9 Correlation & regression

Content

9.1 Correlation coefficient and linear regression

Learning Outcomes

Candidates should be able to:

9.1 Correlation coefficient and linear regression

(a) understand concepts of scatter diagram, correlation coefficient and linear regression;

(b) solve problems involving the calculation and interpretation of the product moment correlation coefficient and of the equation of the least squares regression line;

(c) understand concepts of interpolation and extrapolation;

(d) use a square, reciprocal or logarithmic transformation to achieve linearity.

– Exclude derivation of formulae
– Exclude hypothesis tests

9.1 Correlation coefficient and linear regression

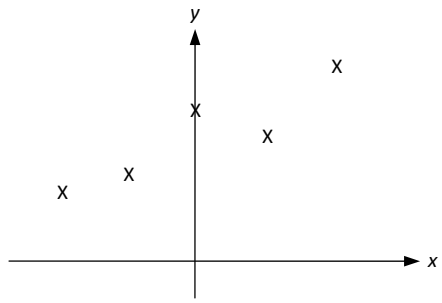
fundamental questions

Example 1

x	–2	–1	0	1	2
y	3	4	7	6	10

Given the following information, we can plot the data as points on a Cartesian plane to get a **scatter diagram**.

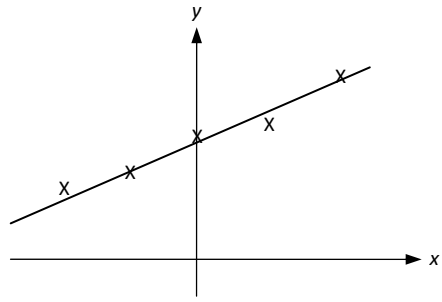
Scatter diagram of y against x



The scatter diagram gives an overview of how the variables are related and enables a **suitable correlation coefficient** to be calculated in measuring the **strength of association** among the variables.

Example 2

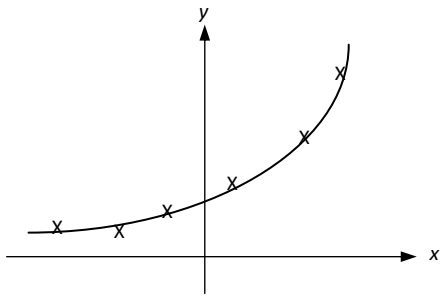
Scatter diagram of y against x



We can also draw a **linear regression line** to represent the set of bivariate data, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, on the scatter diagram.

Similarly, a **non-linear regression curve** can also be used to represent other sets of bivariate data on the scatter diagram.

Scatter diagram of *y* against *x*

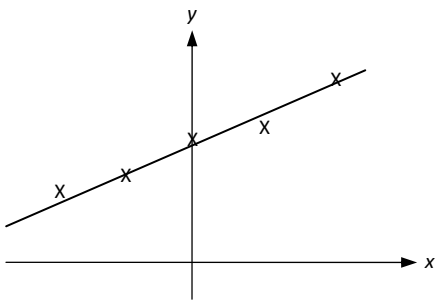


Example 3

Let *X* and *Y* denote two random variables. Both variables *X* and *Y* are deemed to be **correlated** if a **change** in one of them (*X* or *Y*) **leads** to a **change** in the other variable (*Y* or *X*). Otherwise, they are **uncorrelated**.

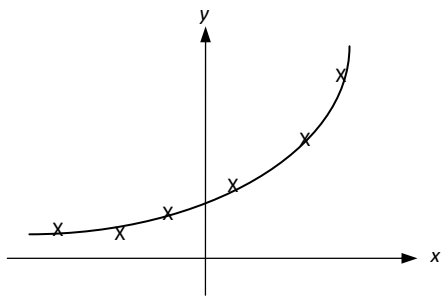
Positive (Direct) **Linear** Correlation

Scatter diagram of *y* against *x*



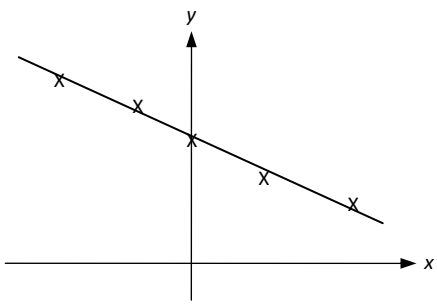
Positive (Direct) **Curvilinear** Correlation

Scatter diagram of *y* against *x*



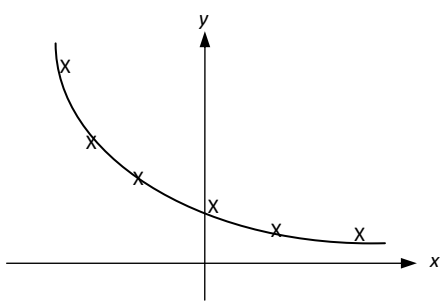
Negative (Inverse) **Linear** Correlation

Scatter diagram of *y* against *x*



Negative (Inverse) **Curvilinear** Correlation

Scatter diagram of *y* against *x*



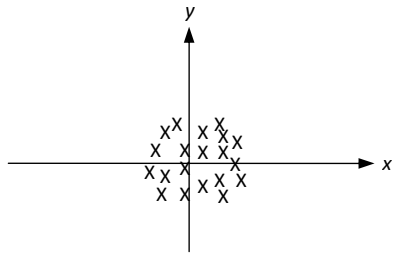
When *x* increases, *y* tends to increase $\Rightarrow$ **positive** correlation between *X* and *Y*

When *x* increases, *y* tends to decrease $\Rightarrow$ **negative** correlation between *X* and *Y*

No fixed relationship

$\Rightarrow$ **zero** correlation between X and Y

Scatter diagram of y against x



Example 4

Given a set of bivariate data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Finding a value of y from **within the range** of x (x_1 to x_n) and finding a value of x from **within the range** of y (y_1 to y_n) is called **interpolation**.

On the other hand, estimating a value from outside the data range is called **extrapolation**. Extrapolation gives an estimated value which is **not reliable** since we are not sure if the relationship between X and Y will hold for larger/smaller values that those recorded.

Refer to example 1:

Estimating the value of y when $x = 0.5 \Rightarrow$ interpolation

Estimating the value of y when $x = -3 \Rightarrow$ extrapolation

Example 5

For any set of bivariate data, there exist two possible linear regression lines.

- **Regression line of *y* on *x***

  This is used to estimate y given a value of x (which is the **independent** variable). We use this regression line when the values of x have been controlled or may be considered to be exact, while the values of y are subjected to experimental errors.

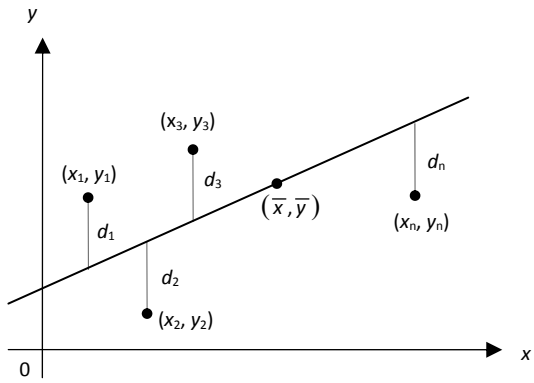
- **Regression line of *x* on *y***

  This is used to estimate x given a value of y (which is the independent variable). We use this regression line when the values of y have been controlled or may be considered to be exact, while the values of x are subjected to experimental errors.

Generally, the 2 regression lines do not coincide. However, the more correlated x and y are, the closer the 2 regression lines. Both lines, however, must pass through the point $(\bar{x}, \bar{y})$ where $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$.

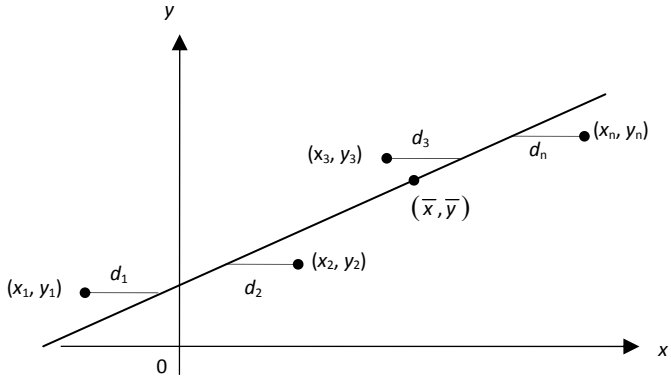
Example 6

To obtain the regression lines, we use the **least squares method**.



Fitting a least squares regression line of ***y* on *x*** means to fit a straight line for the set of data $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ such that the quantity $\sum_{i=1}^{n} d_i^2$ is a **minimum**. $\sum_{i=1}^{n} d_i^2$ is the sum of the **squares of the errors** between the **observed** *y*-values and the **theoretical** *y*-values.

Similarly, we can fit a least squares regression line of ***x* on *y***.



Fitting a least squares regression line of ***x* on *y*** means to fit a straight line for the set of data $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ such that the quantity $\sum_{i=1}^{n} d_i^2$ is a **minimum**. $\sum_{i=1}^{n} d_i^2$ is the sum of the **squares of the errors** between the **observed** *x*-values and the **theoretical** *x*-values.

Note: if the least squares regression line of *y* on *x* is given by $y = 2x - 1$, it does **not** mean $x = \frac{1}{2}y + \frac{1}{2}$ is the least squares regression line of *x* on *y*.

Example 7

We can get the least squares regression line of y on x (or x on y) from the **graphic calculator** if **unsupported** answers are allowed. Otherwise, we can make use of the following equations:

Estimated regression line of ***y* on *x***:

$$y-\overline{y}=b\left(x-\overline{x}\right)$$

where $$b=\frac{\sum\left(x-\overline{x}\right)\left(y-\overline{y}\right)}{\sum\left(x-\overline{x}\right)^2}=\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sum x^2-\frac{\left(\sum x\right)^2}{n}}$$

Estimated regression line of ***x* on *y***:

$$x-\overline{x}=b\left(y-\overline{y}\right)$$

where $$b=\frac{\sum\left(y-\overline{y}\right)\left(x-\overline{x}\right)}{\sum\left(y-\overline{y}\right)^2}=\frac{\sum xy-\frac{\sum y\sum x}{n}}{\sum y^2-\frac{\left(\sum y\right)^2}{n}}$$

Example 8

Given that the least squares regression line of ***y* on *x*** is given by $y = b_1x + c_1$. Gradient **= b_1** and y-intercept **= c_1**.

However, for least squares regression line of ***x* on *y*** given by $x = b_2y + c_2$, to obtain the gradient and intercept, we need to **transform** the above equation to

$y = \frac{1}{b_2}x - \frac{c_2}{b_2}$.

Gradient = $\frac{1}{b_2}$ and y-intercept = $-\frac{c_2}{b_2}$.

b_1 and b_2 are known as the **regression coefficients** of the respective regression lines.

Example 9

Given the following data taken from a certain country,

Year (x)	1948	1950	1952	1954	1956	1958	1960
Population (y) in millions	210	224	302	381	412	467	528

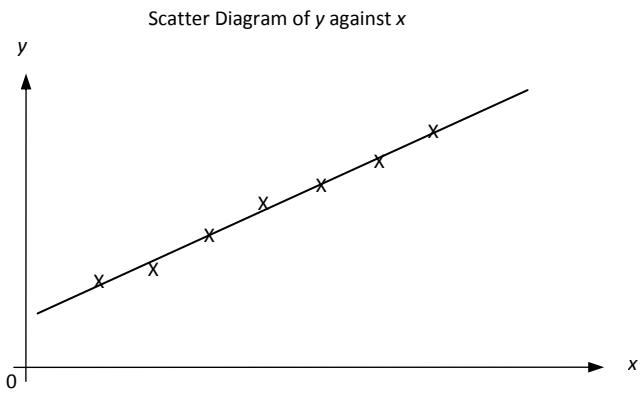
(a) Give a sketch of the scatter diagram. [1]

(b) Find the equation of the least squares regression line of y (population) on x (year), giving the values of the constants to 4 decimal places. Show this line on the scatter diagram. [2]

(c) Predict the population of the country in the year 2000 and comment on your answer. [2]

Solution:

(a)



(b) equation of the least squares regression line of y on x,

$y = 27.6786x - 53723.3571$ (ans)

(c) when $x = 2000$,

$y = 27.7(2000) - 53723$

$y = 1634$

population in the year 2000 is 1634 million. (ans)

Since $x = 2000$ is outside of data range, the calculated y value will not be reliable. (ans)

Mark Scheme:

(a)	correct **s**catter diagram	**B1**	
(b)	$y = 27.6786x - 53723.3571$	**B1**	
	correct regression line	**B1**	
(c)	$y = 1634$	**B1**	
	y value not reliable	**B1**	**[5]**

Example 10

ABC Company keeps records on its salespeople on the premise that sales (y, monthly sales in thousand dollars) should increase with experience (x, number of months on job). A random sample of eight salespeople (n = 8) produced the data on experience and sales as follows:

$\sum x = 109$ $\quad \sum x^2 = 1757$ $\quad \sum y = 115.9$

$\sum y^2 = 1862.77$ $\quad \sum xy = 1801.2$ $\quad n = 8$

By fitting the appropriate regression line, predict the monthly sales that a salesperson would be expected to generate after 10 months on the job. [5]

Solution:

As only summarized data is given, we are unable to use the graphic calculator.

Since we are given the value of x (=10) and asked to find y, the appropriate regression line would be y on x. Let the regression line of y on x be,

$$y - \bar{y} = b(x - \bar{x})$$

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

$$b = \frac{1801.2 - \frac{(109)(115.9)}{8}}{1757 - \frac{(109)^2}{8}} = 0.81678$$

$$\therefore y - \frac{115.9}{8} = 0.81678\left(x - \frac{109}{8}\right)$$

$$y = 0.817x + 3.36$$

When x = 10,

y = (0.817)(10) + 3.36 = 11.53

The salesperson would be expected to generate 11.5 thousand dollars. (ans)

Mark Scheme:

use y on x regression line	**M1**	
correct formulae for b	**M1**	
correct answer for b	**A1**	
y = 0.817x + 3.36	**A1**	
y = 11.53	**A1**	**[5]**

Example 11

Scatter diagrams give a more pictorial view of the linear relationship between 2 variables. However, a more objective method to determine linear relationship is to use a **correlation coefficient**. This correlation coefficient is a measure of the fit of a scatter diagram to a linear model. Its value (between –1 and 1) indicates the **strength** and **direction** of the **linear relationship**.

We denote the **estimated linear (product moment) correlation coefficient** with ***r***:

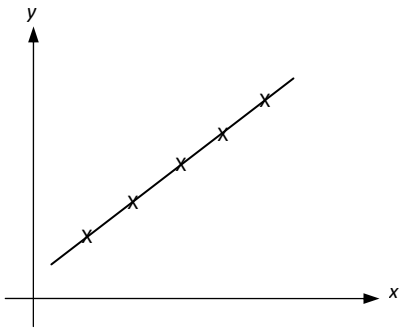
$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\left[\sum (x-\bar{x})^2\right]\left[\sum (y-\bar{y})^2\right]}} = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \dfrac{\left(\sum y\right)^2}{n}\right)}}$$

The **more correlated** the 2 variables are, the **closer** the 2 regression lines will be. Thus when *r* is close to 1 or –1, it is alright to use either the *y* on *x* or *x* on *y* regression lines to estimate the value of *y* or *x*.

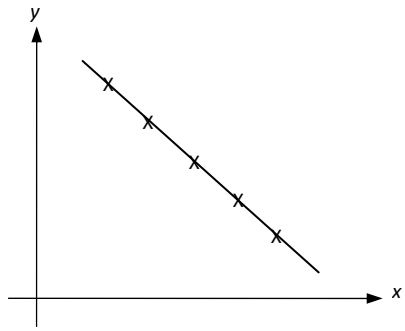
Example 12

- *r* is dimensionless and not affected by the units used in measurements.
- The value of *r* ranges from **–1 and 1**.
  Positive value of $r \Rightarrow Y$ tends to **increase** when *X* **increases**.
  Negative value of $r \Rightarrow Y$ tends to **decrease** when *X* **increases**.
- When $|r| = 1$, the two regression lines of *y* on *x* and *x* on *y* are **identical** and **coincide** with each other.

Scatter diagrams of *y* against *x*:



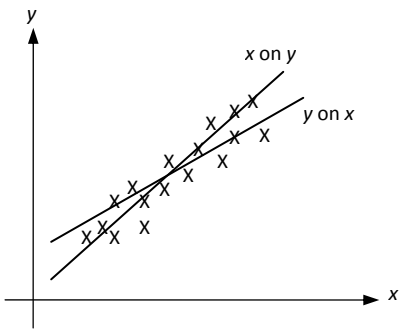
r = 1
perfect positive linear correlation



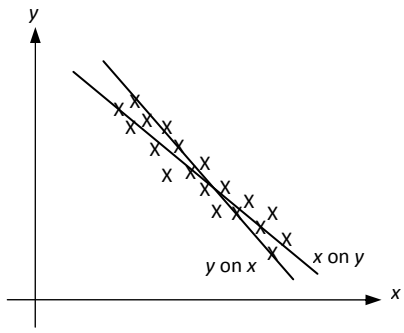
r = –1
perfect negative linear correlation

- When $|r| \approx 1$, the two regression lines of *y* on *x* and *x* on *y* are close **to each other.**

Scatter diagrams of *y* against *x*:



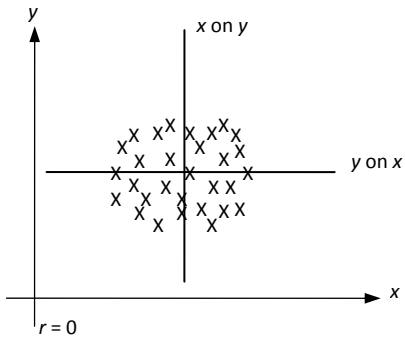
$r \approx 1$
strong positive linear correlation



$r \approx -1$
strong negative linear correlation

- When $r = 0$, the two regression lines of *y* on *x* and *x* on *y* are mutually perpendicular. Zero correlation does **not** necessarily imply that there is no relationship between *X* and *Y*, but rather these 2 variables **do not** share a **linear** relationship.

Scatter diagram of *y* against *x*:



$r = 0$
no linear correlation

Example 13

Consider the following regression lines:

y on *x*: $y = b_1x + a_1$

x on *y*: $x = b_2y + a_2$

then,

$r^2 = b_1b_2$

$\Rightarrow r = \sqrt{b_1b_2}$ if $b_1 > 0, b_2 > 0$

$\Rightarrow r = -\sqrt{b_1b_2}$ if $b_1 > 0, b_2 < 0$

Example 14

If X and Y are two random variables, and U and V are **linear functions** of X and Y where $U = aX + b$ and $V = cY + d$, (a, b, c and d are constants), the correlation coefficient between U and V is the **same** as that between X and Y, provided that $\boldsymbol{ac > 0}$.

Remember that the product moment correlation coefficient, r, measures the degree of scatter. Thus, it is **unchanged** by a **change** of **origin** and **scaling**.

Example 15

Let X represent the number of babies in country A and Y represent the number of houses in country B. If $r \approx 1$, we cannot say that an increase in the number of babies in country A causes an increase in the number of houses in country B.

A high correlation does **not** mean one factor directly causes the other.

Example 16

Let X be the number of cigarette smoked and Y be the number of road accidents over a number of years. If $r \approx 1$, can we say that smoking causes road accidents or road accidents cause smoking? Explain your answer. [2]

Solution:

No. (ans)

There might be a third factor (such as population) to which each might be related. Over the years analyzed, the population might have increased. This causes both smoking and road accidents to increase as well.

Increase in population → more smoking

Increase in population → more cars → more road accidents

As seen, both sets of data might be related to a 3rd factor, which is not stated in the question. (ans)

Always remember that a high correlation between 2 variables does not necessarily imply a **direct casual** relationship between them.

Mark Scheme:

no	**B1**	
correct explanation	**B1**	**[2]**

Example 17

When 2 random variables are **not** linearly correlated, we may use appropriate transformations to **linearize** a set of bivariate data to fit the linear regression model.

- **Square Transformation**

  Given $aV^2 = bU^2 + c$.

  Let $Y = V^2$ and $X = U^2$.

  The given equation is then transformed to:

  $Y = \frac{b}{a}X + \frac{c}{a}$ (straight line equation).

- **Logarithmic Transformation**

  Given $V = ab^U$

  $\Rightarrow \ln V = \ln a + U \ln b$

  Let $Y = \ln V$ and $X = U$.

  The given equation becomes:

  $Y = (\ln b)X + \ln a$

  Given $V = aU^b$

  $\Rightarrow \ln V = \ln a + b \ln U$

  Let $Y = \ln V$ and $X = \ln U$.

  The given equation becomes:

  $Y = bX + \ln a$

- **Reciprocal Transformation**

  Given $\frac{a}{U} + \frac{b}{V} = c$

  Let $Y = \frac{1}{V}$ and $X = -\frac{1}{U}$

  The given equation becomes:

  $Y = \frac{a}{b}X + \frac{c}{b}$

In all 3 kinds of transformation, the end product is the equation of a straight line.

Example 18

The number applicants for a course in Engineering in a particular university in each of six semesters is given below:

Semester (x)	1	2	3	4	5	6
No. of applicants (y)	11	29	68	138	215	560

The administrative office believes that the number of applicants (y) and the semester (x) are related by the equation $y = A(B)^x$, where A and B are constants.

(i) Using a suitable transformation involving $Y = \lg y$, give a sketch of the scatter diagram. Explain whether the scatter diagram provides evidence that the relation is a reasonable model. [2]

(ii) Find the equation of the estimated line of regression of Y on x, and the least squares estimate of A. [2]

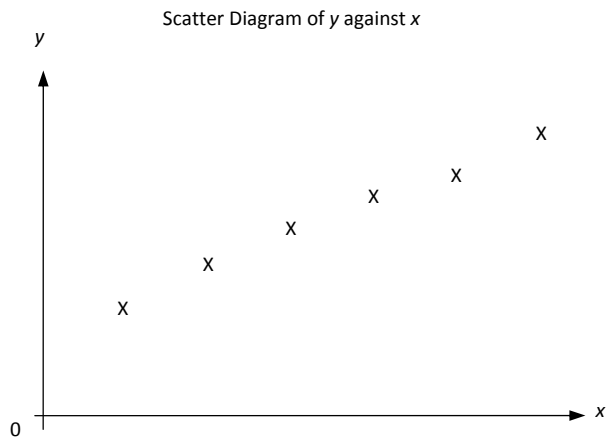
(iii) Explain whether the correlation coefficient supports the reasonability of the model. [2]

Solution:

(i) $y = A(B)^x$

$\lg y = \lg A + x \lg B$

Plot $\lg y$ against x



The points on the scatter diagram lie close to a straight line.

The relation is a reasonable model. (ans)

(ii) From graphic calculator, the regression line of Y on x is

$Y = 0.781 + 0.327x$ (ans)

Since $\lg y = \lg A + x \lg B$,

$\lg A = 0.781$

$A = 6.04$ (ans)

(iii) From graphic calculator,

$r = 0.995 \approx 1$

The correlation coefficient supports the reasonability of the model. (ans)

Mark Scheme:

(i)	scatter diagram	**B1**	
	reasonable model	**B1**	
(ii)	$Y = 0.781 + 0.327x$	**B1**	
	$A = 6.04$	**B1**	
(iii)	r value close to 1	**B1**	
	Supports reasonability of model	**B1**	**[6]**

1. Given $v = \dfrac{au}{b+u}$, show that the equation can be written in a form that is liner in $\dfrac{1}{u}$ and $\dfrac{1}{v}$. [1]

(i) Find the equation of the estimated regression line of $\dfrac{1}{u}$ on $\dfrac{1}{v}$ given the following data. Hence, find the estimates of *a* and *b*, correct to 3 decimal places. [3]

u	0.1	0.2	0.4	0.6	0.8	1.0
v	0.311	0.453	0.681	0.831	0.923	0.986

$\sum \dfrac{1}{u} = 21.417 \quad \sum \left(\dfrac{1}{u}\right)^2 = 136.591 \quad \sum \dfrac{1}{v} = 10.191 \quad \sum \left(\dfrac{1}{v}\right)^2 = 21.015$

$\sum \left(\dfrac{1}{u}\right)\left(\dfrac{1}{v}\right) = 51.233$

(ii) Find the estimated value of *v* when *u* = 2.4. Comment on the reliability of this answer. [2]

2. Research is carried out into how the volume of a liquid in a container changes with time when left at room temperature. Observations at successive times give the data shown in the following table.

Time (*t* days)	0	1	2	3	4	5	6	10	20
Volume (*v* ml)	100	83	64	33	31	21	13	6	2

(i) Show these data on a scatter diagram. [1]

(ii) Calculate the equation of the regression line of *v* on *t*. [1]

(iii) Using the regression line from part (ii), it is estimated that when *t* = 30, *v* = -65ml. This is a poor estimate since volume cannot be negative. Suggest two reasons that resulted in this poor estimation. [2]

The variable *y* is defined by $y = \dfrac{1}{v}$. For the variables *y* and *t*, calculate

(iv) the product moment correlation coefficient and comment on its value. [2]

(v) the equation of a suitable least square regression line, giving a reason for your choice of line.

Use this regression line to estimate the time when the volume is 50ml. Comment on the reliability of your answer. [4]

3. The table shows the profits made by firms in an industry and the corresponding turnovers (sales values).

Turnover \$$x$ (in thousands)	46	53	37	42	34	29	60	44	41	48
Profit \$$y$ (in thousands)	12	14	11	13	10	8	17	12	10	15

(i) Give a sketch of the scatter diagram for the data. [1]

(ii) Find the equation of the least squares regression line of y on x and sketch the line on the scatter diagram. [2]

(iii) Calculate the product moment correlation coefficient. [1]

(iv) Find the profit that a firm is expected to make if it has a turnover of

(a) \$58000, (b) \$10000. [2]

(v) Comment briefly on the reliability of the estimates in (iv). [2]

4. The regression line of y on x is given by $7y = 25 + bx$. The corresponding regression line of x on y is given by $x = 11 + b^*y$. Given that $b^* < b < 0$ and that b^* and b are the roots of the equation $7x^2 + 23x + 6 = 0$, find in any order,

(i) $\left(\bar{x},\bar{y}\right)$,

(ii) r, where r is the product moment correlation coefficient between x and y. [4]

The heartbeat count (x) and the diastolic blood pressure (y), both in similar units, were measured for 10 patients with suspected high blood pressure. The results were as follows:

x	50	52	56	57	63	64	68	70	76	77
y	93	91	87	90	81	82	77	76	70	71

(a) Calculate the least square regression line of y on x. [1]

(b) Calculate the value of the linear (product moment) correlation coefficient for the above data. What can you say about the regression line of y on x and x on y? [2]

(c) Estimate

(i) the diastolic blood pressure for a new patient with heartbeat count of 60

(ii) the heartbeat count of a patient with diastolic blood pressure of 94. [2]

Comment on the reliability of your answers. [1]

(d) Based on the result in (b), a doctor concluded that if a patient's heartbeat count is low, then his diastolic blood pressure must be high. Comment on his conclusion. [1]

5. The table below shows the daily sale of cones of ice-cream in a week by a shop and the maximum daily temperature.

	Mon	Tues	Wed	Thurs	Fri	Sat	Sun
Daily sales, u	94	102	112	53	80	83	88
Temperature °C, t	30.7	31.4	31.8	34.6	25.9	28.5	29.1

(i) Identify a data pair which should be regarded as suspect.
Remove the suspect data pair for the rest of the question. [1]

(ii) Calculate the correlate coefficient for the remaining 6 pairs of data. [1]

(iii) The variable v is defined by $v = \frac{1}{u}$. For the variable v and t, calculate the product moment correlation coefficient and comment on its value. [3]

(iv) Use a regression line to give the best estimate that you can of the daily sales when the temperature is 21.0°C.
State, with a reason, whether the estimation is valid. [3]

6. In a chemical reaction, let x be the weight (in grams) of compound X used and y be the volume (in cm^3) of oxygen released. The following data was obtained in a series of 8 experiments.

$$\sum x = 365.4 \qquad \sum x^2 = 20994 \qquad \sum y = 1557.3 \qquad \sum y^2 = 379997.81$$

$$\sum xy = 89024.47$$

(i) Using only the data above, calculate the equation of the regression line of x on y, in the form of $x = a + by$. [4]

The actual raw data of the experiments are given below:

x	10.3	20.2	29.7	c	52.4	59.5	70.8	81.2
y	30.3	101.4	140.9	185.9	211.7	226.2	328.4	332.5

(ii) Find the value of c. [2]

(iii) It is subsequently found that the x-value of the set of data (81.2, 332.5) was wrong and it should be more than 81.2. Comment on the change on the value of b in (i). [1]

7. Each of a random sample of 10 students are asked about the average number of minutes spent on doing mathematics tutorials in a week (x), and their percentage score for the mathematics final examination (y). The results are tabulated below:

x	20	35	45	60	70	80	100	110	120	140
y	16	25	35	50	60	65	70	75	80	85

(i) Find the equation of the regression line of y on x. [1]

(ii) Find the linear product moment correlation coefficient between y and x, and comment on the relationship between x and y. [2]

(iii) Making use of your answer from part (i), find the equation of the regression line of x on y. [2]

(iv) Use the appropriate regression line to estimate the percentage score of a student who spends 10 minutes doing mathematics tutorial in a week. Comment on the reliability of the estimate. [2]

8. The height, x mm, and the weight, y kg, of each boy in a school were recorded and a scatter diagram was then plotted.

State, with reasons, whether the regression line of Y on X or regression line of X on Y is likely to be more useful. [1]

The equation of the line of regression of Y on X is found to be $Y = 0.09X - 90$. The sample heights are found distributed about a mean of 1600mm with standard deviation 120mm and the standard deviation of the weights is 12kg.

(i) Show that the sample mean of the weights is 54. [1]

(ii) Find the equation of the line of regression of X on Y. [6]

(iii) Find the linear (product moment) correlation coefficient between X and Y. Comment on what this value implies about the regression line. [2]

(iv) Find the expected weight of a boy whose height is 1.5m. [1]

1. Solution:

$$v = \frac{au}{b+u}$$

$$\Rightarrow \frac{1}{v} = \frac{b+u}{au}$$

$$\Rightarrow \frac{1}{v} = \left(\frac{b}{a}\right)\left(\frac{1}{u}\right) + \frac{1}{a}$$

$$\Rightarrow \frac{1}{u} = \left(\frac{a}{b}\right)\left(\frac{1}{v}\right) - \frac{1}{b} \quad \text{(ans)}$$

(i) Using Graphic Calculator, equation of the estimated regression line of $\frac{1}{u}$ on $\frac{1}{v}$ is given as $\frac{1}{u} = 4.00987\frac{1}{v} - 3.24223 \Rightarrow \frac{1}{u} = 4.01\frac{1}{v} - 3.24$.

$$\therefore \frac{a}{b} = 4.00987$$

$$-\frac{1}{b} = -3.24223$$

Solving, we get:

a = 1.237 (ans)

b = 0.308 (ans)

(ii) Using $\frac{1}{u} = 4.00987\frac{1}{v} - 3.24223$, when u = 2.4, v = 1.096 (ans)

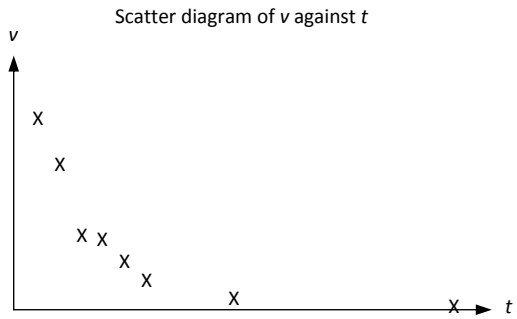
Since u = 2.4 is not within the data range in question, the estimated value of v is not very reliable. (ans)

Mark Scheme:

	$\frac{1}{u} = \left(\frac{a}{b}\right)\left(\frac{1}{v}\right) - \frac{1}{b}$	**B1**	
(i)	$\frac{1}{u} = 4.01\frac{1}{v} - 3.24$	**B1**	
	a = 1.237	**B1**	
	b = 0.308	**B1**	
(iii)	v = 1.096	**B1**	
	outside data range, not reliable	**B1**	**[6]**

2. Solution:

(i)



(ii) From Graphic Calculator,

$v = -4.25t + 63.3$ (ans)

(iii) 2 reasons resulting in poor estimation:

1) From the scatter diagram in part (i), the data do not appear to follow a linear relation. Thus it may not be accurate to estimate results using the regression equation obtained.

2) $t = 30$ is outside of the data range. We are not sure if the relation determined within the data range is still applicable outside the data range. In this case, this relation may not hold outside of data range. (ans)

(iv) $r = 0.975$ (ans)

The product moment correlation coefficient is ≈ 1, indicating a strong positive linear relation between y and t (within data range). (ans)

(v) Equation of suitable least square regression line:

$y = 0.0251t - 0.0431$ (ans)

This line is chosen because from the context of the question, it is implied that v is measured with varying t, meaning v is likely the observed (dependent) variable while t is control (independent variable).

When $v = 50$, $y = 0.02$,

$t = 2.52$ days (ans)

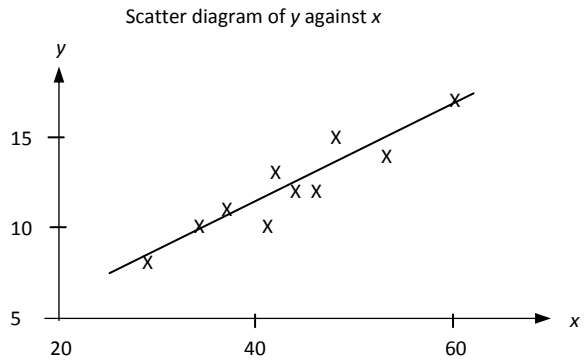
The answer should be quite accurate as the r value is high and $v = 50$ is within the data range.

Mark Scheme:

(i)	correct scatter diagram	**B1**	
(ii)	$v = -4.25t + 63.3$	**B1**	
(iii)	reason 1: no linear relation	**B1**	
	reason 2: outside data range	**B1**	
(iv)	$r = 0.975$	**B1**	
	$r \approx 1$, strong positive linear relation	**B1**	
(v)	$y = 0.0251t - 0.0431$	**B1**	
	v is dependent, t is independent	**B1**	
	$t = 2.52$	**B1**	
	accurate, since $r \approx 1$	**B1**	**[10]**

3. Solution:

(i)



(ii) $y = 0.273x + 0.348$ (ans)

(iii) $r = 0.932$ (ans)

(iv)(a) When $x = 58$, $y = 16.182$

$\therefore$ profit that firm is expected to make = \$16182 (ans)

(iv)(b) When $x = 10$, $y = 3.078$

$\therefore$ profit that firm is expected to make = \$3078 (ans)

For part (iv)(a), $x = 58$ is within the data range. It is an interpolation, hence answer can be accurate.

For part (iv)(b), $x = 10$ is outside of the data range, It is an extrapolation, hence result may not be reliable. (ans)

Mark Scheme:

(i)	correct scatter diagram	**B1**	
(ii)	$y = 0.273x + 0.348$	**B1**	
	regression line	**B1**	
(iii)	$r = 0.932$	**B1**	
(iv)	\$16182	**B1**	
	\$3078	**B1**	
(v)	interpolation, accurate	**B1**	
	Extrapolation, inaccurate	**B1**	**[8]**

4. **Solution:**

Given that b^* and b are the roots of the quadratic equation $7x^2 + 23x + 6 = 0$.

$7x^2 + 23x + 6 = 0$

$\Rightarrow \left(x + \frac{2}{7}\right)(x + 3) = 0$

$x = -\frac{2}{7}$ or $x = -3$

Since $b^* < b < 0$,

$b^* = -\frac{2}{7}$

$b = -3$

(i) The regression line equations become

$7y = 25 - \frac{2}{7}x$ and $x = 11 - 3y$

Solving for the intersection of the 2 lines:

$(\bar{x}, \bar{y}) = \left(\frac{14}{43}, \frac{153}{43}\right)$ (ans)

(ii) $r = -\sqrt{(b^*)\left(\frac{b}{7}\right)} = -\frac{\sqrt{6}}{7}$ (ans)

(a) Using Graphic Calculator, regression line of y on x is
$y = -0.862x + 136.337$ (ans)

(b) From Graphic Calculator, $r = -0.989$.

Since r is close to -1, we can deduce that the two regression lines of y on x and x on y are very close to each other.

(c)(i) Using the regression line y on x, $y = -0.862x + 136.337$,

When $x = 60$, $y = 84.6$

The diastolic blood pressure for a new patient with heartbeat count of 60 is 84.6. (ans)

This estimate is reliable as $x = 60$ is within the data range and the value of r is close to -1.

(c)(ii) As the question did not specify clearly which is the dependent and the independent variable, we compute x on y in the graphic calculator.
$\therefore$ Regression line of x on y:

$x = -1.1357y + 156.19916$

When $y = 94$, $x = 49.4$

The heartbeat count of the patient is 49.4. (ans)

This estimate may not be as reliable as $y = 94$ is slightly out of data range.

(d) The sample size is too small and there will be a danger of over-generalizing the trend. Hence, the conclusion may or may not be reliable. (ans)

Mark Scheme:

(i)	solve quadratic $\left(x+\frac{2}{7}\right)(x+3)=0$	**M1**	
	solve regression line equations	**M1**	
	$(\bar{x},\bar{y})=\left(\frac{14}{43},\frac{153}{43}\right)$	**A1**	
(ii)	r = $\frac{-\sqrt{6}}{7}$	**A1**	
(a)	y = –0.862x + 136.337	**B1**	
(b)	r = –0.989	**B1**	
	Two regression lines are close to each other	**B1**	
(c)	y = 84.6	**B1**	
	x = 49.4	**B1**	
	(i) reliable, (ii) unreliable	**B1**	
(d)	small sample size, over generalizing	**B1**	**[11]**

5. Solution:

(i) The pair of (34.6, 53) should be regarded as suspect. (ans)

(ii) For remaining 6 pairs of data,
r = 0.89981 ≈ 0.900 (ans)

(iii) For variables v and t,
r = –0.93249 ≈ –0.932 (ans)
Since the magnitude of the product moment correlation coefficient (for variables v and t) is larger than the value in part (ii), the linear relationship between $\frac{1}{u}$ and t is stronger than that between u and t.

(iv) Regression line of v against t:

$$v = 0.027842 - 3.7367\times10^{-4}t$$

$$\Rightarrow \frac{1}{u} = 0.027842 - 3.7367\times10^{-4}t$$

When t = 21, u = 63.3
Daily sale when the temperature is 21.0°C is 63.3. (ans)
Since t = 21.0 is not within the data range, the estimate for u may not be reliable.

Mark Scheme:

(i)	(34,6, 53)	**B1**	
(ii)	r = 0.900	**B1**	
(iii)	calculate values of v for corresponding u	**M1**	
	r = –0.932	**A1**	
	stronger relationship between v and t	**B1**	
(iv)	correct regression line $v = 0.027842 - 3.7367\times10^{-4}t$	**M1**	
	u = 63.3	**A1**	
	t = 21.0 not within data range	**B1**	**[8]**

6. **Solution:**

(i) $$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}$$

$$= \frac{89024.47 - \frac{(365.4)(1557.3)}{8}}{379997.81 - \frac{(1557.3)^2}{8}}$$

$$= 0.23285$$

$$\bar{x} = \frac{365.4}{8} = 45.675$$

$$\bar{y} = \frac{1557.3}{8} = 194.66$$

The regression line of x on y is

$$x - \bar{x} = b(y - \bar{y})$$

$$x - 45.675 = 0.23285(y - 194.66)$$

$$x = 0.347 + 0.233y \quad \text{(ans)}$$

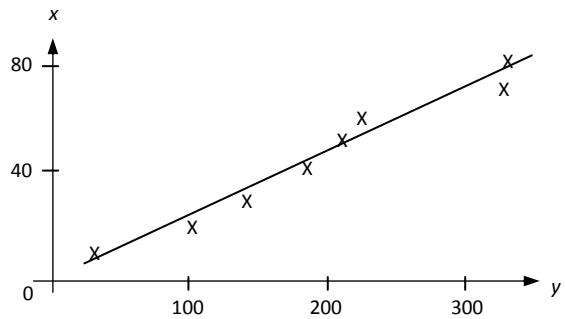
(ii) $$\sum x = 10.3 + 20.2 + 29.7 + c + 52.4 + 59.5 + 70.8 + 81.2$$

$$c = 365.4 - 324.1$$

$$= 41.3 \quad \text{(ans)}$$

(iii)

Scatter diagram of x against y



If $x > 81.2$, the last point will be placed vertically upwards which will result in steeper gradient of new regression line. Value of b will increase. (ans)

Mark Scheme:

(i)	using formulae for b	**M1**	
	finding $\overline{x}$ and $\overline{y}$	**M1**	
	correct values of $\overline{x}$ and $\overline{y}$	**A1**	
	$x = 0.347 + 0.233y$	**A1**	
(ii)	summing x	**M1**	
	$c = 41.3$	**A1**	
(iii)	value of b increases	**B1**	**[7]**

---

7. Solution:

(i) Equation of regression line of y on x:
$y = 10.0 + 0.591x$ (ans)

(ii) $r = 0.972$ (ans)

Since r is very close to 1, there is a strong positive linear correlation between x and y.

(iii) Let $y = a + bx$ be the equation of the regression line of y on x.

Let $x = c + dy$ be the equation of the regression line of x on y.

$\therefore$ from part (i), $b = 0.591$.

Since $r^2 = bd$,

$0.94422 = 0.59102d$

$d = 1.5976 \approx 1.60$

the equation of the regression line of x on y can also be written as $x - \overline{x} = d\left(y - \overline{y}\right)$.

Hence equation of the regression line of x on y is

$x - 78 = 1.5976(y - 56.1)$

$x = 1.60y - 11.6$ (ans)

(iv) when $x = 10$, $y = 15.9$

The percentage score of a student who spends 10 minutes doing mathematics tutorial in a week is 15.9.

This estimate may not be accurate as $x = 10$ is outside of the data range and extrapolation is needed to estimate the percentage score.

Mark Scheme:

(i)	$y = 10.0 + 0.591x$	**B1**	
(ii)	$r = 0.972$	**B1**	
	strong positive linear correlation	**B1**	
(iii)	use $r^2 = bd$	**M1**	
	$x = 1.60y - 11.6$	**A1**	
(iv)	$y = 15.9$	**B1**	
	$x = 10$ outside data range	**B1**	**[7]**

8. **Solution:**

Regression line of Y on X is more useful as we usually want to predict the weight according to the height given. (ans)

(i) X: $\bar{x} = 1600$, sample variance $= 120^2$

Y: sample variance $= 12^2$

$(\bar{x}, \bar{y})$ falls on regression line.

$\bar{y} = 0.09\bar{x} - 90$

$= 0.09(1600) - 90$

$= 54$ (shown) (ans)

(ii) $b = 0.09 = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ --- (1)

To find $d = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$ --- (2)

$\dfrac{\sum(x-\bar{x})^2}{n} = 120^2$ --- (3)

$\dfrac{\sum(y-\bar{y})^2}{n} = 12^2$ --- (4)

Sub (3) into (1):

$\sum(x-\bar{x})(y-\bar{y}) = 0.09\left(120^2\right)n$ --- (5)

Sub (4) and (5) into (2):

$d = \dfrac{0.09\left(120^2\right)n}{12^2 n}$

$d = 9$

$x - \bar{x} = 9(y-\bar{y})$

$x - 1600 = 9(y-\bar{y})$

$x = 1114 + 9y$ (ans)

(iii) $r^2 = bd = 0.81$

$r = 0.9$ (ans)

(r is positive since both d and b are positive)

It implies a strong positive correlation between X and Y which suggests the 2 regression lines are quite identical. (ans)

(iv) Using the y on x regression line,

$y = 0.09(x) - 90$

When x = 1500, y = 45.

The expected weight of a boy whose height is 1.5m is 45kg. (ans)

Mark Scheme:

	predict height according to weight	**B1**	
(i)	show sample mean = 54	**B1**	
(ii)	$d = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$	**B1**	
	$\dfrac{\sum(x-\bar{x})^2}{n} = 120^2$	**B1**	
	$\dfrac{\sum(y-\bar{y})^2}{n} = 12^2$	**B1**	
	solving simultaneous equations	**M1**	
	d = 9	**A1**	
	$x = 9y + 1114$	**A1**	
(iii)	r = 0.9	**B1**	
	regression lines are quite identical	**B1**	
(iv)	y = 45	**B1**	**[11]**