DP IB Maths: AA SL

Your notes

4.2 Correlation & Regression

Contents

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

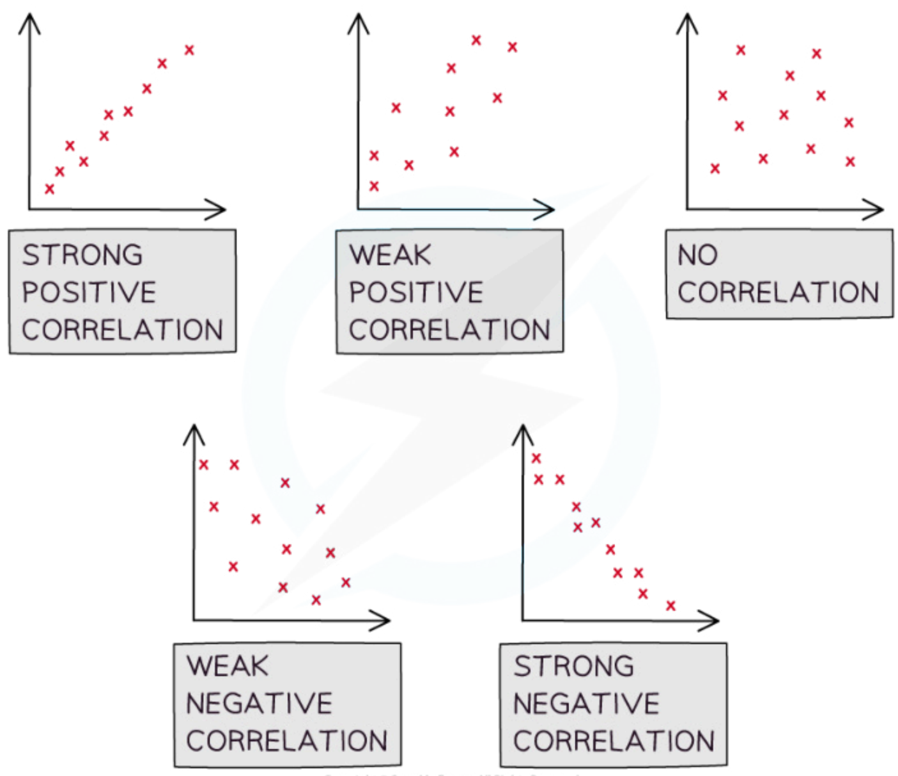
- If you use scatter diagrams in your Internal Assessment then be aware that finding outliers for bivariate data is different to finding outliers for univariate data
  - (x, y) could be an outlier for the bivariate data even if x and y are not outliers for their separate univariate data

Correlation

Your notes

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 l**ine 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



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

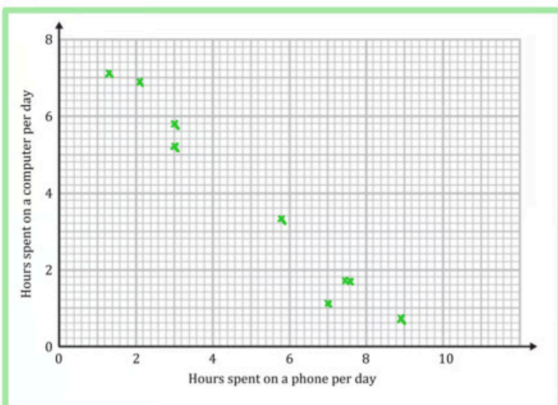
Your notes

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

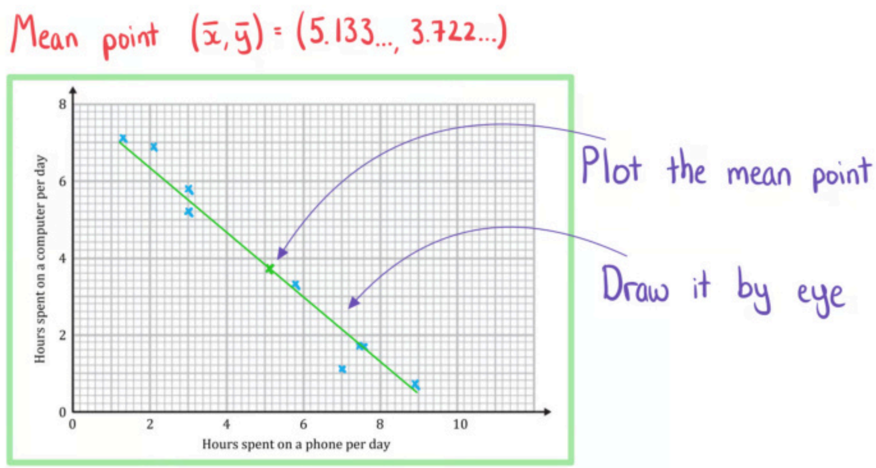
a) Draw a scatter diagram for the data.



b) Describe the correlation.

Strong negative linear correlation

c) Draw a line of best fit.


Mean point (x̄, ȳ) = (5.133..., 3.722...)
Plot the mean point
Draw it by eye
Hours spent on a computer per day
Hours spent on a phone per day
0
2
4
6
8
10

4.2.2 Correlation & Regression

Your notes

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
- It can be calculated by either looking at:
  - **vertical distances** between the line and the data values
    - This is the **regression line of *y* on *x***
  - **horizontal distances** between the line and the data values
    - This is the **regression line of *x* on *y***

How do I find the regression line of *y* on *x*?

- 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 find the regression line of *x* on *y*?

- The **regression line of x on y** is written in the form $x = cy + d$
- c is the **gradient** of the line
  - It represents the change in *x* for each individual unit change in *y*
    - If c is **positive** this means *x* **increases** by c for a unit increase in *y*
    - If c is **negative** this means *x* **decreases** by |c| for a unit increase in *y*
- *d* is the ***x* – intercept**
  - It shows the value of *x* when *y* is zero
- You are expected to use your **GDC** to find the equation of the regression line
  - It is found the same way as the regression line of *y* on *x* but with the two data sets **switched around**
  - Remember the **mean point** $(\bar{x}, \bar{y})$ will lie on the regression line

How do I use a regression line?

- The regression line can be used to decide what type of correlation there is if there is no scatter diagram
  - If the gradient is **positive** then the data set has **positive correlation**
  - If the gradient is **negative** then the data set has **negative correlation**
- The regression line can also be used to **predict** the value of a **dependent variable** from an **independent variable**
  - The equation for the *y* on *x* line should only be used to make predictions for *y*
    - Using a *y* on *x* line to predict *x* is not always reliable
  - The equation for the *x* on *y* line should only be used to make predictions for *x*
    - Using an *x* on *y* line to predict *y* 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
- The *y* on *x* and *x* on *y* regression lines intersect at the mean point $(\bar{x}, \bar{y})$

Examiner Tip

- Once you calculate the values of *a* and *b* store then 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

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.

$r = 0.79433...$

$r = 0.794$ (3sf)

b) 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.

a is the coefficient of x $\quad a = 0.943579...$

b is the constant term $\quad b = -18.05398...$

$y = 0.944x - 18.1$

c) Write down the equation of the regression line of x on y, giving your answer in the form $x = cy + d$ where c and d are constants to be found.

Swap the two sets of data

c is the coefficient of y $\quad c = 0.668700...$

d is the constant term $\quad d = 30.52410...$

$x = 0.669y + 30.5$

d) Use the appropriate regression line to predict the score on the maths test of a student who got a score of 63 on the English test.

Your notes

$y = 63$ so use x on y line

$x = (0.668700\ldots) \times 63 + (30.52410\ldots) = 72.652\ldots$

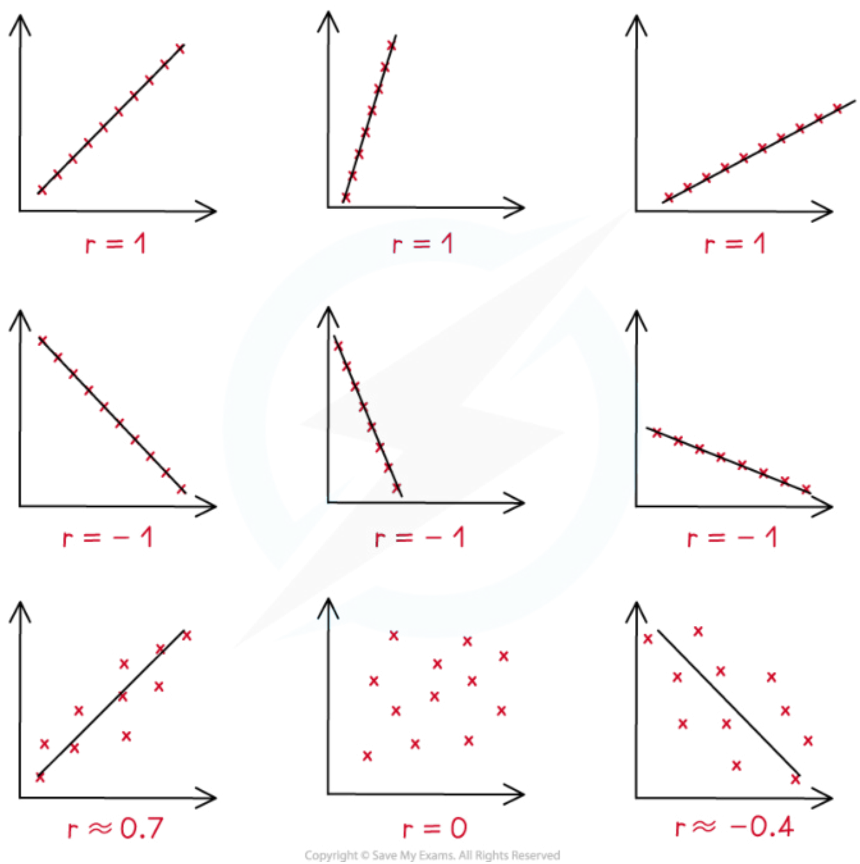
Maths score 72.7

Your notes

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 $-1 \le r \le 1$
  - 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



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

Your notes

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

- $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$ 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