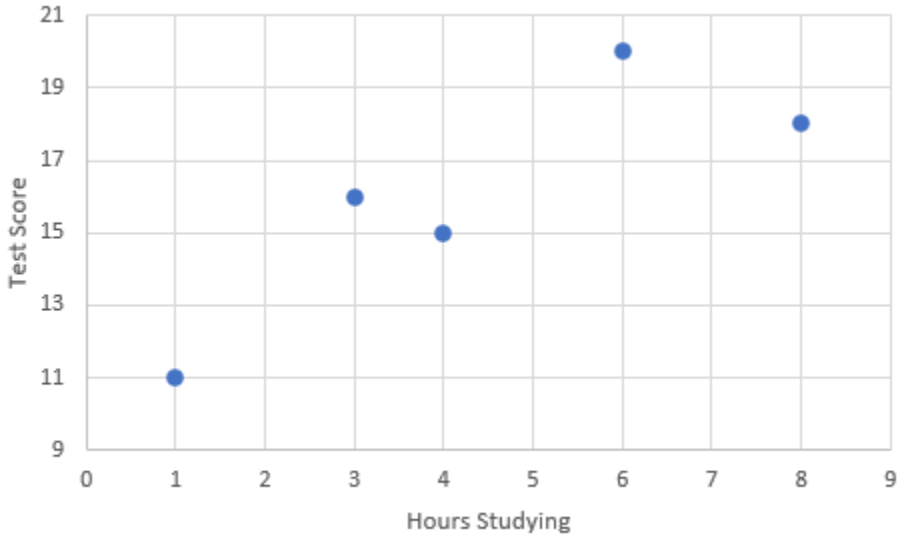
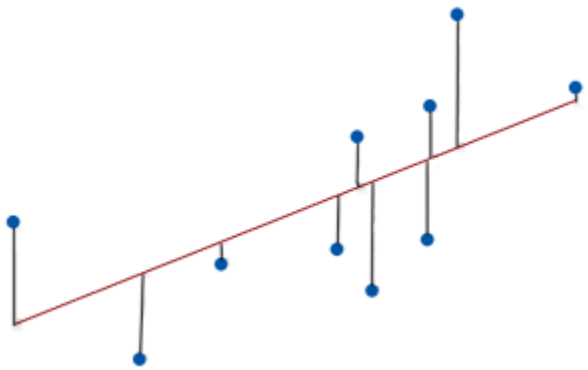
REGRESSION and CORRELATION

Least Square Regression Line

A least squares regression line represents the relationship between variables in a scatterplot. It minimizes the residual sum of squares.



Residuals are the differences between the observed data values and the least squares regression line. It is the difference between the observed value and the model's predicted value.



The **residual** of the ith data point (x_i, y_i) is

$$r_i = y_i - \hat{y}_i$$

where $\hat{y}_i$ is the predicted value of y at $x = x_i$.

$$SS_{\text{res}} = \sum_{i=1}^{n} {r_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Example 1:

The distances travelled of a group of buses are studied. The following table shows the distances travelled x (in km) of five buses and the corresponding amounts of fuel required y (in L).

Distance travelled (x km)	100	105	140	153	180
Amount of fuel required (y L)	32	38	46	50	60

It is suggested that the relationship between the variables can be modelled by the regression equation $y = 0.3x + 2$.

(a) Use the suggested model to write down the estimated amount of fuel required when a bus is travelled by

(i) 100 km;

(ii) 105 km;

(iii) 153 km.

[3]

(b) Hence, calculate SS_{res}, the sum of square residuals.

[2]

It is given that the coefficient of determination of this model is 0.976.

(c) Interpret the coefficient of determination.

[1]

Spearman Rank Correlation

Also called Spearman's rho, the Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. A monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.

Spearman's rank correlation coefficient of a bivariate data set is defined as the Pearson product-moment correlation coefficient of the variables' **ranks**.

Example 2:

The Malvern Aquatic Center hosted a 3 metre spring board diving event. The judges, Stan and Minsun awarded 8 competitors a score out of 10. The raw data is collated in the following table.

Competitors	A	B	C	D	E	F	G	H
Stan's score (x)	4.1	3	4.3	6	7.1	6	7.5	6
Minsun's score (y)	4.7	4.6	4.8	7.2	7.8	9	9.5	7.2

The Commissioner for the event would like to find the Spearman's rank correlation coefficient.

a.i. Write down the value of the Pearson's product–moment correlation coefficient, r. [2]

a.ii. Using the value of r, interpret the relationship between Stan's score and Minsun's score. [2]

b. Write down the equation of the regression line y on x. [2]

c.i. Use your regression equation from part (b) to estimate Minsun's score when Stan awards a perfect 10. [2]

c.ii. State whether this estimate is reliable. Justify your answer. [2]

d. **Copy** and complete the information in the following table. [2]

Competitors	A	B	C	D	E	F	G	H
Stan's Rank		8					1	4
Minsun's Rank		8					1	4.5

e.i. Find the value of the Spearman's rank correlation coefficient, r_s. [2]

e.ii. Comment on the result obtained for r_s. [2]

f. The Commissioner believes Minsun's score for competitor G is too high and so decreases the score from 9.5 to 9.1. [1]

Explain why the value of the Spearman's rank correlation coefficient r_s does not change.