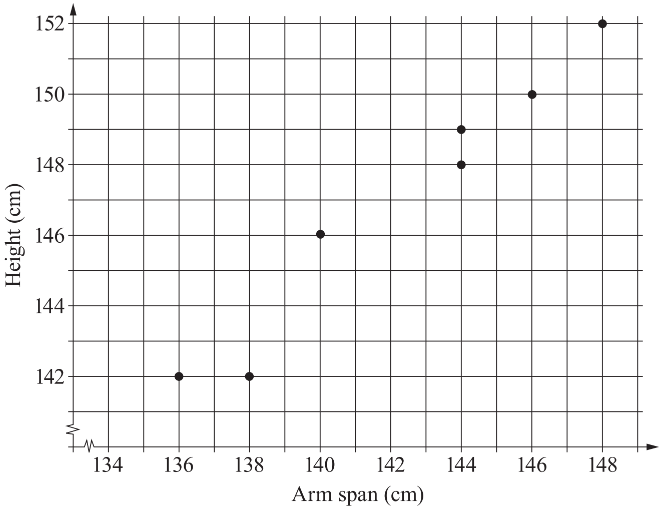
iii. Determine the equation of the line of best fit shown on the graph. ***(2 marks)***

iv. Use the line of best fit to predict the height of a typical 17-year-old male. ***(1 mark)***

v. Why would this model not be useful for predicting the height of a typical 45-year-old male? ***(1 mark)***

12. Statistics, STD2 S4 2019 HSC 23

A set of bivariate data is collected by measuring the height and arm span of seven children. The graph shows a scatterplot of these measurements.



a. Calculate Pearson's correlation coefficient for the data, correct to two decimal places. ***(1 mark)***

b. Identify the direction and the strength of the linear association between height and arm span. ***(1 mark)***

c. The equation of the least-squares regression line is shown.

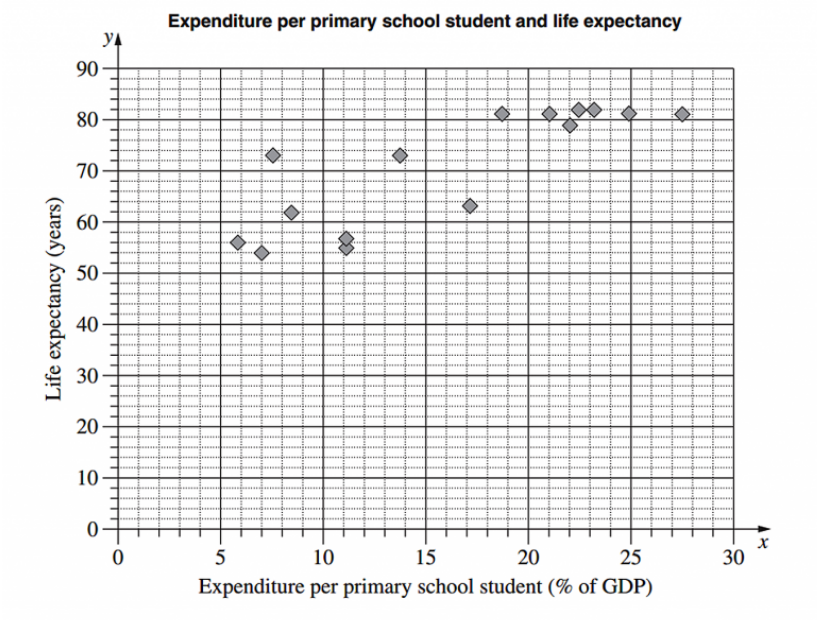
Height = 0.866 × (arm span) + 23.7

A child has an arm span of 143 cm.

Calculate the predicted height for this child using the equation of the least-squares regression line. ***(1 mark)***

13. Statistics, STD2 S4 2014* HSC 30b

The scatterplot shows the relationship between expenditure per primary school student, as a percentage of a country's Gross Domestic Product (GDP), and the life expectancy in years for 15 countries.



i. For the given data, the correlation coefficient, r, is 0.83. What does this indicate about the relationship between expenditure per primary school student and life expectancy for the 15 countries? *(1 mark)*

ii. For the data representing expenditure per primary school student, Q_L is 8.4 and Q_U is 22.5.
What is the interquartile range? *(1 mark)*

iii. Another country has an expenditure per primary school student of 47.6% of its GDP.
Would this country be an outlier for this set of data? Justify your answer with calculations. *(2 marks)*

iv. On the scatterplot, draw the least-squares line of best fit $y = 1.29x + 49.9$. *(2 marks)*

v. Using this line, or otherwise, estimate the life expectancy in a country which has an expenditure per primary school student of 18% of its GDP. *(1 mark)*

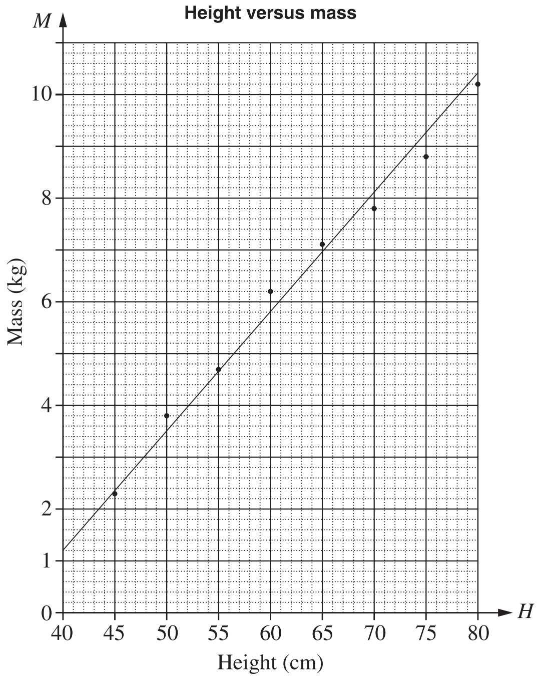
vi. Why is this line NOT useful for predicting life expectancy in a country which has expenditure per primary school student of 60% of its GDP? *(1 mark)*

14. Statistics, STD2 S4 2009 HSC 28b

The height and mass of a child are measured and recorded over its first two years.

Height (cm), H	45	50	55	60	65	70	75	80
Mass (kg), M	2.3	3.8	4.7	6.2	7.1	7.8	8.8	10.2

This information is displayed in a scatter graph.



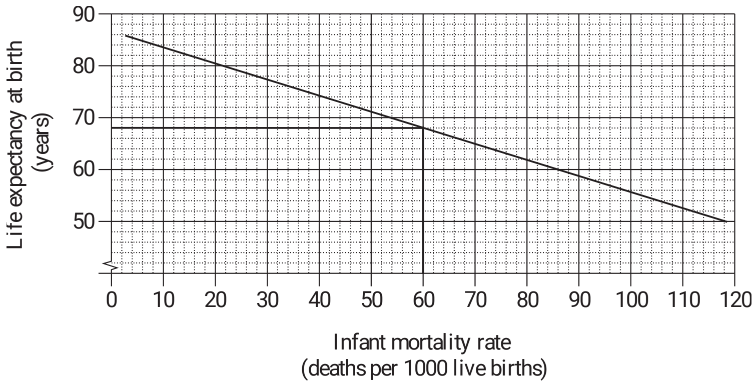
i. Describe the correlation between the height and mass of this child, as shown in the graph. *(1 mark)*

ii. A line of best fit has been drawn on the graph.
Find the equation of this line. *(2 marks)*

Worked Solutions

1. Algebra, STD2 A2 2017 HSC 3 MC

text(When infant mortality rate is 60, life expectancy at birth is 68 years (see below).



$\Rightarrow A$

2. Statistics, STD2 S4 2017 HSC 12 MC

Largest positive correlation occurs when both variables move in tandem. The tighter the linear relationship, the higher the correlation.

$\Rightarrow C$

(Note that B is negatively correlated)

3. Statistics, 2ADV S2 2023 HSC 1 MC

Correlation is positive and strong.

Best option: $r = 0.8$

$\Rightarrow D$

NOTE: Inputting all data points into a calculator is unnecessary and time consuming here.

4. Statistics, STD2 S4 2013 HSC 2 MC

A is correct since the data slopes bottom left to top right (i.e. it's positive).

D also slopes correctly but exhibits a higher correlation co-efficient.

$\Rightarrow A$

5. Statistics, STD2 S4 2012 HSC 11 MC

Increased hours training should reduce the time to complete a race.

$\Rightarrow B$

♦ Mean mark 43%.

6. Statistics, STD2 S4 SM-Bank 1

– 0.73 indicates a strong negative relationship exists.

In this case, it means the more time spent swimming training is associated with a quicker time of running a 1 kilometre time trial.

7. Statistics, 2ADV S2 2021 HSC 17

a.i.

$$\begin{aligned} y &= 29.2–0.011(540) \\ &= 23.26 \\ &= 23.3^{\circ}\text{C} \ \ (1 \text{ d.p.}) \end{aligned}$$

a.ii. On average, the average maximum daily temperature of inland towns drops by 0.011 of a degree for every metre above sea level the town is situated.

b. The correlation co-efficient of the regression line using latitude is significantly stronger than the equivalent co-efficient for the regression line using height above sea level.

$\therefore$ The equation using latitude is preferred.

8. Statistics, 2ADV S2 EQ-Bank 2

i. By calculator:

brain weight $= 49.4 + 2.68 \times$ body weight

COMMENT: Know this critical calculator skill!.

ii. Predicted brain weight of a dingo

$$\begin{aligned} &= 49.4 + 2.68 \times 22.3 \\ &= 109.164 \\ &= 109 \text{ grams} \end{aligned}$$

9. Statistics, 2ADV S2 EQ-Bank 3

i. By calculator:

life span $= 42.89 – 2.85 \times$ sleeping time

ii. Predicted life span of wallaby

$$= 42.89 – 2.85 \times 4.5$$
$$= 30.06\ldots$$
$$= 30 \text{ years}$$

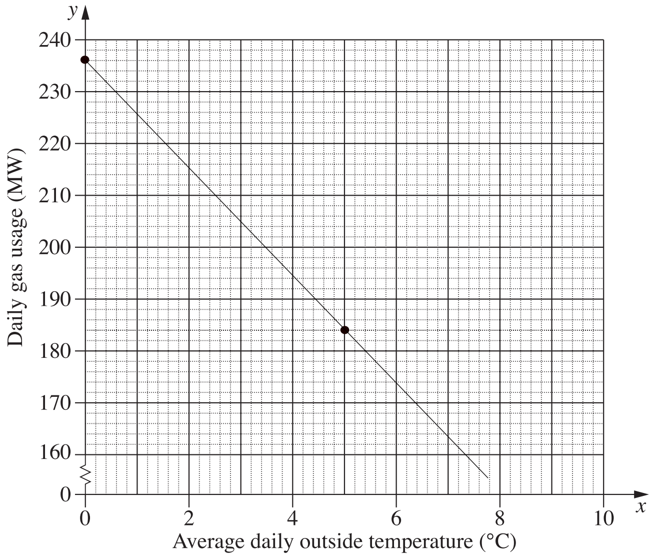
COMMENT: Issues here? YouTube has short and excellent help videos – search your calculator model and topic – eg. "fx-82 regression line".

10. Statistics, 2ADV S2 2023 HSC 18

a. $\overline{x} = \dfrac{0+0+0+2+5+7+8+9+9+10}{10} = 5°\text{C}$

$\overline{y} = \dfrac{1840}{10} = 184$

Regression line passes through: $(0, 236)$ and $(5, 184)$



b. $m = \dfrac{y_2 - y_1}{x_2 - x_1} = \dfrac{184 - 236}{5 - 0} = -\,10.4$

Equation of line $m = -\,10.4$ passing through $(0, 236)$:

$$(y - y_1) = m(x - x_1)$$
$$y - 236 = -\,10.4(x - 0)$$
$$y = -\,10.4x + 236$$

c. Answers could include one of the following:

→ 23˚C is outside the range of the dataset and requires the trend to be extrapolated.

→ At 23˚C, the equation predicts negative daily gas usage.

11. Statistics, STD2 S4 2013 HSC 28b

i. $\text{Gradient} = \dfrac{176-146}{16-11} = \dfrac{30}{5} = 6$

♦♦ *Mean marks of 38%, 26% and 25% respectively for parts (i)-(iii).*
MARKER'S COMMENT: Interpreting gradients has been consistently examined in recent history and almost always poorly answered.

ii. Males should grow 6cm per year between the ages 11–16.

iii. Gradient $= 6$, Passes through $(11, 146)$

$$\begin{aligned} y - y_1 &= m(x - x_1) \\ h - 146 &= 6(a - 11) \\ \therefore h &= 6a - 66 + 146 \\ &= 6a + 80 \end{aligned}$$

iv. Substitue $a = 17$ into equation from part (iii):

$h = (6 \times 17) + 80 = 182$

$\therefore$ A typical 17 year old is expected to be 182cm.

v. People slow and eventually stop growing after they become adults.

12. Statistics, STD2 S4 2019 HSC 23

a. Use "A + Bx" function (fx-82 calc):

$$\begin{aligned} r &= 0.9811\ldots \\ &= 0.98 \quad (2\text{ d.p.}) \end{aligned}$$

♦ *Mean mark 40%.*
COMMENT: Issues here? YouTube has short and excellent help videos – search your calculator model and topic – eg. "fx-82 correlation" .

b. Direction: positive

Strength: strong

c. $$\begin{aligned} \text{Height} &= 0.866 \times 143 + 23.7 \\ &= 147.538 \text{ cm} \end{aligned}$$

13. Statistics, STD2 S4 2014* HSC 30b

i. It indicates there is a strong positive correlation between the two variables

ii. $$\begin{aligned} \text{IQR} &= Q_U - Q_L \\ &= 22.5 - 8.4 \\ &= 14.1 \end{aligned}$$

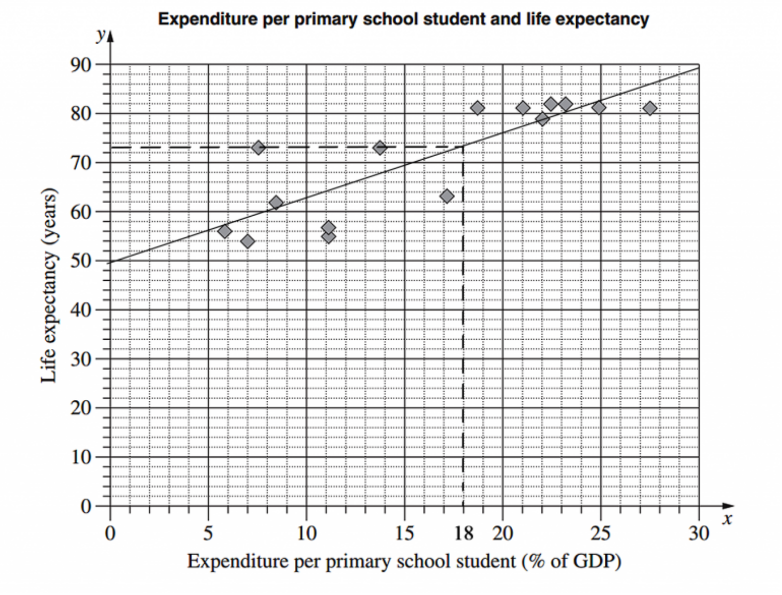
iii. An outlier on the upper side must be more than

$$\begin{aligned} Q_u &+ 1.5 \times IQR \\ &= 22.5 + (1.5 \times 14.1) \\ &= 43.65\% \end{aligned}$$

$\therefore$ A country with an expenditure of 47.6% is an outlier.

♦ *Mean mark 35%*

iv.



v. Life expectancy ≈ 73.1 years (see dotted line)

♦♦ *Mean mark 39%*

Alternative Solution
When $x = 18$
$y = 1.29(18) + 49.9 = 73.12$ years

vi. At 60% GDP, the line predicts a life expectancy of 127.3. This line of best

♦♦♦ *Mean mark 0%. The toughest question on the 2014 paper.*
COMMENT: Examiners regularly ask students to identify and comment

fit is only predictive in a lower range of GDP expenditure.

on outliers where linear relationships break down.

14. Statistics, STD2 S4 2009 HSC 28b

i. **The correlation between height and mass is positive and strong.**

♦ *Mean mark 48%.*

ii. **Using** $P_1(40, 1.2)$ **and** $P_2(80, 10.4)$

$$\begin{aligned}\text{Gradient} &= \frac{y_2 - y_1}{x_2 - x_1}\\ &= \frac{10.4 - 1.2}{80 - 40}\\ &= \frac{9.2}{40}\\ &= 0.23\end{aligned}$$

♦♦♦ *Mean mark 18%.*
MARKER'S COMMENT: Many students had difficulty due to the fact the horizontal axis started at $H = 40\text{cm}$ and not the origin.

Line passes through $P_1(40, 1.2)$

Using $y - y_1 = m(x - x_1)$

$$\begin{aligned}y - 1.2 &= 0.23(x - 40)\\ y - 1.2 &= 0.23x - 9.2\\ y &= 0.23x - 8\end{aligned}$$

$\therefore$ **Equation of the line is** $M = 0.23H - 8$