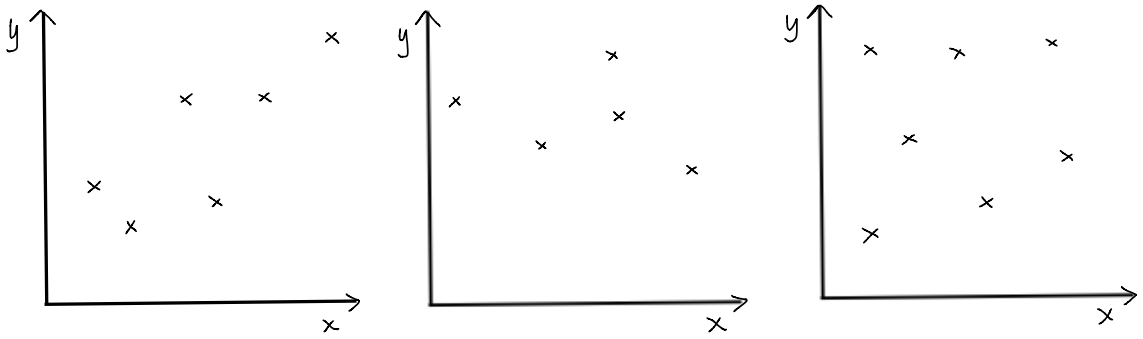
S1 -Chapters 6 and 7 - Correlation and Regression - Summary

* A scatter diagram or plot shows the association between two variables.



Positive correlation
As x increases, y increases

Negative correlation
As x increases, y decreases

No correlation

* From a scatter plot we can only determine if two variables are positively, negatively or not correlated.

To quantify the correlation we need to calculate the product moment correlation coefficient, r defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$$

where $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$

$S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$

$S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$

NOTES:

① S_{xx} and Σx^2 are NOT the same thing

② $\Sigma x^2 = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2$

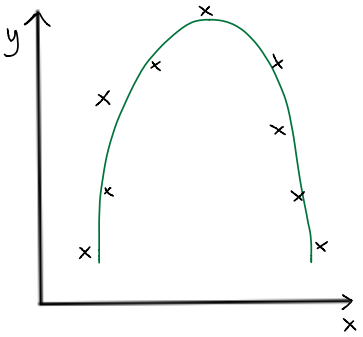
$\Sigma xy = x_1y_1 + x_2y_2 + x_3y_3 + \ldots + x_ny_n$

* r takes values between -1 and 1.
Positive values indicate positive correlation, negative values indicate negative correlation.
The closer r is to zero, the weaker the correlation.
The closer r is to -1 or 1, the stronger the correlation.

* r is NOT affected by coding.

* r is a measure of how close the points lie to a straight line. So, the weaker the correlation the further away the points are from a straight line.

* r only measures linear correlation. If r is close to zero, it means that there is no linear correlation, but this does not exclude any other forms of relationships.



eg. In this case, the relationship is obviously not linear, however r would be close to zero.

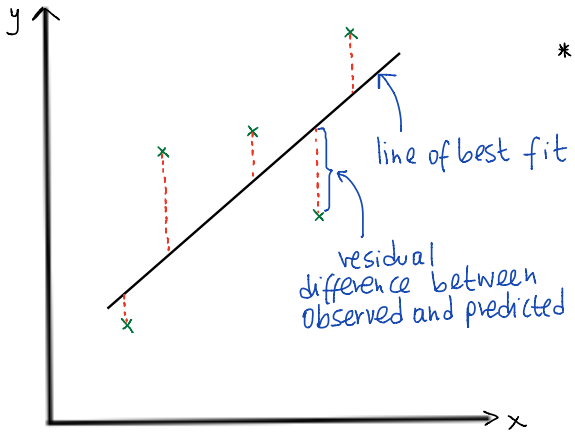
* Correlation does not imply causation ie even if two variables are associated with each other, this does not mean that a causal relationship exists between them.

* If from the scatter plot we see that the points lie roughly across a straight line (follow a linear pattern)
  OR the product moment correlation coefficient indicates strong (positive or negative) correlation indicating that points lie close to a straight line

  then a straight line may be used to model this relationship.

  This line is known as the line of best fit.

* The independent or explanatory variable is set independently of the other variable. It is plotted along the x-axis.
* The dependent or response variable has its values determined by the independent variable. It is plotted along the y-axis.



* The line of best fit or least squares regression line has equation

$$y = a + bx$$

where $b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\bar{x}$.

* b is the gradient. It represents the increase in y for every unit increase in x.

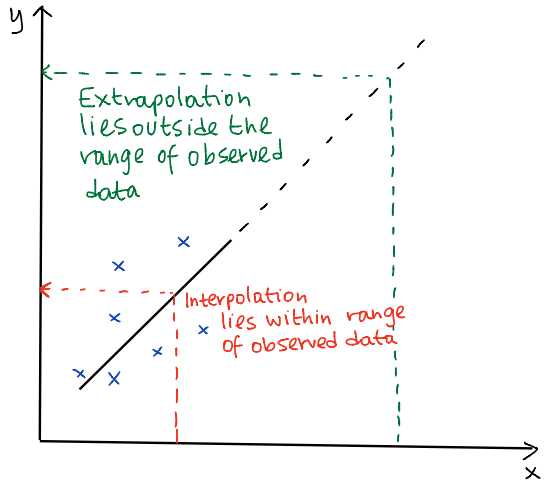
* a is the y-intercept. It represents the value of y when $x=0$.

* In case you work with coded variables, then take the coded regression line and substitute the original variables in order to retrieve the regression line involving the original variables.

* We can use the regression line to make estimates for the value of y given a value for x.

  - Interpolation is when the estimate lies within the range of observed data. Such estimates are considered reliable.

  - Extrapolation is when the estimate lies outside the range of observed data. Such an estimate can be unreliable and should always be viewed with caution.



POINTS TO NOTE

① The product moment correlation coefficient is NOT the gradient. It is a measure of how close the points lie to a straight line.

② If asked to interpret you should give your answers in context ie mention the variables involved in the question.