Section 3.1.15 Least Squares Regression Line

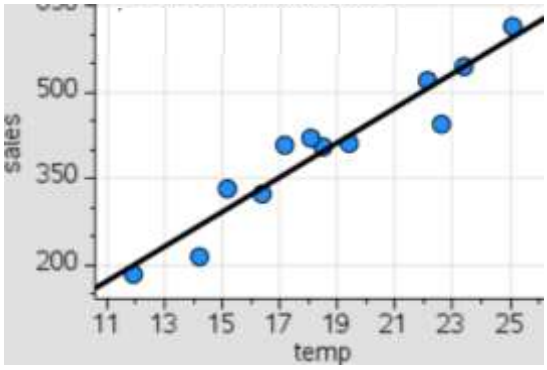
VCAA "Dot Points"
Investigating data distributions, including:

- *least squares line of best fit $y = a + bx$, where x represents the explanatory variable and y represents the response variable; the determination of the coefficients a and b using technology, and the formulas $b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$*
- *modelling linear association between two numerical variables, including the:*
  - *identification of the explanatory and response variables*
  - *use of the least squares method to fit a linear model to the data*

Linear regression

The objective of linear regression is to find the **best fitting straight line** through a series of points upon a scatterplot. This technique is used to model the relationship between two numerical variables.

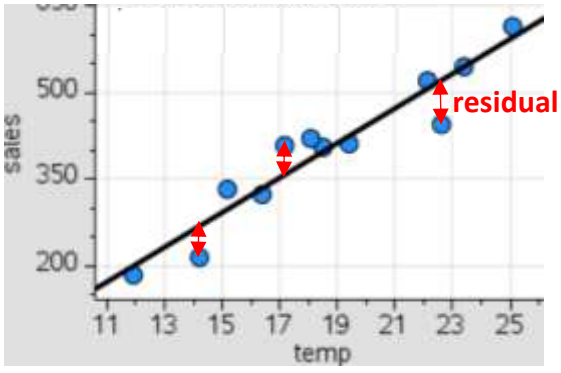
Example 1
Consider the ice cream sales v daily temperature example from Notes 3.1.12. The linear regression for this data is as follows:



Residuals

To understand how a linear regression is determined, one first needs to understand the term **residual**.

A residual is the **difference** in the vertical direction (along y-axes) between the **observed data** (scatterplot dot) and the **predicted value** from the regression line.

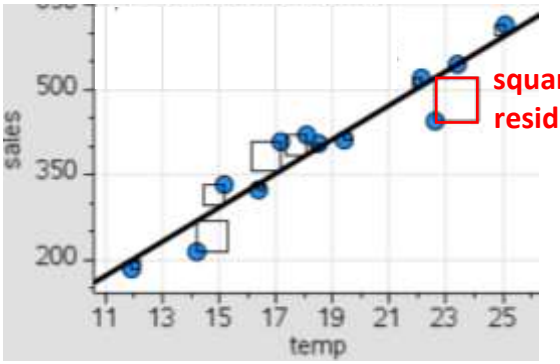


NB: Those points (dots) located **below** the line have a **negative residual** value.

Those points (dots) located **above** the line have a **negative residual** value.

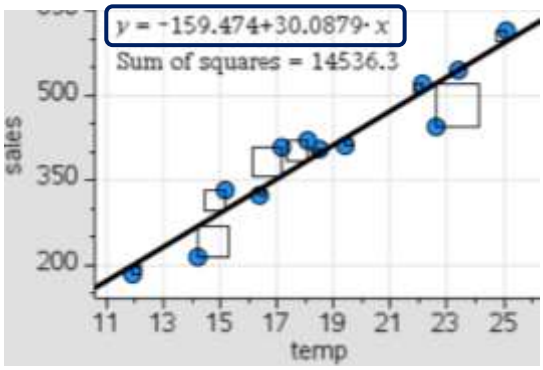
Squared Residuals

If each individual residual line were used to construct a square, there would be as many squares as there are points (dots) on the scatterplot of varying size.



Least Squares Regression

A least squares regression line **minimises the sum of the squared values** of the residual.



So, the least squares regression line for the ice cream sales versus temperature is as follows:

$$Ice\ cream\ sales\ (\$) = -159.474 + 30.0879 \times Temperature(℃)$$

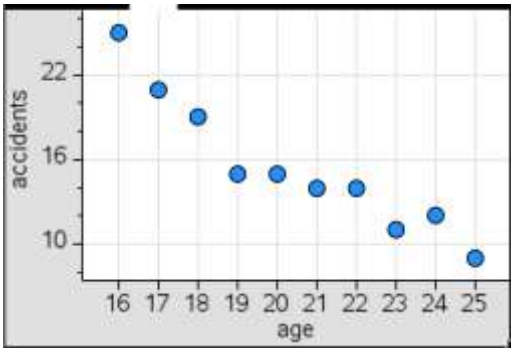
Example 2

Consider the car accidents v driver age example from Notes 3.1.12. The linear regression for this data is as follows:

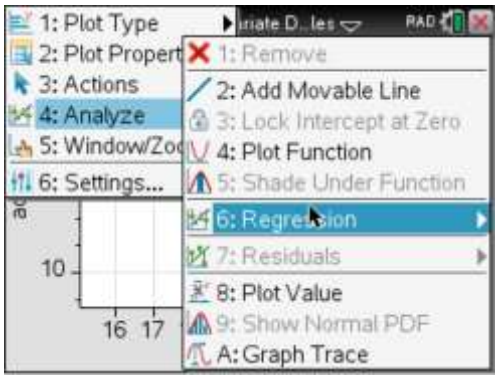
List & Spreadsheet

	A age	B accide...	C	D
1	16	25		
2	17	21		
3	18	19		
4	19	15		
5	20	15		

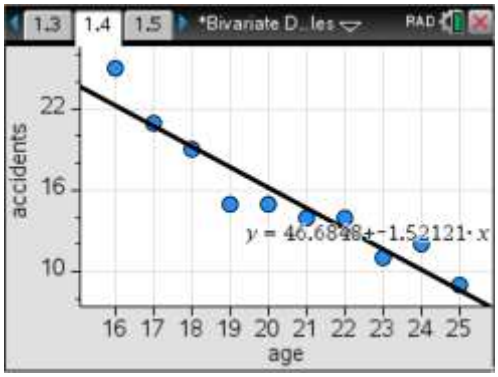
Data & Statistics



menu 4.Analyze /6. Regression



menu ... 2. Show linear $(a + bx)$



Therefore, the least squares regression line for the car accidents versus driver age is as follows:

$$No.of\ car\ accidents = 46.6848 - 1.52121 \times Driver\ Age\ (years)$$

Calculating the least squares regression equation by hand

Recall the general form of the least squares regression line is:

$$y = a + bx$$

Where a is the y-intercept
b is the slope (gradient)

The following pair of equations can be used to calculate the least squares regression equation:

$$b = r\frac{S_y}{S_x}$$

Where b is the slope
r is the Pearson's product-moment correlation coefficient
S_x is the standard deviation of the explanatory variable
S_y is the standard deviation of the response variable

$$a = \bar{y} - b\bar{x}$$

Where a is the y-intercept
$\bar{y}$ is the mean of the response variable
b is the slope
$\bar{x}$ is the mean of the explanatory variable

Example 3

A study was conducted to investigate the effect of drinking coffee on sleep.
In this study, the amount of sleep, in hours, and the amount of coffee drunk, in cups, on a given day were recorded for a group of adults. The following summary statistics were generated.

	Sleep (hours)	*Coffee* (cups)
Mean	7.08	2.42
Standard deviation	1.12	1.56
Correlation coefficient (r)	−0.770	

Calculate the least squares regression equation.

Step.1 Calculate the slope (b)

$$b = r\frac{S_y}{S_x}$$
$$= -0.770 \times \frac{1.12}{1.56}$$
$$= -0.553 \text{ (3 decimal places)}$$

Step.2 Calculate the y-intercept (a)

$$a = \bar{y} - b\bar{x}$$
$$= 7.08 - (-0.553 \times 2.42)$$
$$= 8.418 \text{ (3 decimal places)}$$

Answer: $Sleep\ (hours) = 8.418 - 0.553 \times Coffee\ (cups)$

Once you have calculated a regression line in the format of $y = a + bx$, there are many additional calculations and conclusion that can be formed from the equation.
For example, you could be asked to:

1. Predict a response value from a given explanatory value, or vice versa
2. Interpret the significance of the y-Intercept
3. Interpret the significance of the slope or gradient

Predicting from a regression equation

Predictions can be made using a linear regression equation by substituting either a **response** or **explanatory variable**.

Example 4

Use the following linear regression equation to predict the ice cream sales for a day of temperature 36°C.

$$Ice\ cream\ sales\ (\$) = -159.474 + 30.0879 \times Temperature(°C)$$

Substitute the temperature in to the equation:

$$\begin{aligned} Ice\ cream\ sales\ (\$) &= -159.474 + 30.0879 \times \mathbf{36} \\ &= -159.474 + 1083.1644 \\ &= \mathbf{923.69} \end{aligned}$$

Answer: the predicted ice cream sales for a day of temperature 36°C, is $923.69

Example 5

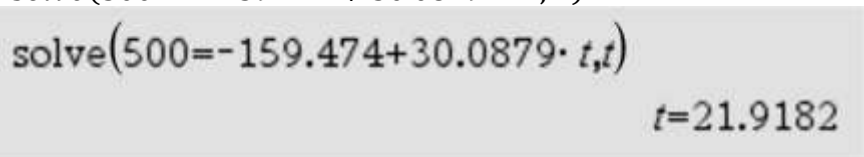
Use the following linear regression equation to predict the daily temperature that would result in $500 ice cream sales.

$$Ice\ cream\ sales\ (\$) = -159.474 + 30.0879 \times Temperature(°C)$$

Substitute the ice cream sales in to the equation:

$$\mathbf{500} = -159.474 + 30.0879 \times Temperature(°C)$$

$$solve(500 = -159.474 + 30.0879 \times T, T)$$



Answer: the predicted daily temperature for ice cream sales of $500, is 21.9°C

Interpret the significance of the y-Intercept

Recall: the y-intercept is found when $x = 0$.

Please use the following template response when interpreting the y intercept:

"A **[explanatory variable]** of 0 **[explanatory units]** has an expected **[response variable]** of **[y-intercept] [response units]**."

Example 6

Interpret the y-intercept for the linear regression equation used to predict the ice cream sales for a daily temperature.

$$Ice\ cream\ sales\ (\$) = \boxed{-159.474} + 30.0879 \times Temperature(℃)$$

Y-intercept

Interpretation:
A **temperature** of 0 ℃ has an expected **Ice Cream Sales** of **-$159.47**. Or rather;
When the daily temperature is 0 ℃ *the ice cream sales are predicted to be -$159.47.*

NB: Clearly, this prediction is nonsense, as ice cream sales cannot be less than $0.

Example 7

Interpret the y-intercept for the linear regression equation used to predict the number of car accidents per 100 drivers given the driver age.

$$No.of\ car\ accidents = \boxed{46.6848} - 1.52121 \times Driver\ Age\ (years)$$

Y-intercept

Interpretation:
A **driver age** of 0 **years** has an expected **No. of car accidents** of 47 **accidents per 100 people.**
Or rather;
When the driver is 0 years old there is expected to be 46.68 accidents per 100 drivers.

NB: Clearly, this prediction is nonsense, as a newborn does not drive a car!

Example 8

Interpret the y-intercept for the linear regression equation used to predict the score on a test given the hours of study.

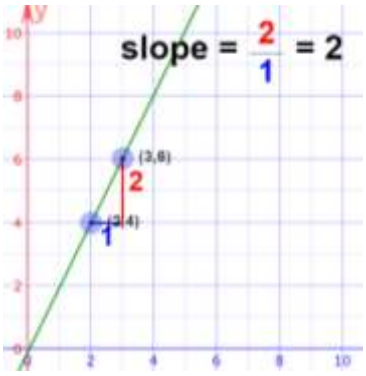
$$Test\ Score\ (\%) = \boxed{54.28} + 7.71 \times Study\ (hours)$$

Y-intercept

Interpretation:
A **study** of 0 **hours** has an expected **test score** of **54 %.** Or rather;
"When there student undertake 0 hours of study they expect to achieve a score of approximately 54%."

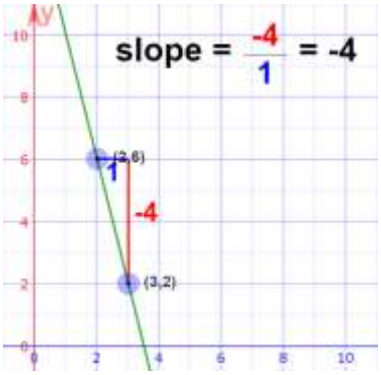
Interpreting the slope (gradient)

The slope, or gradient, indicates the **change in the response variable** for every **one-unit increase** in the **explanatory variable**.



A unit increase along the x-axes produces an increase in the y-axes of 2.

∴ a slope of +2



A unit increase along the x-axes produces a decrease in the y-axes of 4.

∴ a slope of -4

Please use the following template response when interpreting the slope or gradient.

"On average, for every extra **[explanatory unit]** of **[explanatory variable]** the **[response variable] [increases/decreases] by [gradient] [response units]**."

Example 9

Interpret the slope for the linear regression equation used to predict the ice cream sales for a daily temperature.

increases

$$Ice\ cream\ sales\ (\$) = -159.474 + 30.0879 \times Temperature(℃)$$

slope

Interpretation:

"On average, for every extra ℃ of **Temperature** the **Ice cream sales increases by $30.09**."

Or rather;

"On average, for every 1℃ increase in temperature, the ice cream sales will increase by $30.09."

Example 10

Interpret the slope for the linear regression equation used to predict the number of car accidents per 100 drivers given the driver age.

decreases

$$No.of\ car\ accidents = 46.6848 - 1.52121 \times Driver\ Age\ (years)$$

slope

Interpretation:

"On average, for every extra **year** of **driver age** the **number of accidents decreases** by **-1.52 people per 100 drivers**."

Or rather;

On average, for every 1 year increase in driver age, the number of accidents per 100 drivers decreases by 1.55.

Interpolation & Extrapolation

To tell the difference between extrapolation and interpolation, we need to look at the prefixes "extra" and "inter."

- The prefix "**extra**" means "**outside**" or "**in addition to**."
- The prefix "**inter**" means "**in between**" or "**among**."

Just knowing these meanings goes a long way to distinguishing between the two methods.

Interpolation

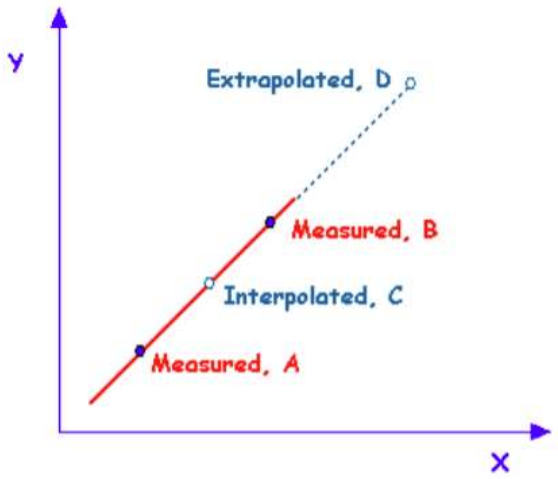
We could use our regression line to **predict** the value of the response variable for a given explanatory variable variable that is **inside the range of our data.**
(ie. between the smallest and largest original data).

Extrapolation

We could use our regression line to **predict** the value of the response variable for a given explanatory variable that is **outside the range of our data**.
(ie. data that is smaller than smallest original data or largest than the largest original data).

Caution

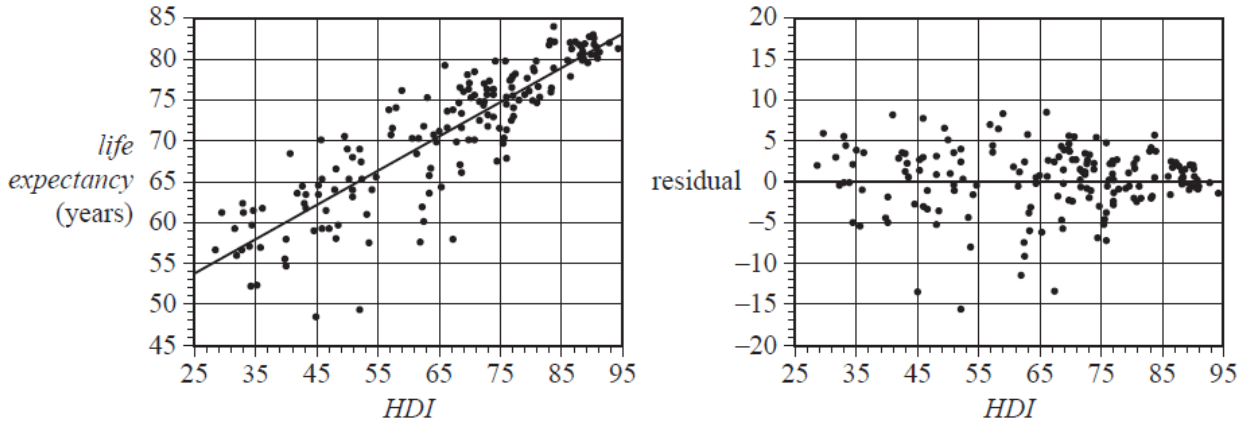
Of the two methods, **interpolation is preferred**. This is because we have a greater likelihood of obtaining a **valid estim**ate. When we use extrapolation, we are assuming that our observed trend continues for values of x outside the range we used to form our model. This may not be the case, and so we must be very careful when using extrapolation techniques.



Exam Styled Questions– Multiple Choice

Use the following information to answer Questions 1 & 2.

The scatterplot below shows life expectancy in years (life expectancy) plotted against the Human Development Index (HDI) for a large number of countries in 2011.
A least squares line has been fitted to the data and the resulting residual plot is also shown.



The equation of this least squares line is

$$life\ expectancy = 43.0 + 0.422 \times HDI$$

The coefficient of determination is $r^2 = 0.875$

Question 1
(2016 Exam 1 Section A – Qn 9)

Given the information above, which one of the following statements is **not** true?

A. The value of the correlation coefficient is close to 0.94
B. 12.5% of the variation in life expectancy is not explained by the variation in the Human Development Index.
C. On average, life expectancy increases by 43.0 years for each 10-point increase in the Human Development Index.
D. Ignoring any outliers, the association between life expectancy and the Human Development Index can be described as strong, positive and linear.
E. Using the least squares line to predict the life expectancy in a country with a Human Development Index of 75 is an example of interpolation.

C

A. $r^2 = 0.875$ therefore $r = \sqrt{0.875} = 0.935 \approx 0.94$. Therefore **TRUE**
B. $r^2 = 0.875$ therefore 87.5% of variation in LE is explained by the variation in HDI. Therefore 12.5% is NOT explained. Therefore **TRUE**.
C. A gradient of 0.442 indicates that a rise of 10 in HDI would produce an increase in LE of 42.2, not 43. Therefore **FALSE**
D. $r = \sqrt{0.875} = 0.935 \approx 0.94$ Therefore a strong, positive, linear association **TRUE**
E. A prediction using a HDI of 75 is interpolation (ie. within the data set) **TRUE**

Therefore Option C

Question 2
(2016 Exam 1 Section A – Qn 10)

In 2011, life expectancy in Australia was 81.8 years and the Human Development Index was 92.9
When the least squares line is used to predict life expectancy in Australia, the residual is closest to

A. –0.6
B. –0.4
C. 0.4
D. 11.1
E. 42.6

B

Actual value (point)
(92.9, 81.8)

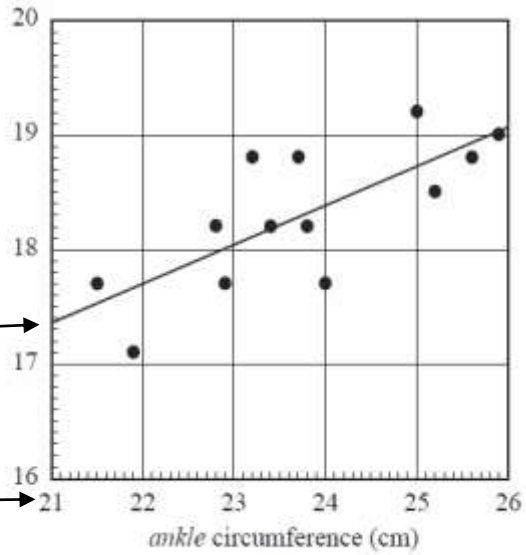
Predicted value (from linear equation)

$$life\ expectancy = 43.0 + 0.422 \times HDI$$
$$= 43.0 + 0.422 \times 92.9$$
$$= 82.2038$$

$$Residual = Actual\ y\ value - Predicted\ y\ value$$
$$= 81.8 - 82.2038$$
$$= -0.4038$$ **Therefore Option B**

Use the following information to answer Questions 3 & 4.

The scatterplot below shows the wrist circumference and ankle circumference, both in centimetres, of 13 people. A least squares line has been fitted to the scatterplot with ankle circumference as the explanatory variable.



17.4 is **NOT** the y intercept

As it is measured when ankle circumference is 21cm **NOT** 0cm

Question 3
(2017 Exam 1 Section A – Qn 8)

The equation of the least squares line is closest to

A. $ankle = 10.2 + 0.342 \times wrist$
B. $wrist = 10.2 + 0.342 \times ankle$
C. $ankle = 17.4 + 0.342 \times wrist$
D. $wrist = 17.4 + 0.342 \times ankle$
E. $wrist = 17.4 + 0.731 \times ankle$

B **Therefore Option B**

A. Wrong response & explanatory variables **FALSE**
B. Correct variables & only option with correct y-intercept **TRUE**
C. Wrong response & explanatory variables **FALSE**
D. Correct variables, but incorrect y-intercept **FALSE**
E. Correct variables, but incorrect y-intercept **FALSE**

NB: You don't even have to calculate the gradient to answer the question.

Question 4
(2017 Exam 1 Section A – Qn 9)

When the least squares line on the scatterplot is used to predict the wrist circumference of the person with an ankle circumference of 24 cm, the residual will be closest to

A. –0.7
B. –0.4
C. –0.1
D. 0.4
E. 0.7

Actual value (point) from off the graph
(24, 17.7)

Predicted value (from linear equation)

$$\begin{aligned} wrist &= 10.2 + 0.342 \times ankle \\ &= 10.2 + 0.342 \times 24 \\ &= 18.408 \end{aligned}$$

$$\begin{aligned} Residual &= Actual\ y\ value - Predicted\ y\ value \\ &= 17.7 - 18.408 \\ &= -0.708 \end{aligned}$$ **Therefore Option A**

Use the following information to answer Questions 5–8.

The table below shows the lean body mass (LBM), percentage body fat (PBF) and body mass index (BMI) of a sample of 12 professional athletes

LBM (kg)	*PBF* (%)	*BMI* (kg/m^2)
63.3	19.8	20.6
58.6	21.3	20.7
55.4	19.9	21.9
57.2	23.7	21.9
53.2	17.6	19.0
53.8	15.6	21.0
60.2	20.0	21.7
48.3	22.4	20.6
54.6	18.0	22.6
53.4	15.1	19.4
61.9	18.1	21.2
48.3	23.3	22.0

Question 5
(2018 NHT Exam 1 Section A – Qn 7)

The mean, $\bar{x}$, and the standard deviation, s_x, for the lean body mass (LBM) of these athletes, in kilograms, are closest to

A. $\bar{x}$ = 48.3 s_x = 4.6
B. $\bar{x}$ = 55.0 s_x = 4.6
C. $\bar{x}$ = 55.0 s_x = 4.8
D. $\bar{x}$ = 55.7 s_x = 4.6
E. $\bar{x}$ = 55.7 s_x = 4.8

Use the TI-Nspire CAS CX calculator

	D	E	F	G
=			=OneVar(	
1		Title	One-Va...	
2		x̄	55.6833	
3		Σx	668.2	
4		Σx²	37460.7	
5		sx := Sn-...	4.79656	

E **Therefore Option E**

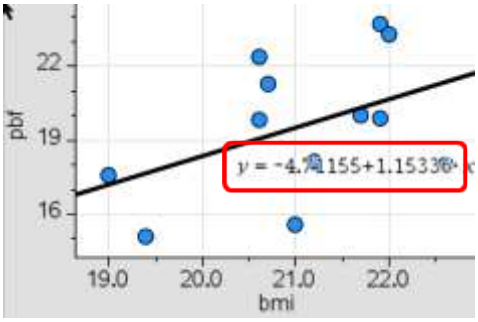
Question 6
(2018 NHT Exam 1 Section A – Qn 8)

A least squares line is fitted to the data using percentage body fat (PBF) as the response variable and body mass index (BMI) as the explanatory variable.

The equation of the least squares line is closest to

A. $PBF = -4.7 + 1.2 \times BMI$
B. $BMI = -4.7 + 1.2 \times PBF$
C. $PBF = 17.8 + 1.7 \times BMI$
D. $BMI = 17.8 + 1.7 \times PBF$
E. $PBF = 23.6 - 0.1 \times BMI$

Use the TI-Nspire CAS CX calculator



A **Therefore Option A**

Question 7
(2018 NHT Exam 1 Section A – Qn 9)

The Pearson correlation coefficient, r, between lean body mass (LBM) and percentage body fat (PBF) is closest to

A. –0.235
B. –0.124
C. 0.124
D. 0.235
E. 0.352

Use the TI-Nspire CAS CX calculator

	G	H	I	J
=			=LinRegB	
2		RegEqn	a+b*x	
3		a	59.8393	
4		b	-0.2123...	
5		r²	0.015431	
6		r	-0.1242...	

B **Therefore Option B**

Question 8
(2018 NHT Exam 1 Section A – Qn 10)

A least squares line is fitted to the data using lean body mass (LBM) as the response variable and body mass index (BMI) as the explanatory variable.
The equation of this line is

$$LBM = 48.9 + 0.320 \times BMI$$

When this line is used to predict the lean body mass (LBM) of an athlete with a body mass index (BMI) of 22.0, the residual will be closest to

A. –7.6 kg
B. –1.5 kg
C. 1.5 kg
D. 33.9 kg
E. 55.9 kg

A

Actual value (point) from table
(22.0, 48.3)

Predicted value (from linear equation)

$$\begin{aligned} LBM &= 48.9 + 0.320 \times BMI \\ &= 48.9 + 0.320 \times 22 \\ &= 55.94 \end{aligned}$$

$$\begin{aligned} Residual &= Actual\ y\ value - Predicted\ y\ value \\ &= 48.3 - 55.94 \\ &= -7.64 \end{aligned}$$ **Therefore Option A**

Question 9
(2018 Exam 1 Section A – Qn 13)

The statistical analysis of a set of bivariate data involving variables x and y resulted in the information displayed in the table below.

Mean	$\bar{x} = 27.8$	$\bar{y} = 33.4$
Standard deviation	$s_x = 2.33$	$s_y = 3.24$
Equation of the least squares line	$y = -2.84 + 1.31x$	

Using this information, the value of the correlation coefficient r for this set of bivariate data is closest to

A. 0.88
B. 0.89
C. 0.92
D. 0.94
E. 0.97

D

The equation for the slope (b)

$$b = r\frac{S_y}{S_x}$$

$$1.31 = r \times \frac{3.24}{2.33}$$

$$\text{solve}\left(1.31 = r \cdot \frac{3.24}{2.33}, r\right) \qquad r = 0.942068$$

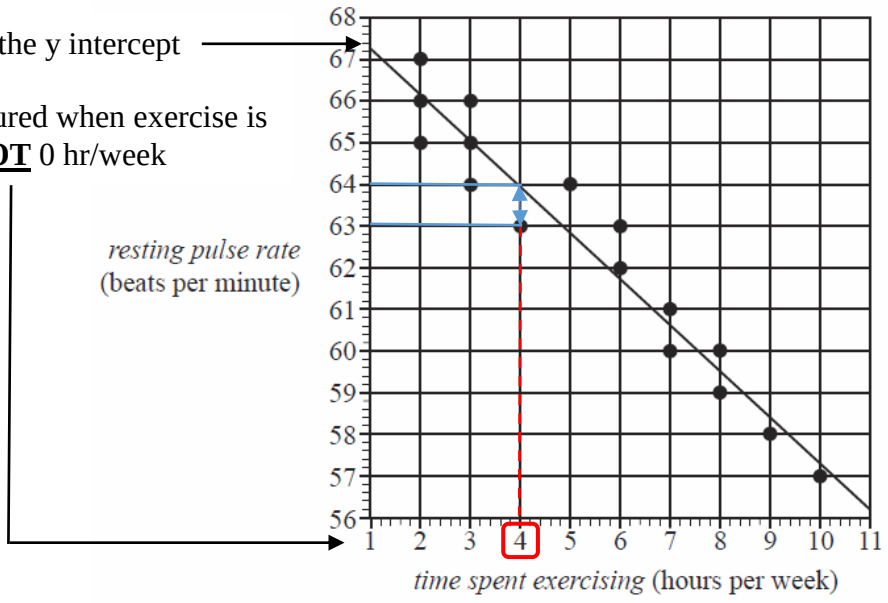
Therefore Option D

Use the following information to answer Questions 10–12.

The scatterplot below displays the resting pulse rate, in beats per minute, and the time spent exercising, in hours per week, of 16 students. A least squares line has been fitted to the data.

67.2 is **NOT** the y intercept

As it is measured when exercise is 1 hr/week **NOT** 0 hr/week



Question 10
(2018 Exam 1 Section A – Qn 7)

Using this least squares line to model the association between resting pulse rate and time spent exercising, the residual for the student who spent four hours per week exercising is closest to

A. –2.0 beats per minute.
B. –1.0 beats per minute.
C. –0.3 beats per minute.
D. 1.0 beats per minute.
E. 2.0 beats per minute.

B

Actual value (point) from off the graph
(4, 63)

Predicted value (from line) @ $x = 4$ is 64

$$Residual = Actual\ y\ value - Predicted\ y\ value$$
$$= 63 - 64$$
$$= -1.0$$ **Therefore Option B**

Question 11
(2018 Exam 1 Section A – Qn 8)

The equation of this least squares line is closest to

A. $resting\ pulse\ rate = 67.2 - 0.91 \times time\ spent\ exercising$ Wrong y-intercept
B. $resting\ pulse\ rate = 67.2 - 1.10 \times time\ spent\ exercising$ Wrong y-intercept
C. $resting\ pulse\ rate = 68.3 - 0.91 \times time\ spent\ exercising$ Wrong gradient
D. $resting\ pulse\ rate = 68.3 - 1.10 \times time\ spent\ exercising$ **Correct y-intercept & gradient**
E. $resting\ pulse\ rate = 67.2 + 1.10 \times time\ spent\ exercising$ Wrong y-intercept

D

Gradient calculation

Step.1 pick 2 point
Point 1 (1, 67.2)
Point 2 (4, 64)

Step 2 Solve

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$
$$= \frac{64 - 67.2}{4 - 1}$$
$$= -1.07 \approx -1.10$$

Therefore Option D

Question 12
(2018 Exam 1 Section A – Qn 9)

The coefficient of determination is 0.8339
The correlation coefficient r is closest to

A. –0.913
B. –0.834
C. –0.695
D. 0.834
E. 0.913

$r^2 = 0.8339$
$r = \pm\sqrt{0.8339}$
$r = \pm 0.913$

NB: the linear regression equation has a **NEGATIVE** gradient

A

$\therefore r = -0.913$
Therefore Option A

Use the following information to answer Questions 9 and 10.

A least squares line is used to model the relationship between the monthly *average temperature* and *latitude* recorded at seven different weather stations. The equation of the least squares line is found to be

$$average\ temperature = 42.9842 - 0.877447 \times latitude$$

Question 9
(2019 Exam 1 Section A – Qn 9)

When the numbers in this equation are correctly rounded to three significant figures, the equation will be

A. $average\ temperature = 42.984 - 0.877 \times latitude$ 3 decimal places NOT significant figures
B. $average\ temperature = 42.984 - 0.878 \times latitude$ 3 decimal places NOT significant figures
C. $average\ temperature = 43.0 - 0.878 \times latitude$ incorrect rounding of gradient
D. $average\ temperature = 42.9 - 0.878 \times latitude$ incorrect rounding of y-intercept
E. $average\ temperature = 43.0 - 0.877 \times latitude$ **CORRECT**

Therefore Option E

E

Question 10
(2019 Exam 1 Section A – Qn 10)

The coefficient of determination was calculated to be 0.893743
The value of the correlation coefficient, rounded to three decimal places, is

A. –0.945
B. –0.898
C. 0.806
D. 0.898
E. 0.945

$r^2 = 0.893743$
$r = \pm\sqrt{0.893743}$
$r = \pm 0.945$

NB: the linear regression equation has a **NEGATIVE** gradient

A

$\therefore r = -0.945$
Therefore Option A