Edexcel International AS Maths: Statistics 1

Your notes

Correlation & Regression

Contents

Scatter Diagrams

What does bivariate data mean?

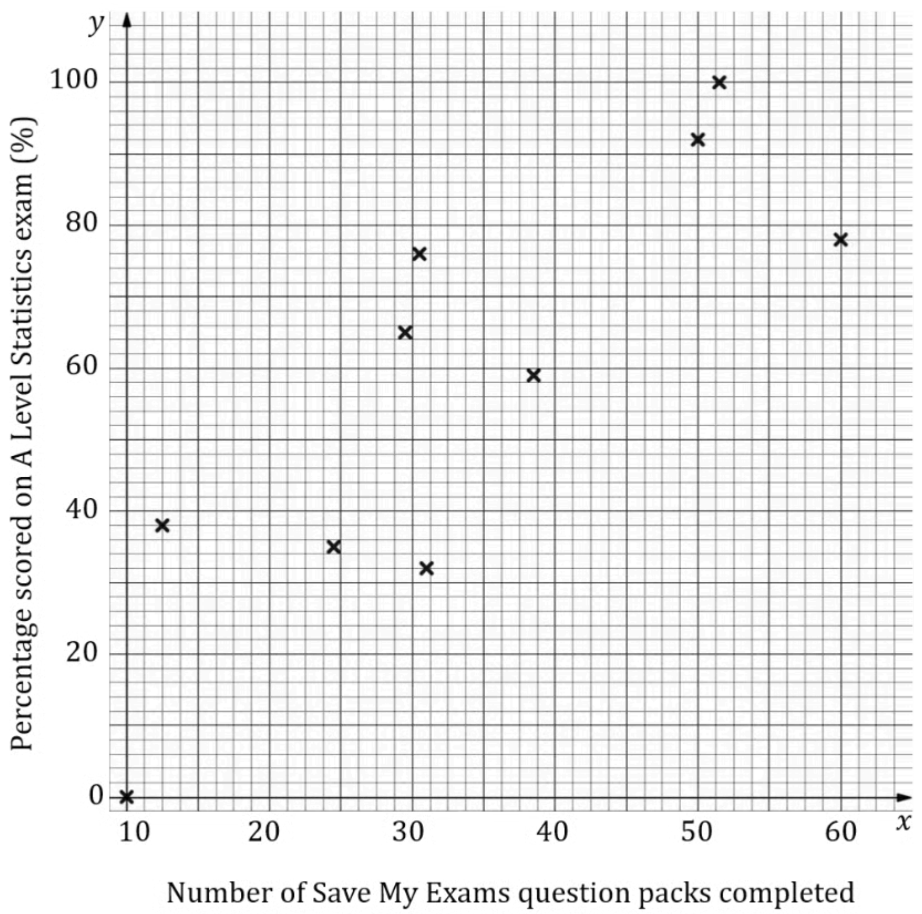
- A lot of statistics is about looking at how different factors, or **variables** change how data behaves
- **Bivariate data** is data which is collected on two variables and looks at how one of the factors affects the other
    - Each data value from one variable will be paired with a data value from 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
    - You may be asked to plot, or add to, a scatter diagram
    - 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 allow statisticians to look for relationships between the two variables
    - Some scatter diagrams will show a clear relationship know as **correlation** (see below)
    - Others will not display on obvious relationship
    - If a scatter diagram shows a relationship you may be asked to identify **outliers**

Worked Example

The scatter diagram below shows the number of Save My Exams question packs completed by a group of students and the percentage score they received in their A-Level Statistics exam.



(i) State which of the variables is the explanatory variable and which is the response variable.

(ii) Explain why the number of question packs completed is on the X – axis.

(iii) Another student completed 50 question packs and scored 80% on their A Level Statistics exam, add this data to the scatter diagram.

Answer:

(i) Explanatory variable: can be controlled $\longrightarrow$ number of question pack completed.

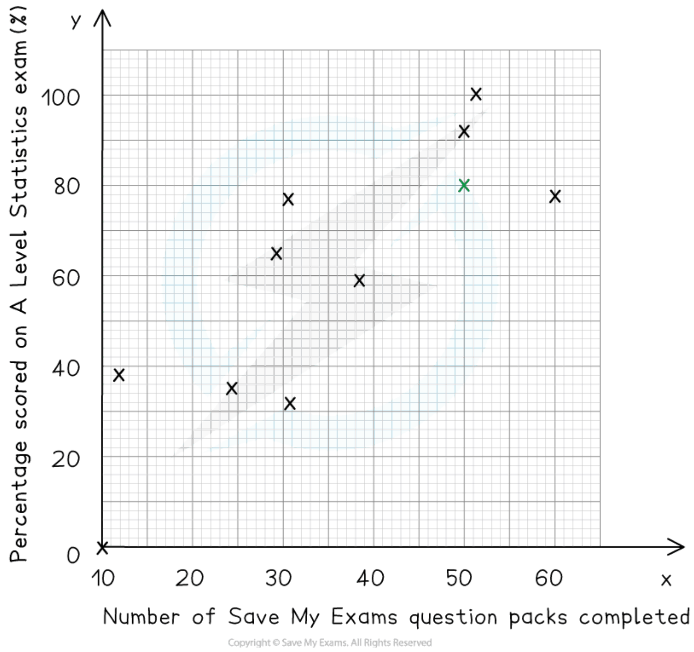
Response variable: is measured $\longrightarrow$ result in exam

The number of question pack completed is the explanatory variable.
The percentage scored on Statistics is the response variable.

(ii) The number of question packs is the variable that can be controlled so it should go on the x-axis.

Your notes

(iii) Plot the point (50, 80)



Your notes

Examiner Tips and Tricks

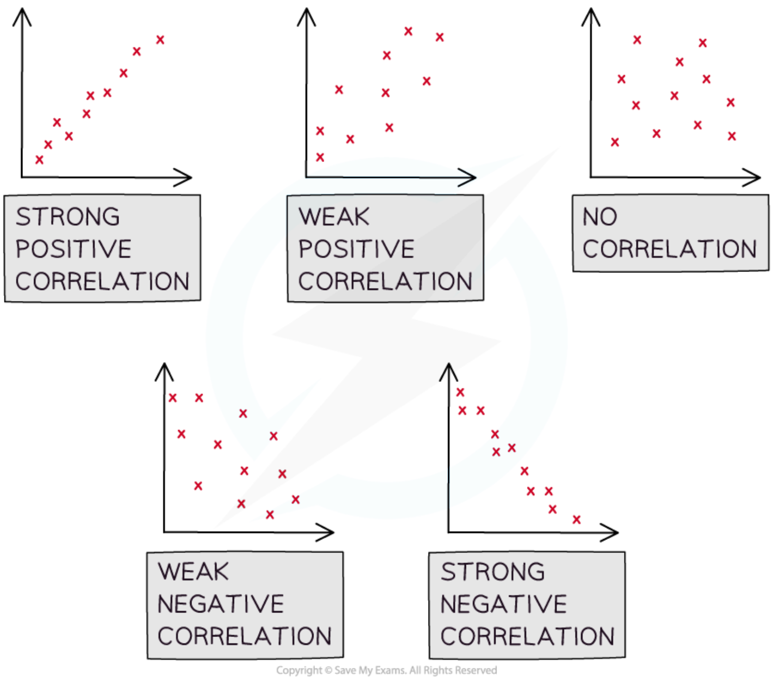
- Learn the vocabulary for the types of variables as you could be asked a question on this. Make sure you check the scales carefully when plotting any points.

Correlation

What is correlation?

- **Correlation** is how the relationship between the two variables is described
- **Perfect linear correlation** means that the bivariate data will all lie on a straight line on a scatter diagram
- Linear correlation can be **positive** or **negative** and it can be **strong** or **weak**
  - **Positive correlation** describes a data set where both variables are increasing
  - **Negative correlation** describes a data set where one variable is increasing and the other is decreasing
- When describing correlation you should say whether it is positive or negative and also say whether it is strong or weak
- If correlation exists then there could be outliers, these will be data points that do not fit the pattern seen on the graph
  - There will likely be a maximum of one or two outliers on any scatter diagram

- You may be asked to identify the outliers



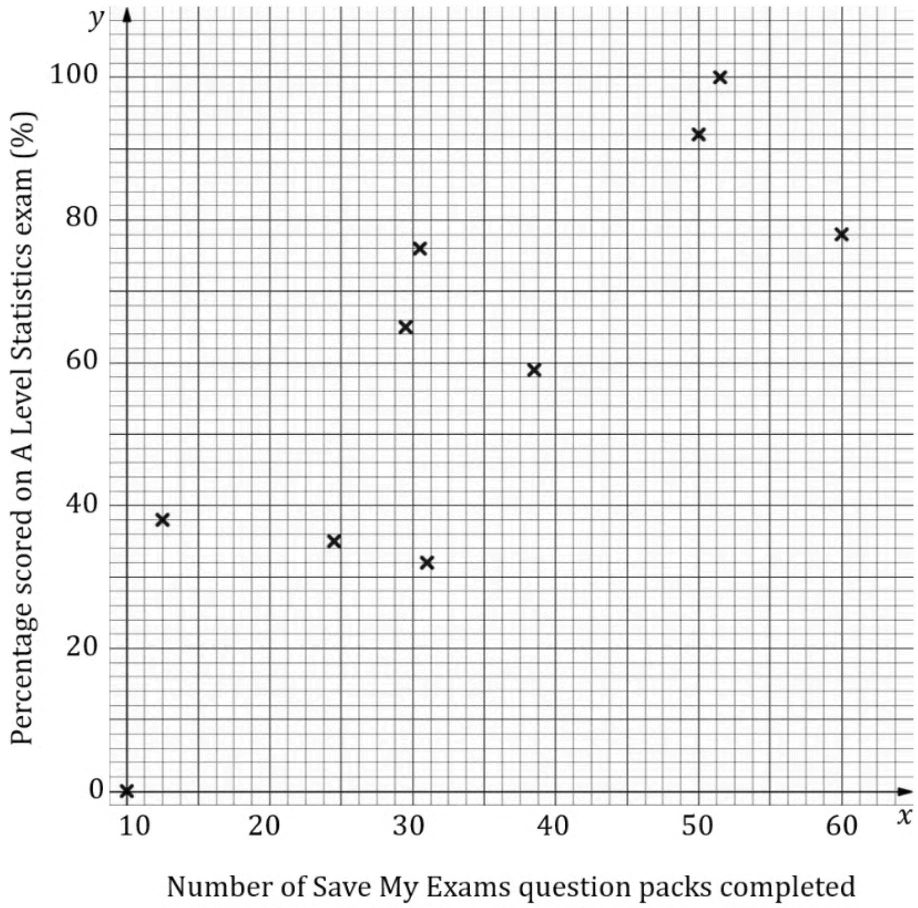
What is the difference between correlation and causation?

- 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
  - 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
  - Observing a relationship between two variables can allow you to create a hypothesis about those two variables

Worked Example

Your notes

The scatter diagram below shows the number of Save My Exams question packs completed by a group of students and the percentage score they received in their A-Level Statistics exam.



(i) Describe the correlation shown in the scatter diagram.

(ii) Decide if you think there could be a causal relationship between the two variables and explain your reasoning.

Answer:

(i) The scatter diagram shows fairly strong, positive correlation

(ii) The scatter diagram shows that the more Save My Exams question packs that are completed the greater the exam result.

It is likely that there is a causal relationship between the number of Save My Exams question packs completed and the result in the student's A level Statistics exam.

Linear Regression

Your notes

What is linear regression?

- If strong linear correlation exists on a scatter diagram, then a **line of best fit** can be drawn
  - This is a **linear graph** added to the scatter diagram that best approximates the relationship between the two variables
  - Previously will have been drawn by eye as a line that fits closest to the data values
  - The data can be used to calculate the equation of the straight line that represents the best fit of the relationship between the two variables
- 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 and can be calculated either by looking at the vertical or the horizontal distances between the line and the data values
  - If the regression line is calculated by looking at the vertical distances it is called the **regression line of y on x**
  - If the regression line is calculated by looking at the horizontal distances it is called the **regression line of x on y**
    - The regression line of x on y is rarely used and you are unlikely to come across it at this level
- The **regression line of y on x** is written in the form $y = a + bx$
  - *b* is the **gradient** of the line and represents the change in y for each individual unit change in x
  - *a* is the y – intercept and shows the value of y for which x is zero
  - It is useful to know that the point $(\bar{x}, \bar{y})$ will lie on the regression line

How to use a regression line?

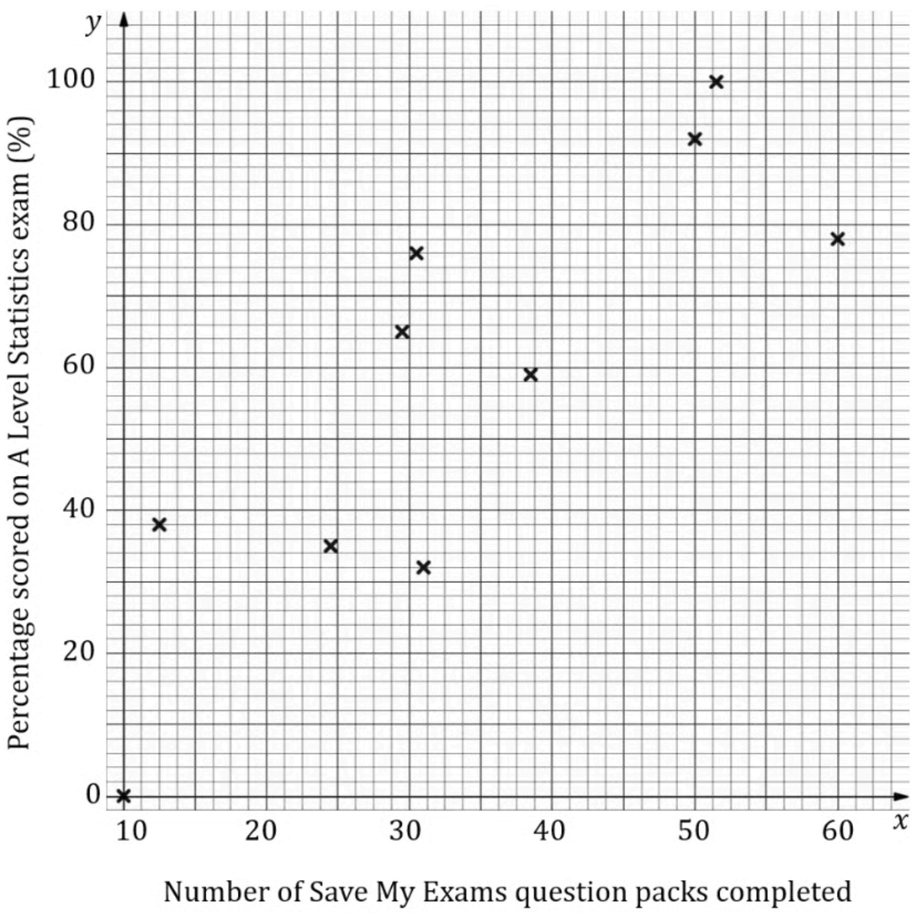
- Drawing a regression line is done in the same way as drawing a straight line graph, substitute some values from the **independent data set** to help you
- The regression line can be used to decide what type of correlation there is if there is no scatter diagram
  - If b is positive then the data set has **positive correlation** and if b is negative then the data set has **negative correlation**
- The value of *b* can be used to interpret how the data is changing
  - *b* is the **gradient** of the line and represents the change in *y* for each individual unit change in x
- The regression line can also be used to **predict** the value of a **dependent variable** from an **independent variable**
  - Predictions should only be made for values of the dependent variable that are within the range of the given data

- Making a prediction within the range of the given data is called **interpolation**
- Making a prediction outside of the range of the given data is called **extrapolation** and is much less reliable
- The prediction will be more reliable if the number of data values in the original sample set is bigger

Your notes

Worked Example

The scatter diagram below shows the number of Save My Exams question packs completed by a group of students, x, and the percentage score they received in their A-Level Statistics exam, y.



The equation of the regression line of y on x is $y = 18 + 1.3x$

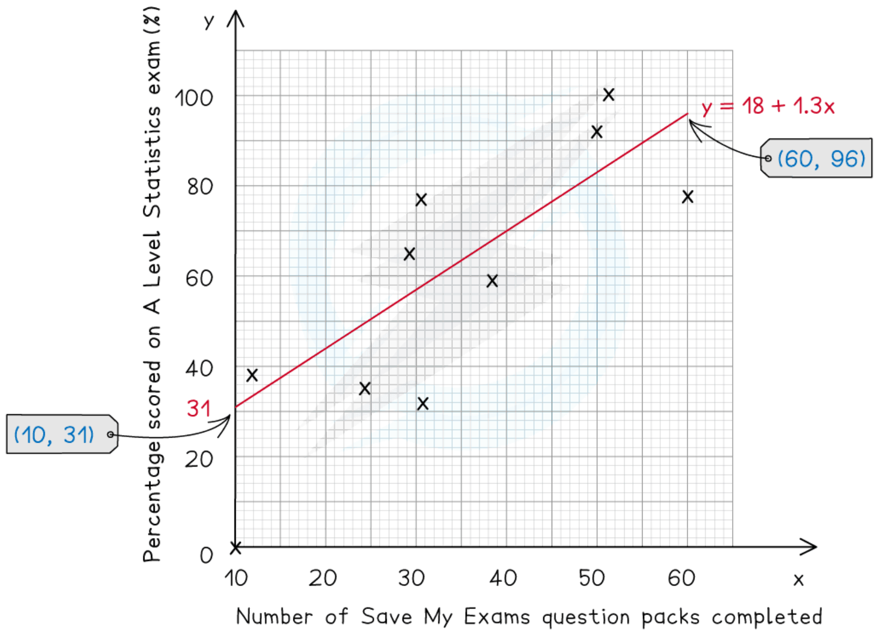
(i) Draw the regression line onto the scatter diagram.

(ii) Interpret the meaning of the values 18 and 1.3 in the scatter diagram.

(iii) Explain why the regression line given should not be used to estimate the percentage when someone has completed 80 question packs.

Answer:

(i) Be careful here as the x–axis does not begin at zero.
at $x = 10$, $y = 18 + 1.3(10) = 31\%$
at $x = 60$, $y = 18 + 1.3(60) = 96\%$



(ii) 18 is the expected score if no question packs were completed.
1.3 is the increase in percentage score for every completed question pack.

(iii) The regression line should only be used to make predictions for data values that are within the range of the original data.

80 is outside of the range of the data used to form the regression line, this would be extrapolation and should not be done.

Examiner Tips and Tricks

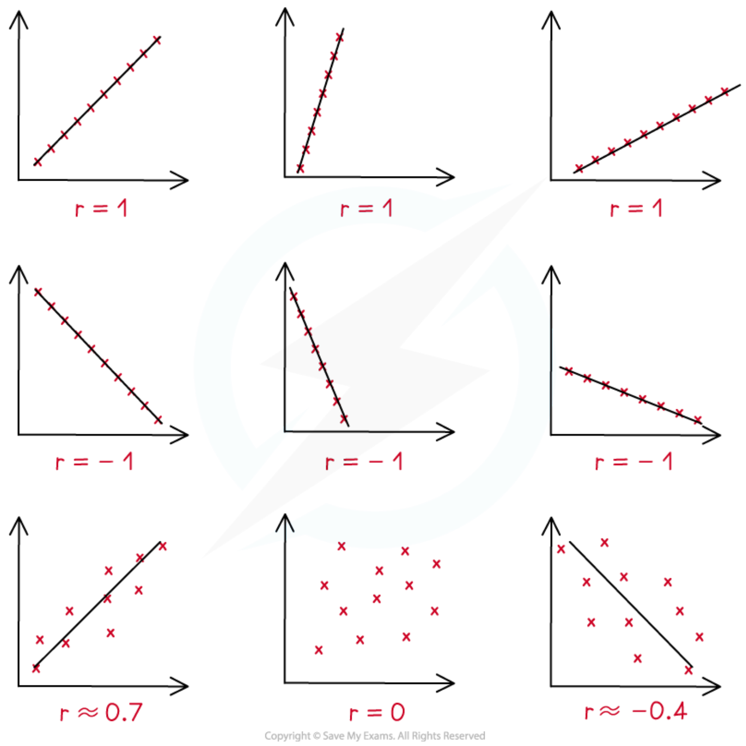
- Remember that the value of b is the gradient of the regression line, a greater value of b **does not** mean stronger correlation. When using a regression line to make a prediction make sure that the value you are predicting from falls within the range of the data used to calculate the regression line.

Your notes

PMCC

What is the product moment correlation coefficient?

- The product moment correlation coefficient (PMCC) is a way of giving a numerical value to **linear correlation** of bivariate data
- The PMCC of a sample is denoted by the letter *r*
  - *r* can take any value such that $-1 \leq r \leq 1$
    - Can be written as $|r| \leq 1$
  - A positive value of *r* describes positive correlation
  - A negative value of *r* describes negative correlation
  - If *r = 0* there is no correlation
  - *r = 1* means perfect positive correlation and *r = -1* means perfect negative correlation
  - The closer to 1 or -1, the stronger the correlation
- The gradient of the regression line does not change the value of *r*



How is the product moment correlation coefficient (PMCC) calculated?

- For *n* pairs of bivariate data (*x*, *y*) we define the following statistics
  - $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$
  - $S_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$
  - $S_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$
  - These are given in the formula booklet
- These are related to variance and can be written in several different ways:
  - S_{xx}
    - $\Sigma(x - \bar{x})^2$
    - $\Sigma x^2 - n\bar{x}^2$
    - $n\sigma_x^{\ 2}$
  - S_{xy}
    - $\Sigma(x - \bar{x})(y - \bar{y})$
    - $\Sigma xy - n\bar{x}\,\bar{y}$
- The product moment correlation coefficient (PMCC) is then calculated using the formula
  - $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$
  - This is given in the formula booklet

Calculating Regression Line

If the PMCC is close to 1 or -1 then this suggests the data follows a linear model. In this case a regression line of the form *y* = *a* + *bx* is appropriate.

How do I calculate the equation of the regression line of y on x?

- The **gradient *b*** of the **regression line** is calculated using the formula
  - $b = \frac{S_{xy}}{S_{xx}}$
  - This is given in the formulae booklet

Your notes

- The **y-intercept** *a* of the **regression line** is calculated using the formula
  - $a = \bar{y} - b\bar{x}$
  - This is given in the formulae booklet
  - This is found using the fact that the point $(\bar{x}, \bar{y})$ lies on the regression line
- If you are asked to find the equation of the regression line of ***x* on *y***
  - $x = c + dy$
  - $d = \frac{S_{xy}}{S_{yy}}$
  - $c = \bar{x} - d\bar{y}$
  - These are not given in the formulae booklet

Your notes

Worked Example

Ashika is a football coach to 20 children. She records how long it takes each of them to run a lap of the football pitch, p seconds, and the distance that they can kick the football, d metres.

Ashika calculates the following summary statistics:

$$S_{pp} = 687.2 \quad \bar{p} = 62.8 \quad \Sigma d = 1566 \quad \Sigma d^2 = 124240 \quad \Sigma pd = 99127$$

.

(a) Calculate S_{pd}.

(b) Calculate the product moment correlation coefficient between p and d.

(c) Calculate the equation of the regression line of d on p giving your answer in the form $d = a + bp$

Answer:

a) Calculate Σp using $\bar{p} = \frac{\Sigma p}{n}$

$62.8 = \frac{\Sigma p}{20} \Rightarrow \Sigma p = 1256$

Calculate S_{pd} using $S_{pd} = \Sigma pd - \frac{(\Sigma p)(\Sigma d)}{n}$

$S_{pd} = 99127 - \frac{(1256)(1566)}{20}$

$S_{pd} = 782.2$

b) Calculate S_{dd} using $S_{dd} = \Sigma d^2 - \frac{(\Sigma d)^2}{n}$

$S_{dd} = 124240 - \frac{(1566)^2}{20} = 1622.2$

Calculate r using $r = \frac{S_{pd}}{\sqrt{S_{pp} \times S_{dd}}}$

$r = \frac{782.2}{\sqrt{687.2 \times 1622.2}} = 0.74083...$

$r = 0.741$ (3.s.f.)

c) Identify x and y

$y = a + bx$

$d = a + bp$

$y = d$ and $x = p$

Calculate b using $b = \frac{S_{xy}}{S_{xx}} = \frac{S_{pd}}{S_{pp}}$

$b = \frac{782.2}{687.2} = 1.1382...$

Calculate $\bar{d}$ using $\bar{d} = \frac{\Sigma d}{n}$

$\bar{d} = \frac{1566}{20} = 78.3$

Calculate a using $a = \bar{y} - b\bar{x} = \bar{d} - b\bar{p}$

$a = 78.3 - (1.1382...)(62.8)$

$a = 6.8183$

Use answer button on calculator

$a = 6.82 + 1.14p$ (3.s.f.)

Your notes

Examiner Tips and Tricks

- Questions typically use different variables instead of *x* and *y*. It might help to label the independent variable as *x* and the dependent variable as *y*, this will help you when calculating the equation of the regression line.

Coding with PMCC

Does coding affect the product moment correlation coefficient (PMCC)?

- We can code data using **linear transformations**
  - $X = px + q$
  - $Y = my + n$
    - The new variables won't necessarily be the upper case versions of the old variables - they could be any letter
- S_{xx}, S_{yy} and S_{xy} are related to **variances** so they are affected when the coding involves a multiplication
  - $S_{XX} = p^2 \times S_{xx}$
  - $S_{YY} = m^2 \times S_{yy}$
  - $S_{XY} = pm \times S_{xy}$
- Coding **does not affect** the product moment correlation coefficient
  - The factors of p and m cancel out in the formula
  - $r_{XY} = r_{xy}$

Coding Linear Regression

Does coding affect the equation of the regression line of y on x?

- Coding **does affect** the equation of the regression line of y on x
- Coding is used to make numbers simpler to work with
- The equation of the regression line of Y on X can be calculated using the coded data
- This equation can then be used to find the equation of the regression line of y on x

How do I use coding to find the equation of the regression line of y on x?

- Given the variables x and y are coded using
  - $X = px + q$
  - $Y = my + n$

- The equation of the regression line of Y on X can be calculated as $Y = A + BX$
- To find the equation of the regression line of y on x
  - Substitute the codes into the equation
    - $my + n = A + B(px + q)$
  - Rearrange into the form $y = a + bx$

Your notes

Worked Example

Stewart collects data to compare the salaries, £s , and lengths of service, l years, of employees in a business. Stewart codes the data using the formulae

$$m = \frac{s - 30000}{12} \text{ and } w = 52l.$$

Stewart finds that:

- the product moment correlation coefficient between m and w is 0.739,
- the equation of the regression line of m on y is $m = -83.1 + 3.65w$.

(a) Write down the product moment correlation coefficient between s and l.

(b) The equation of the regression line of s on l can be written as $s = a + bl$. Find the values of a and b to three significant figures.

Answer:

a) Coding does not affect correlation

$r_{sl} = r_{mw}$

$r_{sl} = 0.739$

b) $m = -83.1 + 3.65w$

Substitute in the codes

$m = \frac{s - 30000}{12}$ and $w = 52L$

$\frac{s - 30000}{12} = -83.1 + 3.65(52L)$

Rearrange into required form

$S - 30000 = -997.2 + 2277.6L$

$S = \underbrace{29002.8}_{a} + \underbrace{2277.6L}_{b}$

$a = 29000$ (3.s.f.) $b = 2280$ (3.s.f.)

Your notes

Examiner Tips and Tricks

- When rearranging the equation of the regression it is important that you don't round your coefficients until the very end. Use your ANS button on your calculator to keep the accuracy.