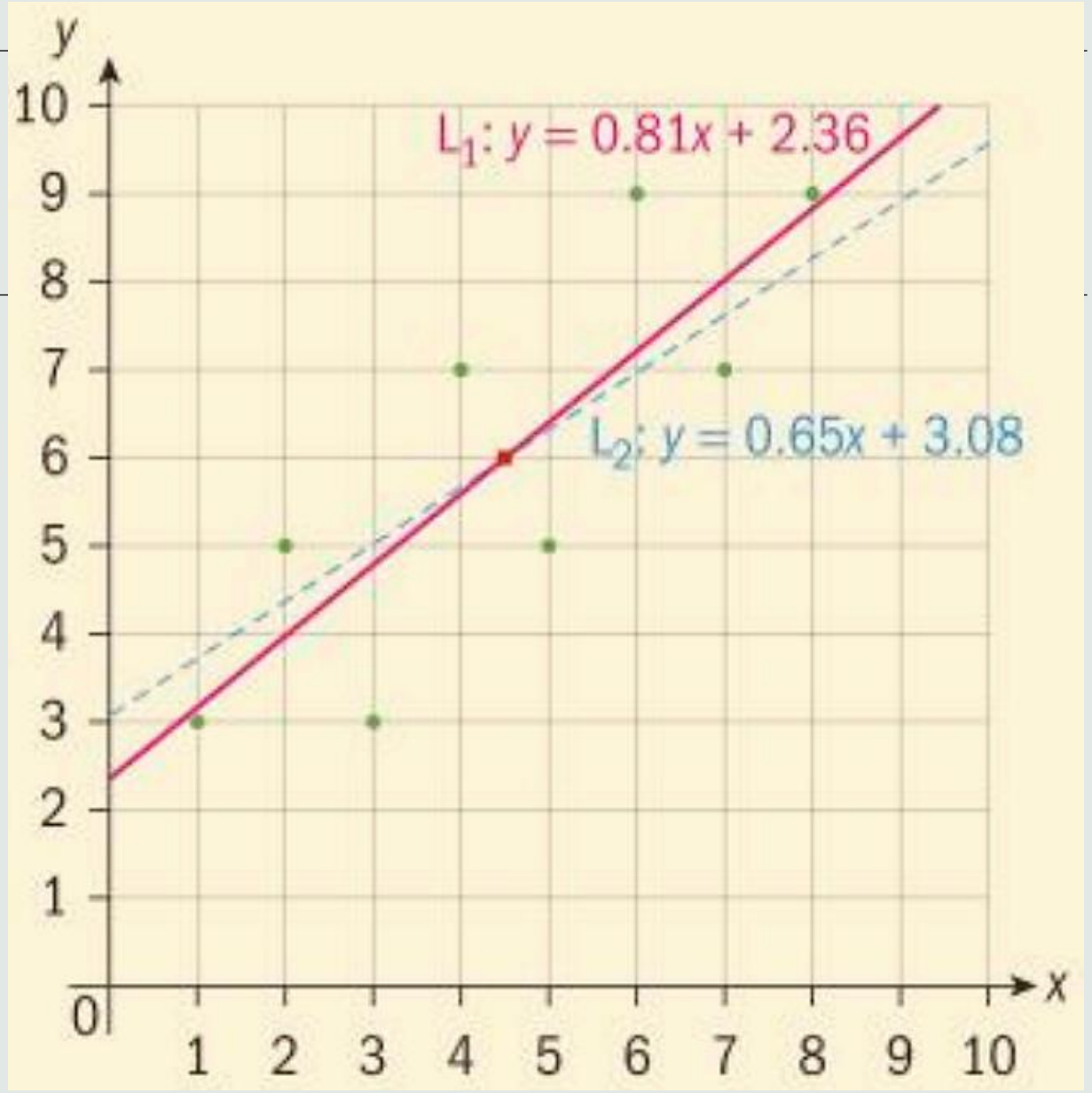
Linear Regression

WEEK 8

Investigation

- Which line do you think better fits the data on the right? Why?

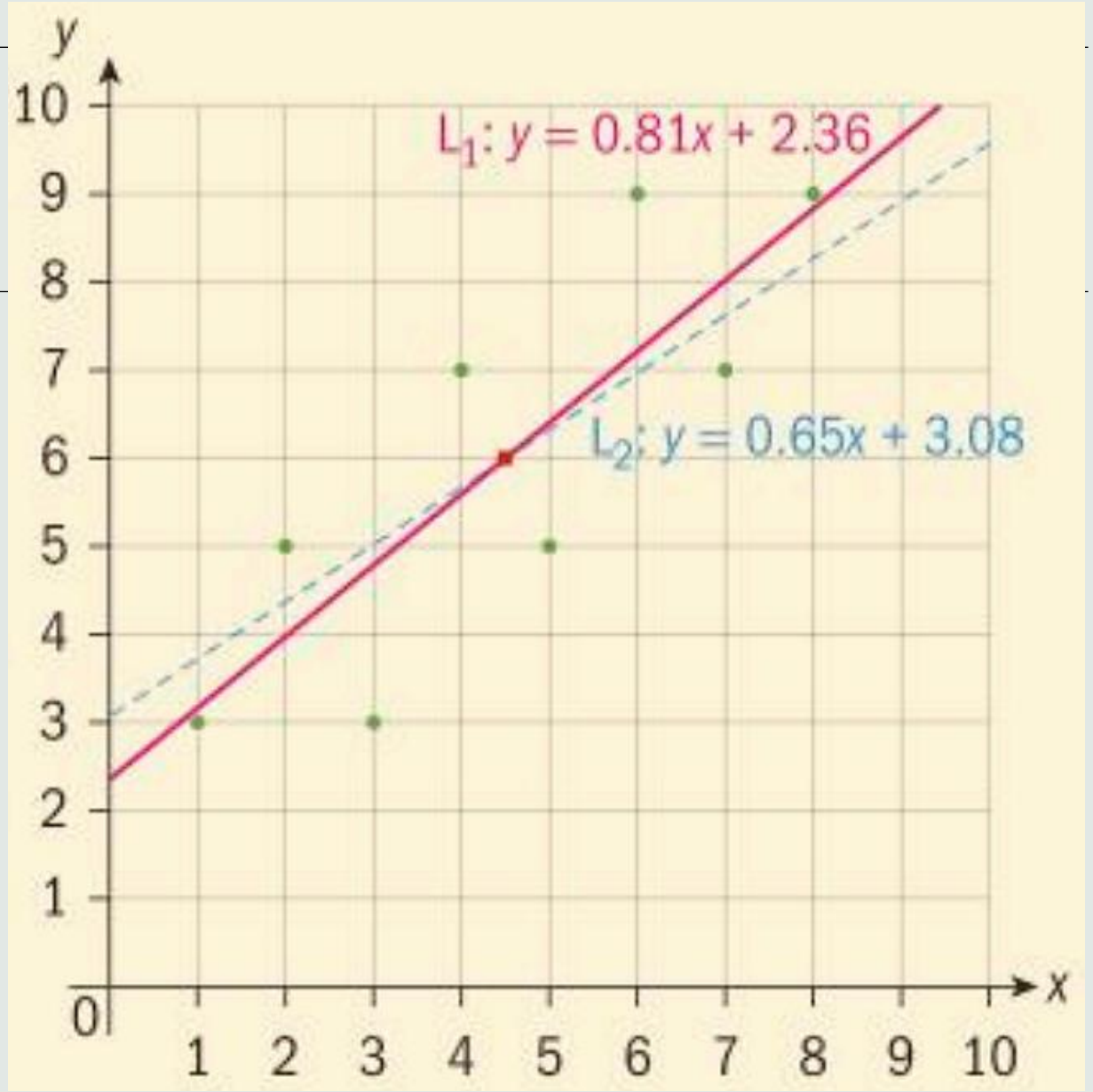


Investigation

- One way to measure the fit of a line to a data set is to calculate residuals.

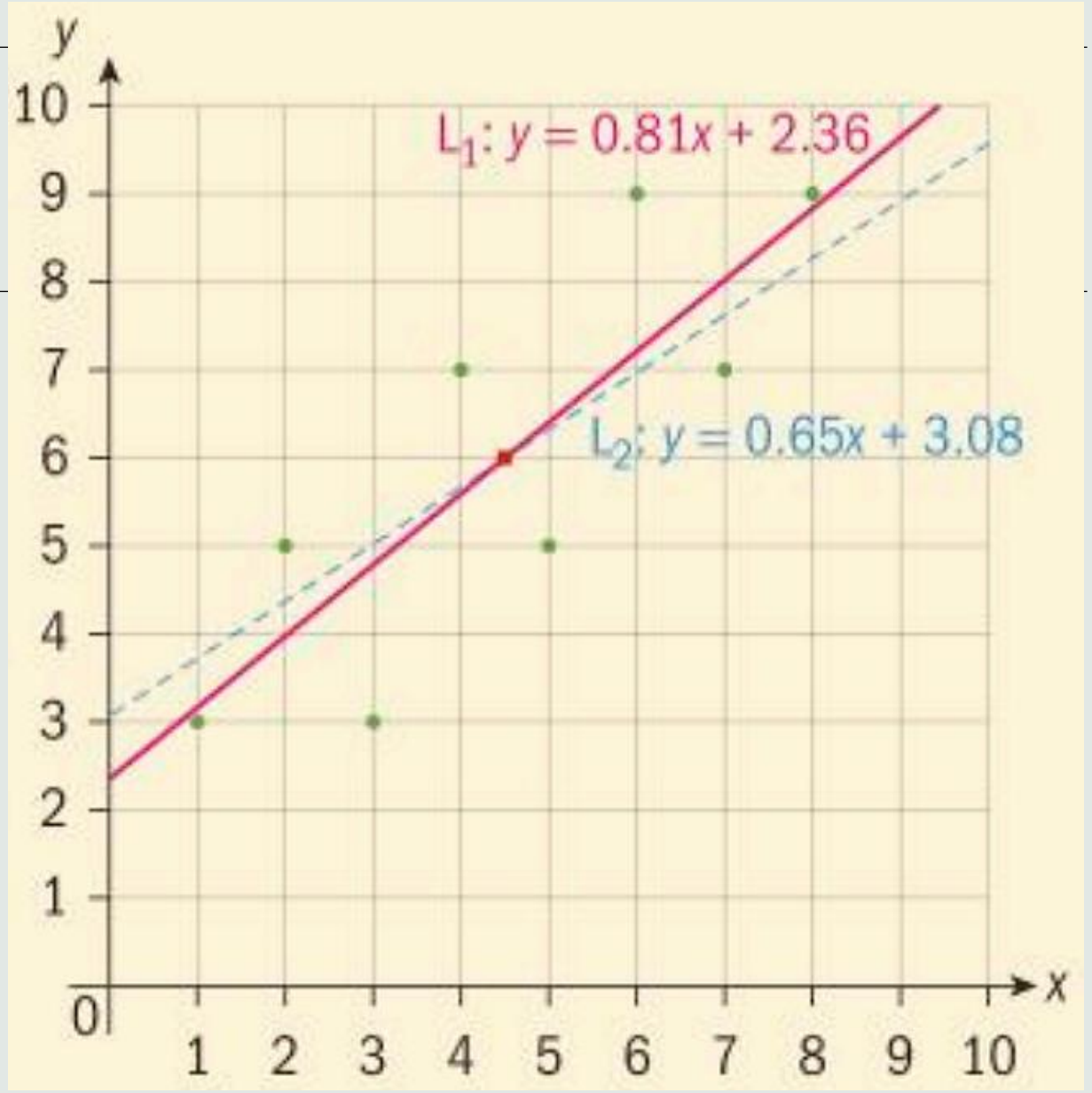
- **Residuals** – the difference between the actual y-value and the predicted y-value.

  - these are errors made when using best-fit lines to make predictions.



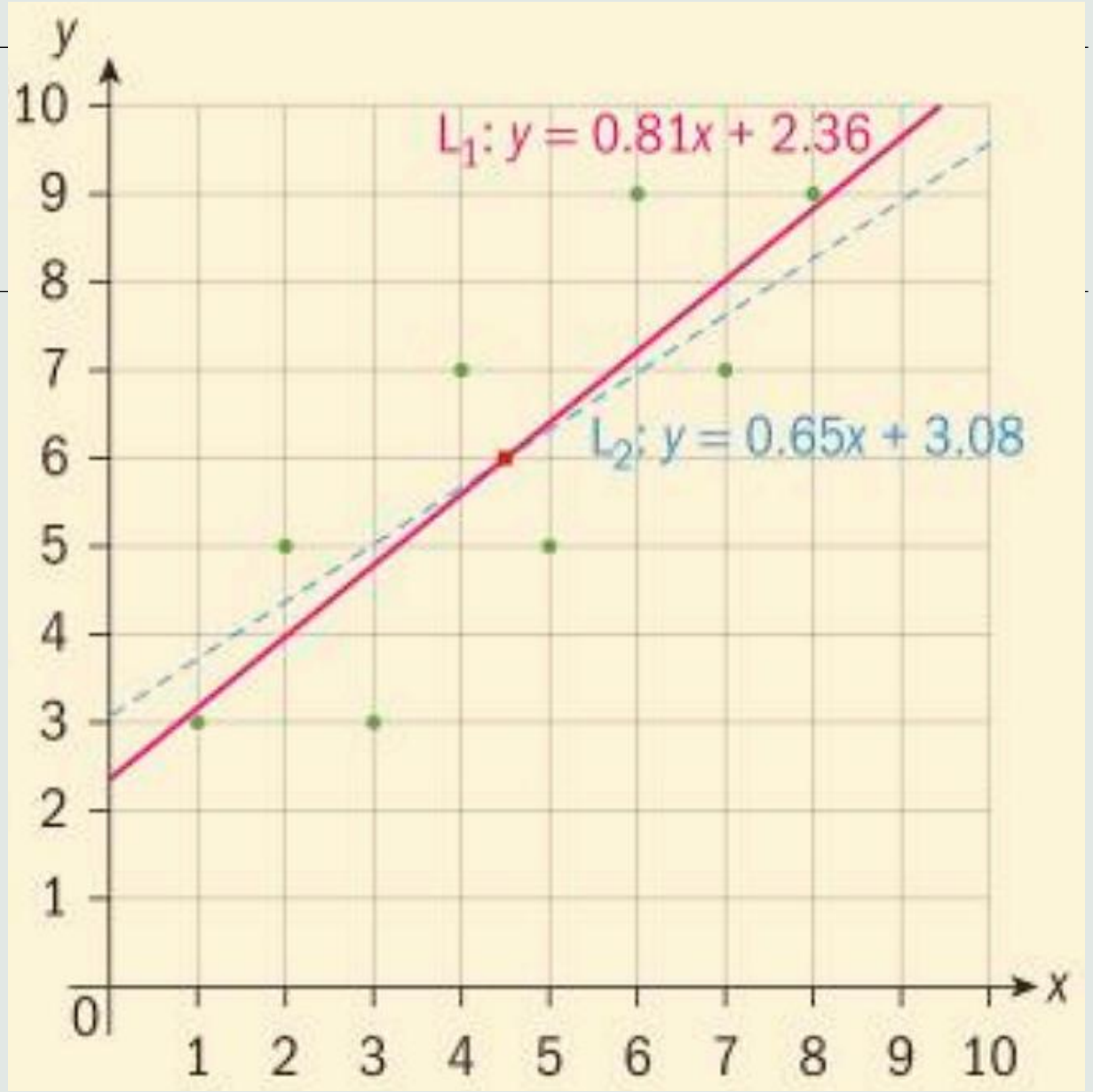
Investigation

- Will the line that best fits the data have residuals with generally smaller or larger values? Why?



Investigation

- Calculate the residuals for each line by completing the table on the next slide.



Investigation

Point	x	y	Predicted y using L_1	Residual using L_1	Square of residual using L_1	Predicted y using L_2	Residual using L_2	Square of residual using L_2
(1, 3)	1	3	$0.81\times1+2.36=3.17$	$3-3.17=-0.17$	0.0289	$0.65\times1+3.08=3.73$	$3-3.73=-0.73$	0.5329
(2, 5)	2							
(3, 3)	3							
(4, 7)	4							
(5, 5)	5							
(6, 9)	6							
(7, 7)	7							
(8, 9)	8							
					$SS_{res}=$			$SS_{res}=$

- What does a positive residual tell you about the predicted y-value compared with the actual y-value?

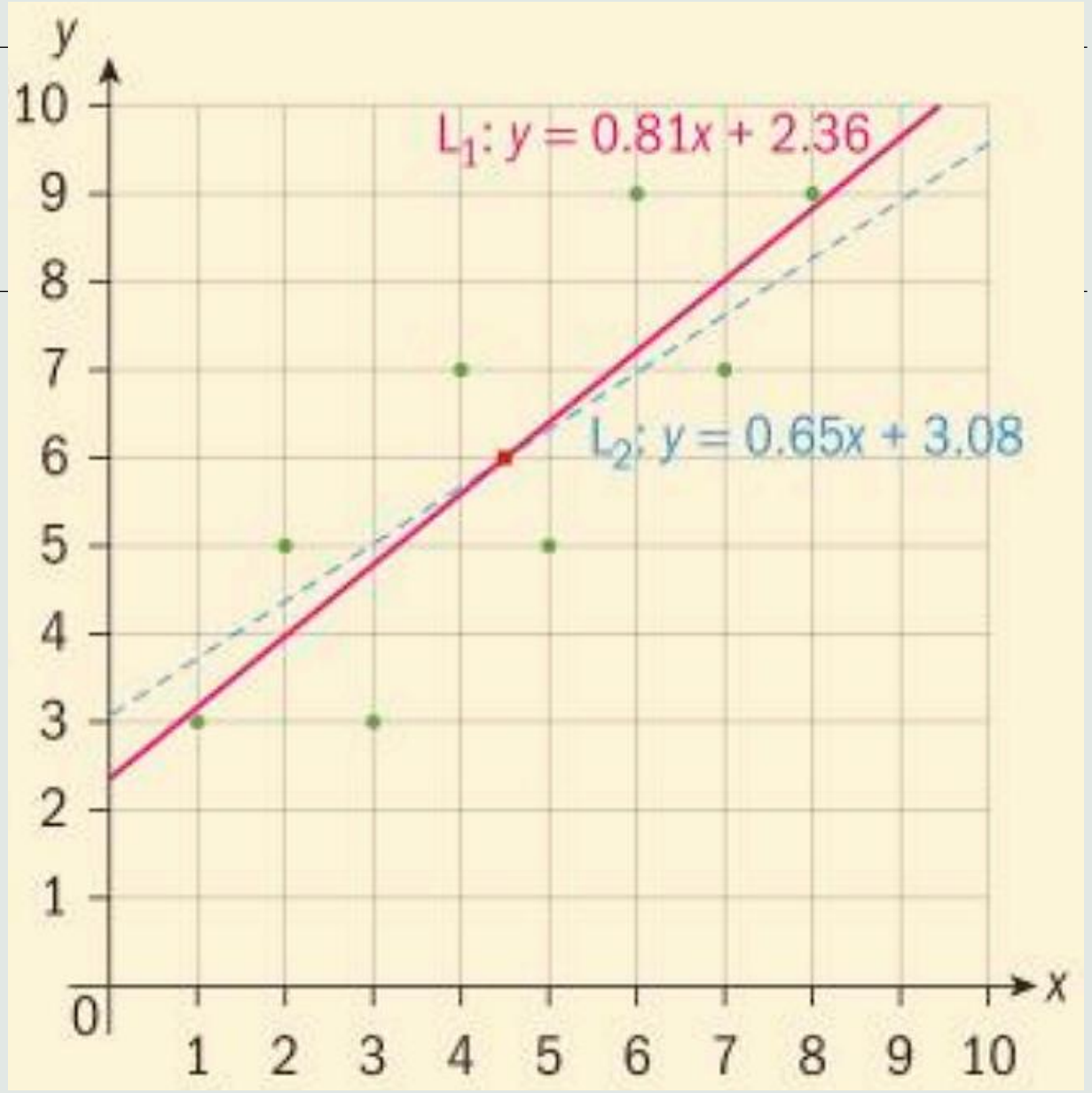
Point	x	y	Predicted y using L1	Residual using L1	Square of residual using L1	Predicted y using L2	Residual using L2	Square of residual using L2
(1,3)	1	3	3.17	-0.17	0.0289	3.73	-0.73	0.5329
(2,5)	2	5	3.98	1.02	1.0404	4.38	0.62	0.3844
(3,3)	3	3	4.79	-1.79	3.2041	5.03	-2.03	4.1209
(4,7)	4	7	5.6	1.4	1.96	5.68	1.32	1.7424
(5,5)	5	5	6.41	-1.41	1.9881	6.33	-1.33	1.7689
(6,9)	6	9	7.22	1.78	3.1684	6.98	2.02	4.0804
(7,7)	7	7	8.03	-1.03	1.0609	7.63	-0.63	0.3969
(8,9)	8	9	8.84	0.16	0.0256	8.28	0.72	0.5184
				SS res =	**12.4764**		SS res =	**13.5452**

Investigation

- Based on the SS res, which line has the better fit?

Investigation

- What happens to the size of the squares as you make the line a worse fit for the data? A better fit?



Definition

- The residual for a point (x_i, y_i) in a data set modelled by the linear function $f(x)$ is given by residual of $x_i = y_i - f(x_i)$.
- For a set of n data points, $\{(x_i\,,\ y_i)\}$ and approximating linear function $f(x), SS_{res} = \sum_{i=1}^{n}(y_i - f(x))^2$.
- The ***linear regression equation*** or the ***least squares regression line*** fits a straight line or surface that minimizes the discrepancies (also the sum of square residuals) between predicted and actual output values.
- If the vertical (y) residuals are minimized, the regression line is said to be "y on x". This line is used for predicting y-values from given x-values.

Example 1

At a coach station, the maximum temperature in °C (x) and the number of bottles of water sold (y) were recorded over 10 consecutive days. The collected data are summarized in the table.

Day	1	2	3	4	5	6	7	8	9	10
x	20	19	21	21.3	20.7	20.5	21	19.3	18.5	18
y	140	130	140	145	143	145	145	125	120	123

a. Use a graph of the data to justify why a linear regression is appropriate.

When is it appropriate to use a linear regression for prediction?

Predictions from linear regression are more accurate when the correlation coefficient is stronger. At least a moderate correlation and linear relationship should be established before making predictions from a linear regression.

Example 1

At a coach station, the maximum temperature in °C (x) and the number of bottles of water sold (y) were recorded over 10 consecutive days. The collected data are summarized in the table.

Day	1	2	3	4	5	6	7	8	9	10
x	20	19	21	21.3	20.7	20.5	21	19.3	18.5	18
y	140	130	140	145	143	145	145	125	120	123

b. Find the regression line of y on x.

Example 1

At a coach station, the maximum temperature in °C (x) and the number of bottles of water sold (y) were recorded over 10 consecutive days. The collected data are summarized in the table.

Day	1	2	3	4	5	6	7	8	9	10
x	20	19	21	21.3	20.7	20.5	21	19.3	18.5	18
y	140	130	140	145	143	145	145	125	120	123

c. Interpret the gradient and y-intercept of the regression equation in context.

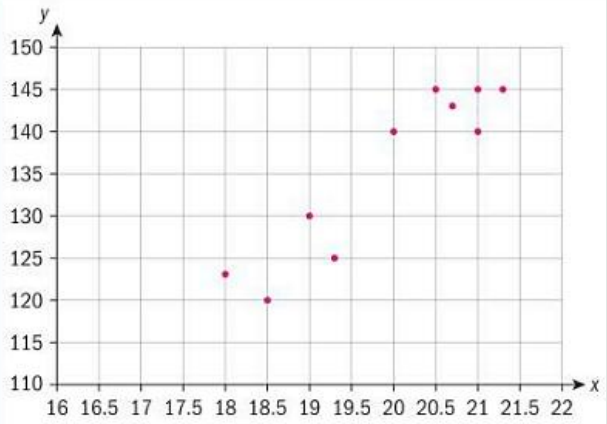
Example 1

At a coach station, the maximum temperature in °C (x) and the number of bottles of water sold (y) were recorded over 10 consecutive days. The collected data are summarized in the table.

Day	1	2	3	4	5	6	7	8	9	10
x	20	19	21	21.3	20.7	20.5	21	19.3	18.5	18
y	140	130	140	145	143	145	145	125	120	123

d. Use the regression equation to predict the number of bottles that will be sold at a temperature of 19.5℃.

a Because the data is approximately linear, linear regression is appropriate.



b $y = 8.05x - 24.7$

The GDC shows the general form of the equation as $y = ax + b$ with $a = 8.05$ (3 s.f.) and $b = -24.7$ (3 s.f.).

c The gradient of 8.05 indicates that an increase of 1°C corresponds to an increase of about 8 bottles sold.

The y-intercept is outside the range of the data set and is negative, so it does not have meaning in the context.

d 132

Substitute $x = 19.5$ in the regression equation, and solve either with technology or algebraically:

$y = 8.05(19.5) - 24.7 = 132.2$

Answers

Example 2

- Find out the relation between the number of articles written by journalists in a month and their number of years of experience. Here, the dependent variable (y) is the number of articles written and the independent variable (x) is the number of years of experience.

X	Y
5	24
10	30
4	22
1	10
3	18
$\sum x = 23$	$\sum y = 104$

$$R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$R = \frac{(5*572) - (23*104)}{\sqrt{[5*151 - (23)^2][5\sum 2384 - (104)^2]}}$$

$$= \frac{2860 - 2392}{\sqrt{(755 - 529)(11920 - 10816)}}$$

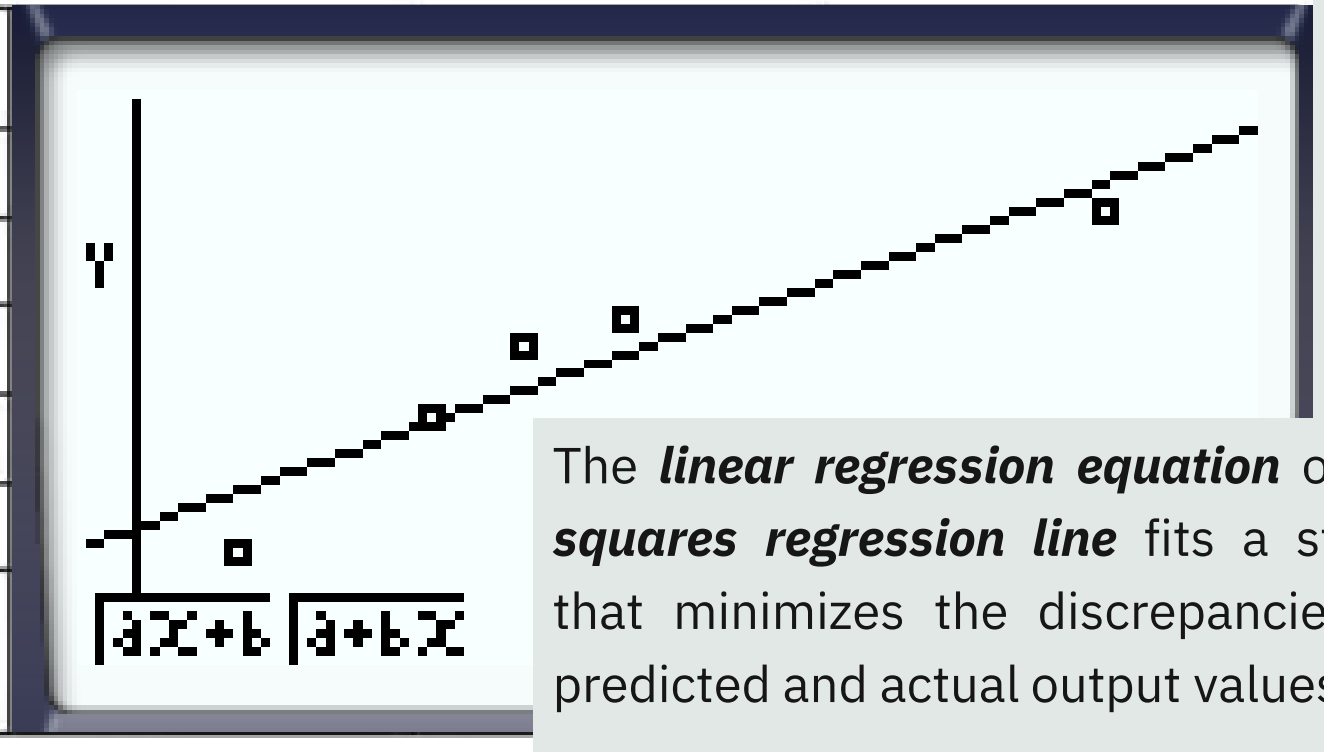
$$= 0.93693$$

$$0.93693^2 \qquad 0.8778$$

Example 2

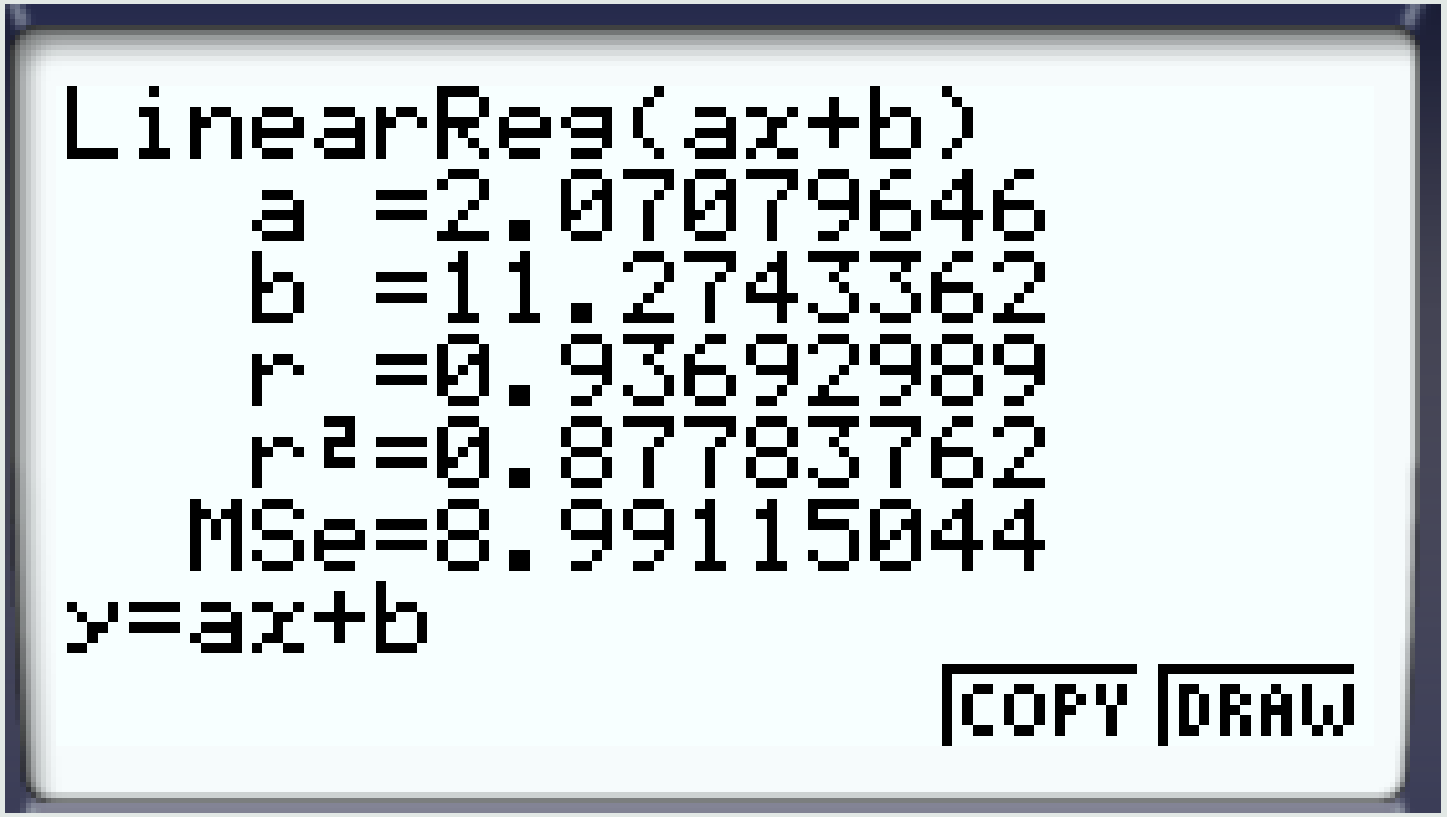
```
LinearReg(ax+b)
    a =2.07079646
    b =11.2743362
    r =0.93692989
    r²=0.87783762
  MSe=8.99115044
y=ax+b
                COPY DRAW
```

X	Y
5	24
10	30
4	22
1	10
3	18
$\sum x = 23$	$\sum y = 104$



The ***linear regression equation*** or the ***least squares regression line*** fits a straight line that minimizes the discrepancies between predicted and actual output values.

Example 2



- The correlation r measures the strength of the linear relationship between two quantitative variables.
- r^2 shows how well the data fit the regression model (the goodness of fit)

Your Turn

WEEK 8

Your Turn

1 The travel time in minutes (x) and the price in euros (y) of ten different train journeys between various places in Spain are shown in the table.

x	128	150	102	140	140	98	75	130	80	132
y	25.95	40	24.85	31.8	30.2	28.95	21.85	34.5	23.25	26

a Plot the data points on a scatter diagram. Use your diagram to justify why a linear regression is appropriate.

b Write down the equation of the regression line of y on x.

c Predict the price of a train journey of 2 hours.

d Comment on whether the regression equation be more reliable in predicting the price for a journey of 10 minutes or 100 minutes. Justify your answer.

Your Turn

2 The heights in metres (x) and weights in kilograms (y) of ten male gorillas are shown in the table.

x	1.9	1.83	1.81	1.79	1.74	1.91	1.93	1.86	1.81	1.95
y	275	267	260	257	258	272	273	268	261	273

a Plot the data points on a scatter diagram. Use your diagram to justify why a linear regression is appropriate.

b Write down the equation of the least squares regression line for this data.

c Predict the weight of a gorilla that is 1.8 m tall.

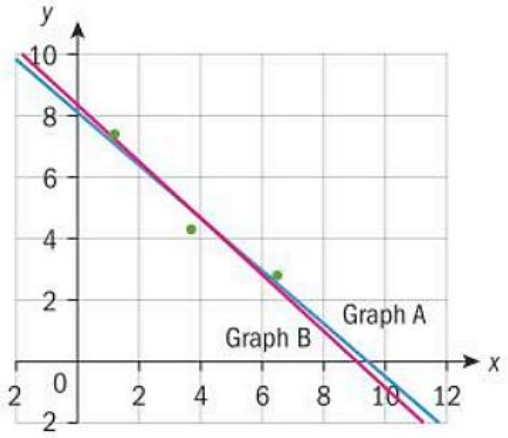
d Interpret the meaning of the gradient in context.

Your Turn

3 Two potential lines of fit for the data set shown in the table are $f_1(x) = -0.861x + 8.11$ and $f_2(x) = -0.913x + 8.30$. One of these is the linear regression equation.

x	1.2	6.5	3.7
y	7.4	2.8	4.3

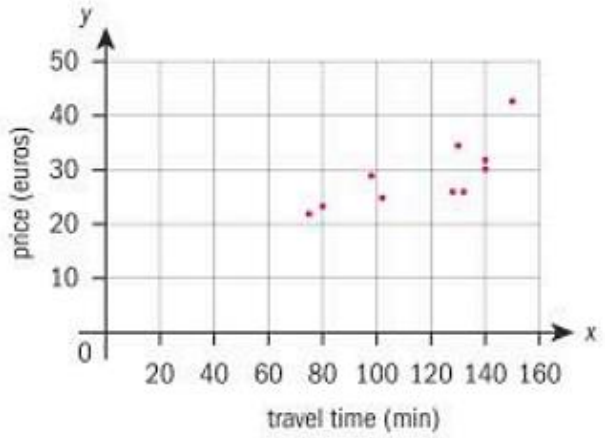
a Match each equation to its graph, with reasons.



b Calculate the sum of square residuals for each equation and hence determine which is the linear regression equation.

Answers

1 **a** A linear regression is appropriate because the data displays a roughly linear trend.

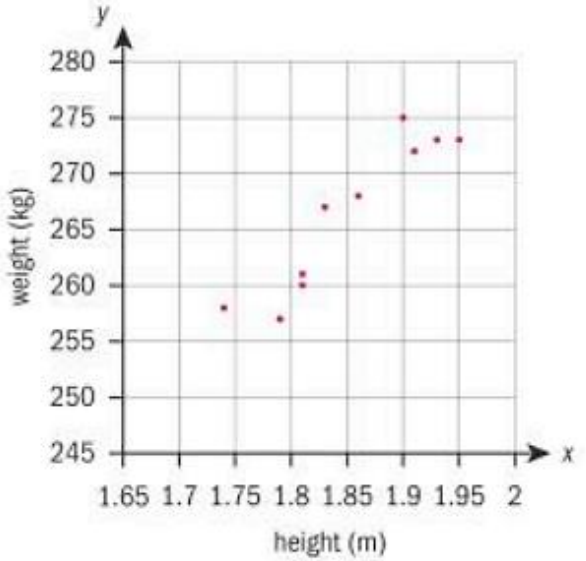


b $y = 0.16x + 9.84$

c $x = 120$, $y = 29.4$ euros

d 100 minutes by interpolation (within the data set). Predicting for 10 minutes would be extrapolation beyond the data set.

2 **a** A linear regression is appropriate because the data displays a roughly linear trend.



b $y = 93.7x + 92.8$

c $x = 1.8$, $y = 261$ kg

d A 1 cm increase in height corresponds to a 0.937 kg increase in weight.

Answers

3 **a** $f_1(x)$ is graph A because it has the lower y-intercept (8.11). $f_2(x)$ is Graph B.

b $SSR_1 = 0.576$, $SSR_2 = 0.614$
Since $f_1(x)$ has the smaller sum of square residuals, it is the least squares regression equation.

Point	x	y	Predicted y using L1	Residual using L1	Square of residual using L1	Predicted y using L2	Residual using L2	Square of residual using L2
(1.2, 7.4)	1.2	7.4	7.0768	0.3232	0.10445824	7.2044	0.1956	0.0382594
(6.5, 2.8)	6.5	2.8	2.5135	0.2865	0.08208225	2.3655	0.4345	0.1887903
(3.7,4.3)	3.7	4.3	4.9243	-0.6243	0.38975049	4.9219	-0.6219	0.3867596
	3.8	**4.833**		SS res =	**0.57629098**		SS res =	**0.6138092**