

DP IB Maths: AI SL



Your notes

4.2 Correlation & Regression

Contents

- * 4.2.1 Bivariate data
- * 4.2.2 Correlation Coefficients
- * 4.2.3 Linear Regression



Your notes

4.2.1 Bivariate data

Scatter Diagrams

What does bivariate data mean?

- **Bivariate data** is data which is collected on **two variables** and looks at how one of the factors affects the other
 - Each data value from one variable will be **paired** with a data value from the other variable
 - The two variables are often related, but do not have to be

What is a scatter diagram?

- A **scatter diagram** is a way of graphing bivariate data
 - One variable will be on the x-axis and the other will be on the y-axis
 - The variable that can be **controlled** in the data collection is known as the **independent** or **explanatory variable** and is plotted on the x-axis
 - The variable that is **measured** or discovered in the data collection is known as the **dependent** or **response variable** and is plotted on the y-axis
- Scatter diagrams can contain **outliers** that do not follow the trend of the data

Examiner Tip

- If you use scatter diagrams in your Internal Assessment then be aware that finding outliers for bivariate data is different to finding outliers for univariate data
 - (x, y) could be an outlier for the bivariate data even if x and y are not outliers for their separate univariate data

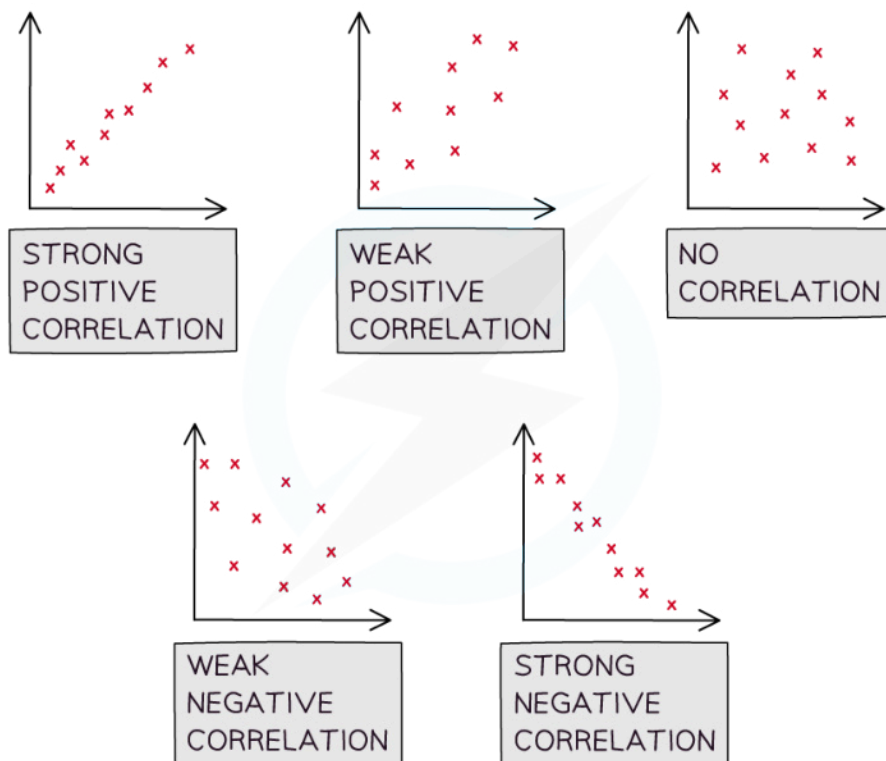


Your notes

Correlation

What is correlation?

- **Correlation** is how the **two variables change in relation to each other**
 - Correlation could be the result of a **causal relationship** but this is not always the case
- **Linear correlation** is when the changes are proportional to each other
- **Perfect linear correlation** means that the bivariate data will all lie on a straight line on a scatter diagram
- When describing correlation mention
 - The type of the correlation
 - **Positive correlation** is when an **increase** in one variable results in the other variable **increasing**
 - **Negative correlation** is when an **increase** in one variable results in the other variable **decreasing**
 - **No linear correlation** is when the data points don't appear to follow a trend
 - The strength of the correlation
 - **Strong linear correlation** is when the data points lie **close** to a **straight line**
 - **Weak linear correlation** is when the data points are **not close** to a **straight line**
- If there is **strong linear correlation** you can draw a **line of best fit** (by eye)
 - The line of best fit will pass through the mean point (\bar{x}, \bar{y})
 - If you are asked to draw a line of best fit
 - Plot the mean point
 - Draw a line going through it that follows the trend of the data



Copyright © Save My Exams. All Rights Reserved

What is the difference between correlation and causation?

- It is important to be aware that just because correlation exists, it does not mean that the change in one of the variables is **causing** the change in the other variable
 - **Correlation does not imply causation!**
- If a change in one variable **causes** a change in the other then the two variables are said to have a **causal relationship**
 - Observing correlation between two variables does **not always** mean that there is a causal relationship
 - There could be **underlying factors** which is causing the correlation
 - Look at the two variables in question and consider the context of the question to decide if there could be a causal relationship
 - If the two variables are temperature and number of ice creams sold at a park then it is likely to be a causal relationship
 - Correlation may exist between global temperatures and the number of monkeys kept as pets in the UK but they are unlikely to have a causal relationship



Your notes



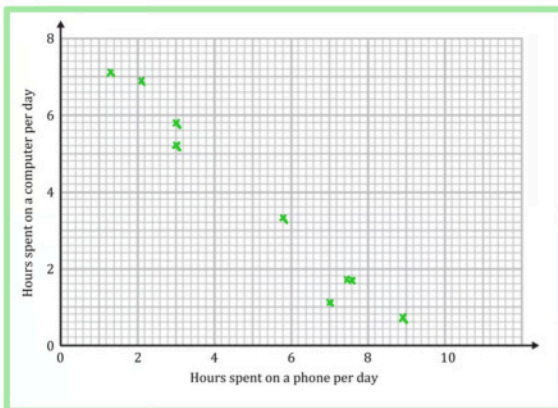
Your notes

 **Worked example**

A teacher is interested in the relationship between the number of hours her students spend on a phone per day and the number of hours they spend on a computer. She takes a sample of nine students and records the results in the table below.

Hours spent on a phone per day	7.6	7.0	8.9	3.0	3.0	7.5	2.1	1.3	5.8
Hours spent on a computer per day	1.7	1.1	0.7	5.8	5.2	1.7	6.9	7.1	3.3

a) Draw a scatter diagram for the data.



b) Describe the correlation.

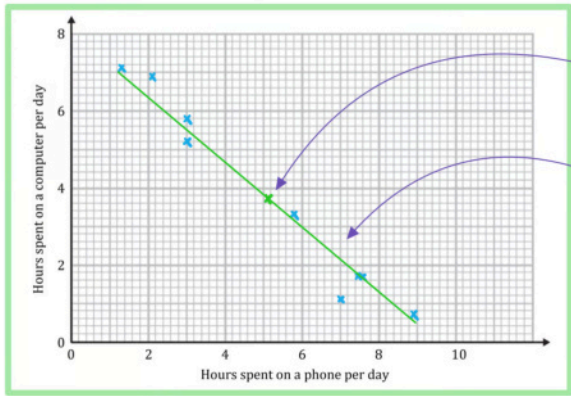
Strong negative linear correlation

c) Draw a line of best fit.



Your notes

Mean point $(\bar{x}, \bar{y}) = (5.133..., 3.722...)$



Plot the mean point

Draw it by eye



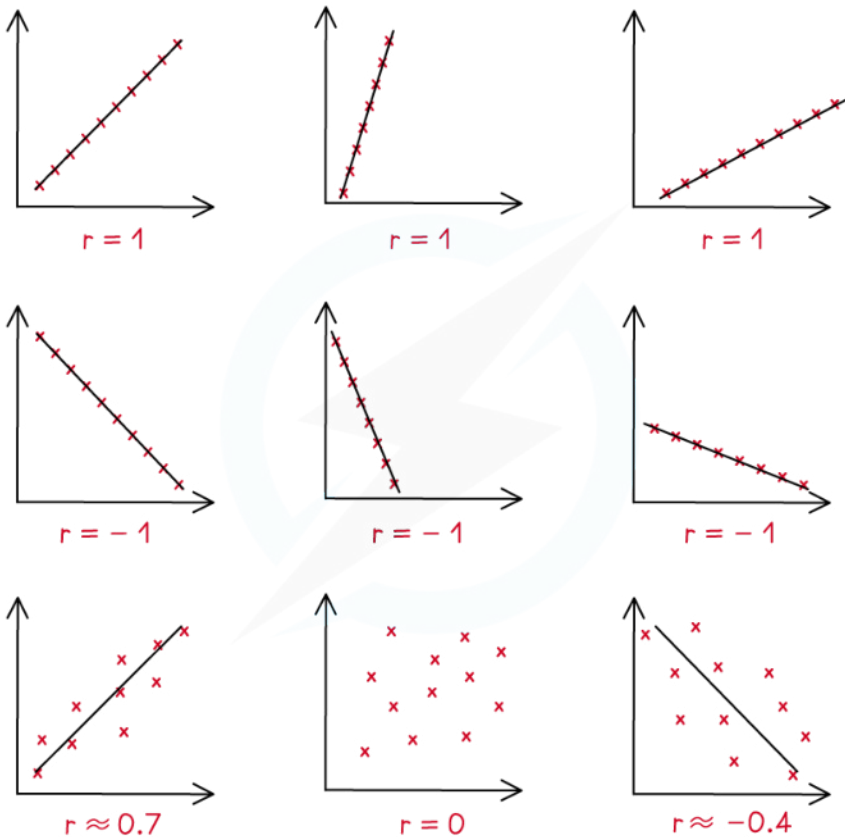
Your notes

4.2.2 Correlation Coefficients

PMCC

What is Pearson's product-moment correlation coefficient?

- Pearson's product-moment correlation coefficient (PMCC) is a way of giving a numerical value to a **linear relationship** of bivariate data
- The PMCC of a sample is denoted by the letter r
 - r can take any value such that $-1 \leq r \leq 1$
 - A **positive value** of r describes **positive correlation**
 - A **negative value** of r describes **negative correlation**
 - $r = 0$ means there is **no linear correlation**
 - $r = 1$ means **perfect positive linear** correlation
 - $r = -1$ means **perfect negative linear** correlation
 - The closer to 1 or -1 the stronger the correlation



Copyright © Save My Exams. All Rights Reserved

How do I calculate Pearson's product-moment correlation coefficient (PMCC)?

- You will be expected to use the statistics mode on your GDC to calculate the PMCC
- The formula can be useful to deepen your understanding

$$r = \frac{S_{xy}}{S_x S_y}$$

- $S_{xy} = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$ is linked to the **covariance**
- $S_x = \sqrt{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}$ and $S_y = \sqrt{\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2}$ are linked to the **variances**
- You **do not need to learn this** as using your GDC will be expected

When does the PMCC suggest there is a linear relationship?

- **Critical values** of r indicate when the PMCC would suggest there is a linear relationship
 - In your exam you will be given critical values where appropriate
 - Critical values will depend on the size of the sample
- If the **absolute value** of the **PMCC** is **bigger** than the **critical value** then this suggests a linear model is appropriate



Your notes

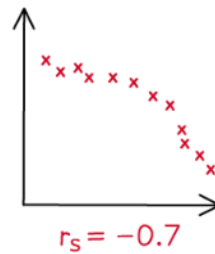
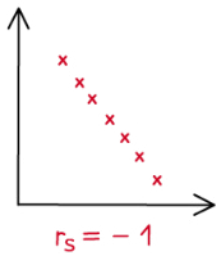
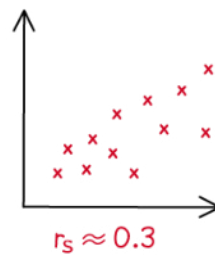
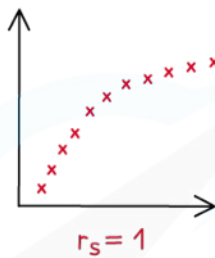
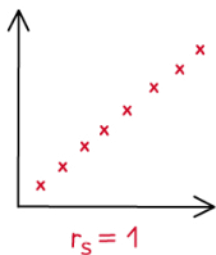


Your notes

Spearman's Rank

What is Spearman's rank correlation coefficient?

- Spearman's rank correlation coefficient is a measure of how well the relationship between two variables can be described using a **monotonic** function
 - **Monotonic** means the points are either always increasing or always decreasing
 - This can be used as a way to **measure correlation in linear models**
 - Though Spearman's Rank correlation coefficient can also be used to assess a non-linear relationship
- Each data is ranked, from biggest to smallest or from smallest to biggest
 - For n data values, they are ranked from 1 to n
 - It doesn't matter whether variables are ranked from biggest to smallest or smallest to biggest, but they must be ranked in the **same order for both variables**
- Spearman's rank of a sample is denoted by r_s
 - r_s can take any value such that $-1 \leq r_s \leq 1$
 - A **positive value** of r_s describes a **degree of agreement** between the rankings
 - A **negative value** of r_s describes a **degree of disagreement** between the rankings
 - $r_s = 0$ means the data shows **no monotonic behaviour**
 - $r_s = 1$ means the rankings are in complete agreement: the data is **strictly increasing**
 - An increase in one variable means an increase in the other
 - $r_s = -1$ means the rankings are in complete disagreement: the data is **strictly decreasing**
 - An increase in one variable means a decrease in the other
 - The **closer to 1 or -1** the **stronger the correlation** of the rankings



Copyright © Save My Exams. All Rights Reserved

How do I calculate Spearman's rank correlation coefficient (PMCC)?

- Rank each set of data independently

- 1 to n for the x -values
- 1 to n for the y -values
- If some values are equal then give each the average of the ranks they would occupy
 - For example: if the 3rd, 4th and 5th highest values are equal then give each the ranking of 4
 - $\frac{3 + 4 + 5}{3} = 4$
- Calculate the PMCC of the **rankings** using your GDC
 - This value is **Spearman's rank correlation coefficient**



Your notes



Your notes

Appropriateness & Limitations

Which correlation coefficient should I use?

- **Pearson's PMCC** tests for a **linear relationship** between two variables
 - It will not tell you if the variables have a non-linear relationship
 - Such as exponential growth
 - Use this if you are interested in a linear relationship
- **Spearman's rank** tests for a **monotonic relationship** (always increasing or always decreasing) between two variables
 - It will not tell you what function can be used to model the relationship
 - Both linear relationships and exponential relationships can be monotonic
 - Use this if you think there is a non-linear monotonic relationship

How are Pearson's and Spearman's correlation coefficients connected?

- If there is **linear correlation** then the relationship is also **monotonic**
 - $r = 1 \Rightarrow r_s = 1$
 - $r = -1 \Rightarrow r_s = -1$
 - However the **converse is not true**
- It is possible for Spearman's rank to be 1 (or -1) but for the PMCC to be different
 - For example: data that follows an **exponential growth model**
 - $r_s = 1$ as the points are always increasing
 - $r < 1$ as the points do not lie on a straight line

Are Pearson's and Spearman's correlation coefficients affected by outliers?

- Pearson's PMCC **is** affected by outliers
 - as it uses the numerical value of each data point
- Spearman's rank is **not usually** affected by outliers
 - as it only uses the ranks of each data point

Examiner Tip

- You can use your GDC to plot the scatter diagram to help you visualise the data



Your notes

Worked example

The table below shows the scores of eight students for a maths test and an English test.

Maths (x)	7	18	37	52	61	68	75	82
English (y)	5	3	9	12	17	41	49	97

- a) Write down the value of Pearson's product-moment correlation coefficient, r .

Enter data into GDC.

$$r = 0.79433\dots$$

$$r = 0.794 \text{ (3sf)}$$

- b) Find the value of Spearman's rank correlation coefficient, r_s .

Rank the data

x rank	8	7	6	5	4	3	2	1
y rank	7	8	6	5	4	3	2	1

Find PMCC of ranks

$$r_s = 0.97619\dots$$

$$r_s = 0.976 \text{ (3sf)}$$

- c) Comment on the values of the two correlation coefficients.

The value of r suggests there is strong positive linear correlation. The value of r_s suggests strong positive correlation, which is not necessarily linear.



Your notes



Your notes

4.2.3 Linear Regression

Linear Regression

What is linear regression?

- If **strong linear correlation** exists on a scatter diagram then the data can be modelled by a **linear model**
 - Drawing lines of best fit by eye is not the best method as it can be difficult to judge the best position for the line
- The **least squares regression line** is the line of best fit that minimises the **sum of the squares** of the gap between the line and each data value
 - This is usually called the **regression line of y on x**
 - It can be calculated by looking at the vertical distances between the line and the data values
- The **regression line of y on x** is written in the form $y = ax + b$
- a is the **gradient** of the line
 - It represents the change in y for each individual unit change in x
 - If a is **positive** this means y **increases** by a for a unit increase in x
 - If a is **negative** this means y **decreases** by $|a|$ for a unit increase in x
- b is the **y – intercept**
 - It shows the value of y when x is zero
- You are expected to use your **GDC** to find the equation of the regression line
 - Enter the bivariate data and choose the **model “ $ax + b$ ”**
 - Remember the **mean point** (\bar{x}, \bar{y}) will lie on the regression line

How do I use a regression line?

- The equation of the regression line can be used to decide what type of correlation there is if there is no scatter diagram
 - If a is **positive** then the data set has **positive correlation**
 - If a is **negative** then the data set has **negative correlation**
- The equation of the regression line can also be used to **predict** the value of a **dependent variable (y)** from an **independent variable (x)**
 - The equation should **only be used** to make **predictions for y**
 - Using a y on x line to **predict x is not always reliable**
 - Making a prediction **within the range** of the given data is called **interpolation**
 - This is usually reliable
 - The stronger the correlation the more reliable the prediction
 - Making a prediction **outside of the range** of the given data is called **extrapolation**
 - This is much less reliable
 - The prediction will be more reliable if the number of data values in the original sample set is bigger

Examiner Tip

- Once you calculate the values of a and b store them in your GDC
 - This means you can use the full display values rather than the rounded values when using the linear regression equation to predict values
 - This avoids rounding errors



Your notes



Your notes

Worked example

Barry is a music teacher. For 7 students, he records the time they spend practising per week (X hours) and their score in a test (y %).

Time (X)	2	5	6	7	10	11	12
Score (y)	11	49	55	75	63	68	82

- a) Write down the equation of the regression line of y on X , giving your answer in the form $y = ax + b$ where a and b are constants to be found.

Enter data into GDC

a is the coefficient of x $a = 5.5680\dots$

b is the constant term $b = 15.4136\dots$

$$y = 5.57x + 15.4$$

- b) Give an interpretation of the value of a .

$a = 5.57$ means that the model suggests that the score increases by 5.57% for every extra hour of practice.

- c) Another of Barry's students practises for 15 hours a week, estimate their score. Comment on the validity of this prediction.

Substitute $x = 15$

$$y = (5.5680...) \times 15 + (15.4136...) = 98.93..$$

The model predicts a score of 98.9% but this is unreliable as $x = 15$ is outside the range of data. Therefore extrapolation is being used.



Your notes