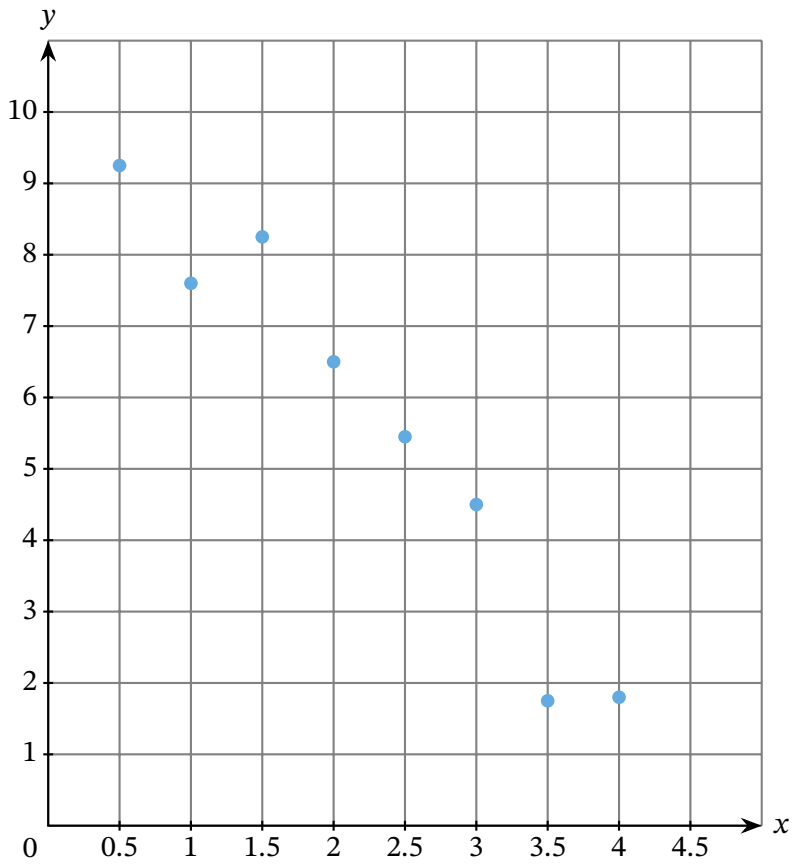
Worksheet: Least Squares Regression Line

Q1: The scatterplot shows a set of data for which a linear regression model appears appropriate.



The data used to produce this scatterplot is given in the table shown.

x	0.5	1	1.5	2	2.5	3	3.5	4
y	9.25	7.6	8.25	6.5	5.45	4.5	1.75	1.8

Calculate the equation of the least squares regression line of y on x, rounding the regression coefficients to the nearest thousandth.

A $y = 6.819 - 0.525x$

B $y = 10.657 - 2.231x$

C $y = 9.973 - 2.150x$

D $y = 4.094 + 0.686x$

2 E $y = 10.235 - 1.078x$

Q2: Using the information in the table, find the regression line $\hat{y} = a + bx$. Round a and b to 3 decimal places.

Cultivated Land in Feddan	126	13	104	180	38	161	14	99	55	177
Production of a Summer Crop in Kilograms	160	40	80	340	260	200	280	280	140	100

A $\hat{y} = 168.563x + 0.201$

B $\hat{y} = 0.201x + 168.563$

C $\hat{y} = 0.201x + 207.437$

D $\hat{y} = 0.034x + 168.563$

Q3: The table shows the price of a barrel of oil and the economic growth. Using the information in the table, find the regression line $\hat{y} = a+bx$. Round a and b to 3 decimal places.

Price of One Barrel of Oil in Dollars	50.40	55.30	63	70.70	83.60	94.10	102.50	118
Economic Growth Rate	-1	0.5	0.5	1	2.8	3.9	4.9	5

A $\hat{y} = 0.092x - 5.132$

B $\hat{y} = 0.004x - 5.132$

C $\hat{y} = -5.132x + 0.092$

D $\hat{y} = 0.092x + 9.532$

Q4: The table shows the relation between the variables x and y. Find the equation of the regression line in the form $\hat{y} = a + bx$. Approximate a and b to 3 decimal places.

x	10	22	22	13	16	21
y	25	18	24	25	12	17

A $\hat{y} = -0.376x + 26.684$

B $\hat{y} = 26.684x - 0.376$

C $\hat{y} = -0.376x + 13.649$

D $\hat{y} = -0.013x + 26.684$

Q5: Using the information in the table, estimate the value of y when $x = 13$. Give your answer to the nearest integer.

x	23	9	24	15	7	12
y	22	24	25	13	21	9

A -12

B 26

C 19

D 18

Q6: Using the information in the table, find the error in y if $x = 22$. Give your answer to the nearest integer.

x	26	22	28	15	30	10	25	29
y	5	4	12	7	14	10	13	15

A 1

B 17

C 6

D 0

Q7: The table shows the price of a barrel of oil and the economic growth. Using the information in the table, estimate the economic growth if the price of a barrel of oil is 35.40 dollars.

Price of a Barrel of Oil in Dollars	26	13.30	22.90	12.40	26.70	17.90	23.60	37.40
Economic Growth Rate	1.8	0.4	3.7	2.3	3.2	2.7	0.5	0.3

A 0.2

B 1.5

C 2.4

D 2.5

Q8: Given that points $(3, -9)$ and $(2, -4)$ lie on a regression line y on x, which of the following points does not lie on the same line?

Question Video

A $(20, -94)$

B $(16, -69)$

C $(-10, 56)$

D $(12, -54)$

Q9: The following table shows the relation between the lifespan of cars in years and their selling price in thousands of pounds. Find the equation of the line of regression in the form $\hat{y} = a + bx$, writing a and b to 3 decimal places.

Car's Lifespan (x)	5	2	2	3	5	5	1	2
Selling Price (y)	71	83	60	90	93	70	41	45

A $\hat{y} = 6.828x + 47.788$

B $\hat{y} = 0.736x + 47.788$

C $\hat{y} = 47.788x + 6.828$

D $\hat{y} = 6.828x + 90.463$

Q10: For a given data set, $\sum x = 47$, $\sum y = 45.75$, $\sum x^2 = 329$, $\sum y^2 = 389.3125$, $\sum xy = 310.25$, and $n = 8$. Calculate the value of the regression coefficient b in the least squares regression model $y = a + bx$. Give your answer correct to three decimal places.

A $b = 0.616$

B $b = -0.176$

C $b = 0.989$

D $b = -0.188$

E $b = 0.784$

Q11: The latitude (x) and the average temperatures in February (y, measured in °C) of 10 world cities were measured. The calculated least squares linear regression model for this data was $y = 35.7 - 0.713x$.

What is the interpretation of the value of -0.713 in the model?

A For every additional degree of latitude, the average temperature increased by 0.713°C.

B For every additional 0.713 degrees of latitude, the average temperature decreased by 1°C.

C It is the y-intercept of the regression line.

D It is the average temperature in February for a city of latitude 0 (on the equator).

E For every additional degree of latitude, the average temperature decreased by 0.713°C.

What is the interpretation of the value of 35.7 in the model?

A For every additional degree of latitude, the average temperature decreased by 0.713°C.

B For every additional degree of latitude, the average temperature increased by 0.713°C.

C It is the gradient of the regression line.

D For every additional 0.713 degrees of latitude, the average temperature decreased by 1°C.

E It is the average temperature in February for a city of latitude 0 (on the equator).

Q12: A city council is investing in improving their bus services. Over a five-year period, they collect data on the amount of money invested in each bus route (x, measured in 100s of dollars) and the percent of bus services that run on time (y, measured in %). They find that the data can be described by the linear regression model $y = 52.3 + 2.7x$.

- What is the interpretation of the value of 2.7 in the regression model?

A. It represents the percent of bus services that would run on time with no investment.

B. For every additional \$100 of investment, an additional 2.7% of bus services run on time.

C. It is the y-intercept of the regression line.

D. For every additional \$52.3 of investment, an additional 2.7% of bus services run on time.

- What is the interpretation of the value of 52.3 in the regression model?

A. For every additional \$100 of investment, an additional 2.7% of bus services run on time.

B. It represents the percent of bus services that would run on time with no investment.

C. It represents the percent of bus services that would run on time with \$100 of investment.

D. It is the gradient of the regression line.

Q13: The relationship between the distances jumped by competitors in the long jump (x meters) and high jump (y meters) during the women's heptathlon at the 2016 Rio Olympics can be modeled by the regression line $y = 0.218x + 0.483$.

What is the interpretation of the value 0.218 in the regression model?

A For every extra meter jumped in the high jump, the competitors jumped on average an extra 0.218 meters in the long jump.

B It is the y-intercept of the regression line.

C This is the predicted high jump result for a competitor who jumped 0 meters in the long jump competition.

D For every extra meter jumped in the long jump, the competitors jumped, on average, an extra 0.218 meters in the high jump.

What is the interpretation of the value 0.483 in the regression model?

A This is the predicted long jump result, in meters, for a competitor who jumped 0 meters in the high jump competition.

B It is the slope of the regression line.

C It is the x-intercept of the regression line.

D For every extra meter jumped in the long jump, the competitors jumped, on average, an extra 0.483 meters in the high jump.

E This is the predicted high jump result, in meters, for a competitor who jumped 0 meters in the long jump competition.

Does the interpretation of the value 0.483 seem reasonable in the context of the data?

A No, the model has been extrapolated a long way and is therefore unreliable.

B yes

Estimate, to the nearest hundredth of a meter, the expected high jump result for a competitor who jumped 6.03 m in the long jump competition.

A 5.55 meters

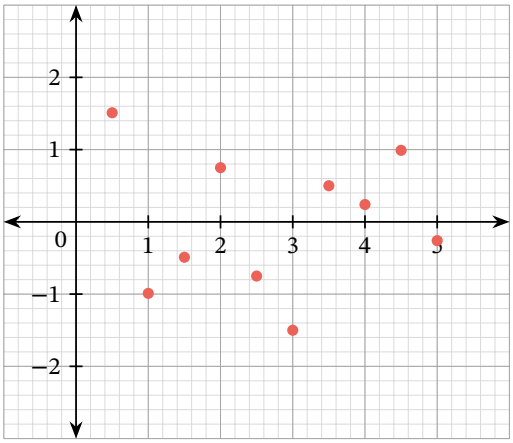
B 3.13 meters

C 4.22 meters

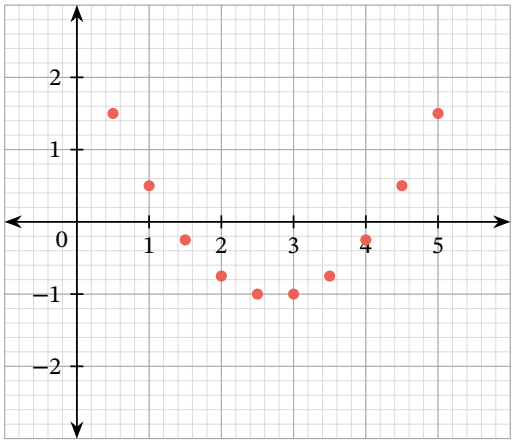
D 1.31 meters

E 1.80 meters

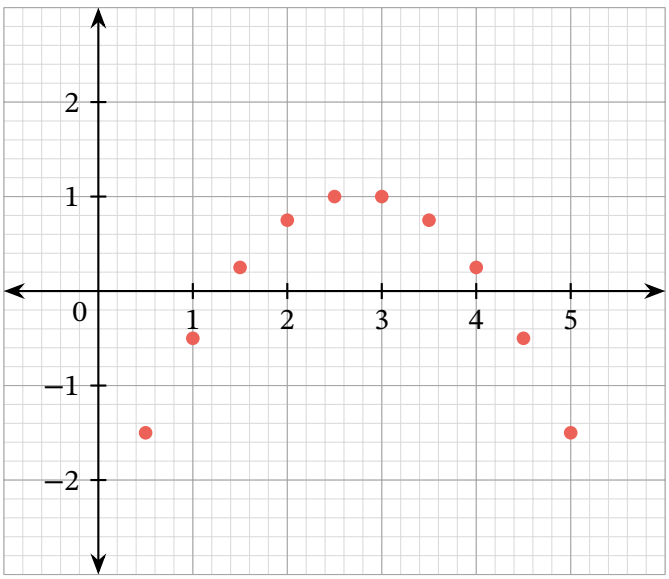
Q14: A linear model was fitted to three data sets. The residual plot for each data set is shown. For which data set is a linear model appropriate?



A



B



C

A A

B B

C C

Q15: Given the regression line $\hat{y} = 7.3x - 5.9$, find the expected value of y when $x = 30$.

A -213.1

B -224.9

C 213.1

D 224.9

Q16: An ice cream salesman records data on the number of ice creams sold each day and the temperature at midday during the April-November period. He fits a linear regression model of the form $y = a+bx$ to the data. Would you expect the regression coefficient b to be positive or negative in this context?

A negative

B positive

Q17: Two variables X and Y have a correlation coefficient of r and their mean and standard deviations are denoted by $\overline{X}, \overline{Y}, s_X$, and s_Y, respectively. Which of the following is the formula for calculating the slope, b, of the least squares regression line $Y = a + bX$?

A $\frac{s_X}{s_Y}$

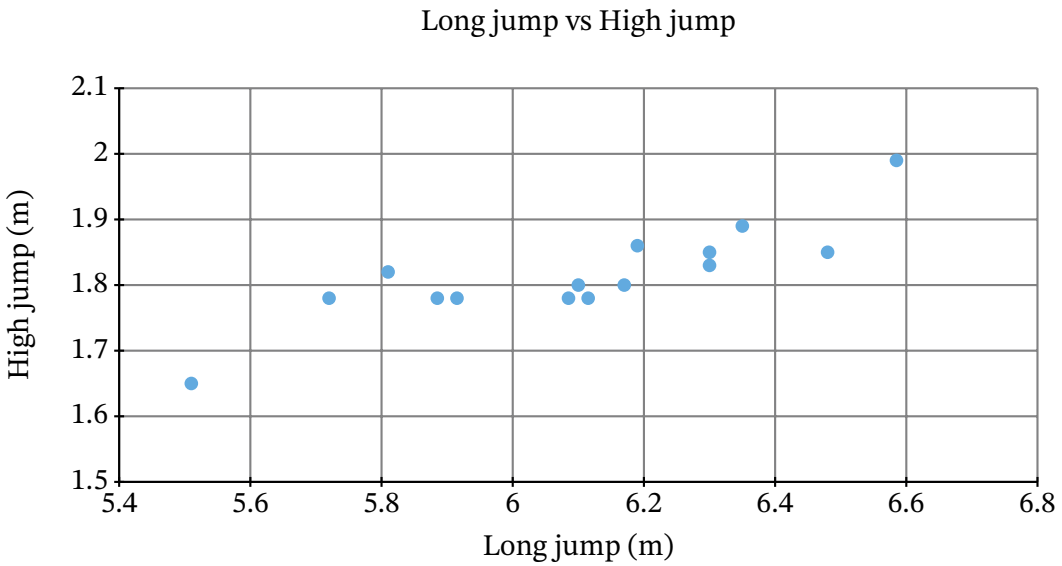
B $\frac{s_Y}{s_X}$

C $r\frac{s_X}{s_Y}$

D $\frac{s_Y}{r \cdot s_X}$

E $r\frac{s_Y}{s_X}$

Q18: The scatterplot shows the high jump and long jump results achieved by 15 competitors in the women's heptathlon competition in the 2016 Rio Olympics.



Does a linear model appear to be appropriate for modeling this data set?

A yes

B no

Would you expect the regression coefficient of this model to be positive or negative?

A positive

B negative

The data table shows the numerical data used to produce the scatter diagram.

Long Jump (m)	5.51	5.72	5.81	5.88	5.91	6.05	6.08	6.10	6.16	6.19	6.31	6.31	6.34	6.48	6.58
High Jump (m)	1.65	1.77	1.83	1.77	1.77	1.77	1.8	1.77	1.8	1.86	1.86	1.83	1.89	1.86	1.98

Representing long jump by x and high jump by y, find the values of S_{xx}, S_{yy}, and S_{xy} to the nearest thousandth.

A $S_{xx} = 1.392, S_{yy} = 0.11, S_{xy} = 0.057$

B $S_{xx} = 0.407, S_{yy} = 1.157, S_{xy} = 0.471$

C $S_{xx} = 1.002, S_{yy} = 0.121, S_{xy} = 0.637$

D $S_{xx} = 1.196, S_{yy} = 0.077, S_{xy} = 0.261$

E $S_{xx} = 0.164, S_{yy} = 0.55, S_{xy} = 1.237$

- Hence, calculate the equation of the regression line of y on x.

A $y = 1.245x + 0.483$

B $y = 0.483x + 0.218$

C $y = 0.349x + 1.157$

D $y = 0.218x + 0.483$

E $y = 1.157x + 0.349$

Q19: Liam conducted a statistical experiment to measure the number of goals as a function of the number of soccer games. With the number of soccer games as his independent variable and the number of goals as his dependent variable, the line of best fit had a slope of 2.28. What does this mean?

A The unit of the slope is 2.28 games per goal.

B For every goal, 2.28 games were played.

C The unit of the slope is 2.28 goals per game.

Q20: A variable X has a mean of 67.9 with a standard deviation of 3.1.
A variable Y has a mean of 29.3 with a standard deviation of 1.2.
Given that the correlation coefficient between X and Y is 0.37, calculate the least squares regression line of Y on X. Round the final values for a and b to 3 decimal places.

A $y = 19.575 + 0.143x$

B $y = 0.704 + 0.160x$

C $y = 28.920 + 0.857x$

D $y = 35.612 + 0.956x$

E $y = 3.0227 + 0.387x$

Q21: Use the information in the table to calculate the least squares regression line of y on x. Write the final values of the correlation coefficient and constant accurate to three decimal places.

	x	y	xy	x^2	y^2
1	22	18	396	484	324
2	22	19	418	484	361
3	23	20	460	529	400
4	26	18	468	676	324
5	31	23	713	961	529
6	32	24	768	1,024	576
7	34	22	748	1,156	484
8	37	25	925	1,369	625
9	41	29	1,189	1,681	841
10	42	27	1,134	1,764	729
Sum	310	225	7,219	10,128	5,193

A $y = 0.251x + 4.209$

B $y = 0.939x + 15.746$

C $y = 0.236x + 3.965$

D $y = 3.725x + 62.467$

E $y = 0.471x + 7.898$

Q22: If s_x and s_y are the standard deviations of x and y, and r_{xy} is the sample correlation coefficient between x and y, which of the following is the slope of a simple linear regression $y = \alpha + \beta x$?

A r_{xy}

B $r_{xy}\dfrac{s_y}{s_x}$

C $r_{xy}\left(\dfrac{s_y}{s_x}\right)^2$

D $\dfrac{s_y}{s_x}$

Q23: Amelia hits a golf ball. She knows that the height, h, of the golf ball above the ground is 0 at time $t = 0$. She suspects that subsequently h is a quadratic function of t. If Amelia is correct, plotting which of the following would give a straight line graph?

A $\dfrac{h}{t}$ against t^2

B $\dfrac{h}{t}$ against t

C h^2 against t

D h against t

E h against t^2

Q24: Jacob has collected data on the amount of fertilizer, x, he uses for each of his tomato plants, and the amount they grow as a result, y. He suspects there is a linear relationship between these two variables. If Jacob is correct, plotting which of the following would give a straight line graph?

A $\log y$ against x

B y against $\sqrt{x}$

C y against x^2

D $\sqrt{y}$ against x

E y against x

Q25: A gardener is investigating what effect the volume of weed killer used (x) has on the number of weeds (y) in her garden. She collects data and then fits a linear regression model of the form $y = a + bx$ to the data. Would you expect the regression coefficient b to be positive or negative in this context?

A positive

B negative