

Chapter 7 (Statistics)**Correlation and Linear Regression****Objectives**

At the end of the chapter, students should be able to:

- (a) Understand that bivariate data consists of the values of two variables (independent and dependent variables) obtained from the same sample expressed in ordered pairs;
- (b) Use the graphic calculator to plot the scatter diagram for a set of bivariate data to determine if there is a linear relationship between the two variables;
- (c) Understand that the correlation coefficient is a measure of the fit of a scatter diagram to a linear model;
- (d) Calculate the product moment correlation coefficient for a set of bivariate data using a graphic calculator, and relate the value (in particular, values close to -1, 0 and 1) to the appearance of the scatter diagram; [Note: zero correlation does not necessarily imply ‘no relationship’, but rather ‘no linear relationship’.]
- (e) Understand that a high correlation between two variables does not necessarily imply one directly causes the other;
- (f) Understand the concepts of linear regression and ‘least squares’ with reference to the scatter diagram;
- (g) Use a graphic calculator to find the equation of the least squares regression line, and interpret its slope and intercept; [Note: A different line will be obtained if we interchange the independent and dependent variables.]
- (h) Understand the concepts of extrapolation and interpolation of data, and use the appropriate regression line to make prediction or estimate a value in practical situations.

Contents

- 7.1 Introduction
 - 7.1.1 Bivariate Data
 - 7.1.2 Independent vs Dependent Variables
- 7.2. Scatter Diagrams
 - 7.2.1 Drawing Scatter Diagrams
 - 7.2.2 Interpreting Scatter Diagrams
- 7.3 Product Moment Correlation Coefficient, r
 - 7.3.1 What is Product Moment Correlation Coefficient, r ?
 - 7.3.2 Calculation of Product Moment Correlation Coefficient, r using Formula
 - 7.3.3 Calculation of Product Moment Correlation Coefficient, r using Graphic Calculator
 - 7.3.4 Properties of Product Moment Correlation Coefficient, r
 - 7.3.5 Limitations of Product Moment Correlation Coefficient, r
 - 7.3.6 The Effects of Transformation on Product Moment Correlation Coefficient, r
- 7.4 Linear Regression
 - 7.4.1 Types of Regression Lines
 - 7.4.2 Using Graphic Calculator to find Equations of Regression Lines
 - 7.4.3 Least Squares Regression Lines of y on x
 - 7.4.4 Least Squares Regression Lines of x on y
 - 7.4.5 Properties of Regression Lines
- 7.5 Interpolation and Extrapolation
- 7.6 Self-Reading Examples

Resources

1. A Concise Course in A Level Statistics with Worked Examples
By J. Crawshaw and J. Chambers
2. A Comprehension Guide: H2 Mathematics for “A” Level, Volume 2
By Frederick Ho, David Khor, Yui-P’ng Lam and B.S. Ong
3. Worked Examples in Statistical Inference
By R.N.D. Publications

7.1 Introduction

Regression analysis and correlation analysis have been developed to study and measure the statistical relationship that exists between two or more variables. In the A Level syllabus, we will investigate the linear regression analysis and linear correlation analysis of two variables.

In **linear correlation analysis**, we measure the strength or closeness of the linear relationship between the two variables.

In **linear regression analysis**, we prepare an estimating (or regression) linear equation to estimate the values of one variable from given values of another.

In other words, correlation analysis reveals the extent to which two variables are related whilst regression analysis tells us how they are related.

EXAMPLE

We are interested in the relationship between the number of hours studying and the marks obtained. The table below shows the amount of time (x hours) that 6 average students spent studying for a test and the marks that they obtained (y out of 100 marks).

x	1	2.5	3	4	4.2	5
y	20	46	48	60	61	65

Correlation and regression analyses will help us to answer the following questions:

- Is there a relationship between the amount of time spent on revision and marks obtained? – Correlation
- Is the relationship a linear one? – Linear Correlation
- If there is a linear relationship, can we reasonably predict the value of one of the variables from the knowledge of the other? – By finding the regression line
- If another student studies for 2 hours, can you predict how many marks he will obtain? What about if he studies for 7 hours? – By using the regression line
- How certain are you of your prediction? – Interpolation, Extrapolation

7.1.1 Bivariate Data

Bivariate data refers to data connecting two variables. Each observation thus comprises a pair of values taken by the two variables. It is customary to express the bivariate data as **ordered pairs** (x, y).

An example of bivariate data is the data given in the table above, which shows a set of 6 pairs of values for the two variables, x and y .

7.1.2 Independent vs Dependent Variables

In the physical sciences, we often set up investigations or experiments in which we try to find a relationship between two variables.

For such experiments, it is fairly common for the researcher to assign arbitrary values to one variable so that corresponding values of the other variable can be measured. For this reason, the variable that the researcher has control over is known as the independent variable (or controlled variable), while the other variable under investigation (whose values we want to predict) is called the dependent variable.

On the other hand, there are situations (called observational studies) where the values of the independent variable cannot be pre-selected. An example of this is the study on the effect of time taken to prepare for a test (independent variable) on result of the test (dependent variable).

An **independent variable** (or input/controlled/predetermined variable) is usually the one whose value can be set or controlled. It is usually denoted by x .

A **dependent variable** (or output/response variable) is usually the one whose value we want to predict. It is usually denoted by y .

However, it is sometimes not possible to decide what the independent and dependent variables are, given a set of data. For example, History and Geography scores of a student.

Example 1 Identifying independent and dependent variables

In each of the cases below, identify the independent and dependent variables.

- (a) The drying time of a particular type of hobby paint depends on various factors including ambient temperature. The following table shows eight pairs of data obtained in an experiment on the ambient temperature in degree Celsius ($^{\circ}\text{C}$) and the corresponding drying time (in hours) of the paint.

Ambient temperature, x ($^{\circ}\text{C}$)	32.6	6.6	23.5	12.7	35.1	20.3	27.8	17.4
Drying time, y (hours)	5.9	24.9	3.8	17.7	3.4	12.5	8.6	16.2

Independent variable : Ambient temperature

Dependent variable : Drying time

- (b) The table shows a Verbal Reasoning test score, x , and an English test score, y , for each of a random sample of 10 students who took both tests.

Verbal Reasoning test score (x)	112	106	110	115	109	119	102	100	116	95
English test score (y)	69	63	75	79	72	85	90	58	75	60

Either x or y can be considered as the independent variables

7.2 Scatter Diagrams

7.2.1 Drawing Scatter Diagrams

A graph of the set of observations of the ordered pairs (x, y) on the x - y plane is known as a **scatter diagram** (or scatter plot).

By convention, we plot the independent variables along the horizontal axis and the dependent variables on the vertical axis. Note that scatter diagrams are used only for **quantitative** variables (those that are numerically comparable).

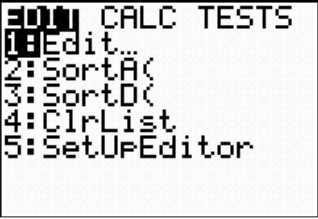
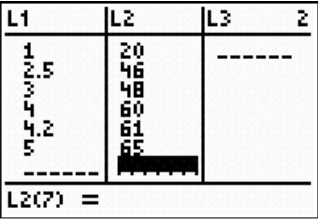
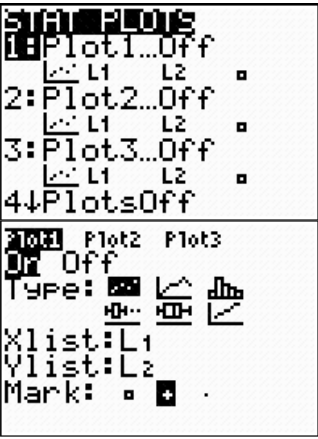
Your graphic calculator can be used to draw scatter diagrams.

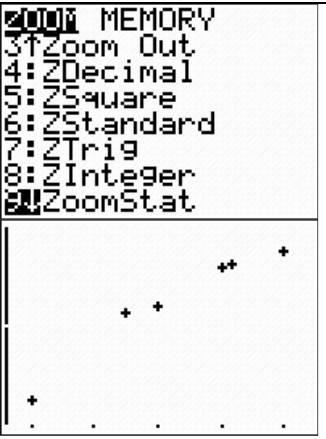
Example 2

The table shows the number of hours (x hours) that 6 average students spent studying for a test and the marks obtained (y out of 100 marks). Sketch the scatter diagram for the data.

x	1	2.5	3	4	4.2	5
y	20	46	48	60	61	65

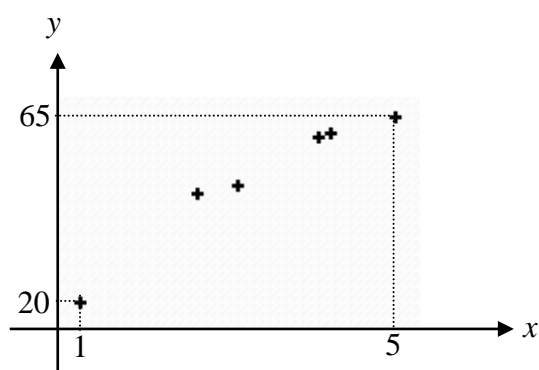
To sketch the scatter diagram, we may use the following GC keystrokes:

Steps	Screenshot	Remarks
Select [STAT], followed by [EDIT]. Select [1]: [Edit] and press [ENTER].		
Enter the values of x in list L_1 and values of y in L_2 .		
Press [STATPLOT]. Select [1]: [Plot1]. Set Plot to On.		<p>Type: Choose the first type of graph.</p> <p>XList: Refers to the list of variables on the horizontal axis, in this case it is L_1.</p> <p>YList: Refers to the list of variables on the vertical axis, in this case it is L_2.</p> <p>Mark: Choose any of the three styles.</p> <p>L_1: Press [2ND] [1].</p> <p>L_2: Press [2ND] [2].</p>

Press [Zoom] and select [9] to view the scatter diagram with statistical data points displayed.	 <p>2000 MEMORY 3:Zoom Out 4:ZDecimal 5:ZSquare 6:ZStandard 7:ZTrig 8:ZInteger 9:ZoomStat</p> <p>The scatter plot shows data points with a positive correlation. The x-axis ranges from 0 to 5, and the y-axis ranges from 0 to 65.</p>	<p>Note:</p> <p>To read the data points on the screen, press [TRACE]. Use the left or right arrow keys to move from point to point.</p>
---	--	---

When drawing a scatter diagram, ensure that the **axes are labeled** and the **relative position** of the points are correct. The **max and min x and y values observed** must be **labeled** as well.

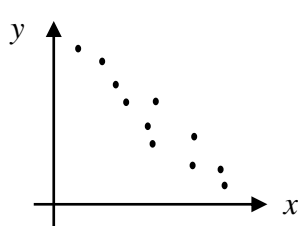
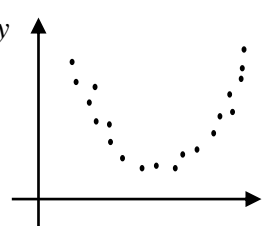
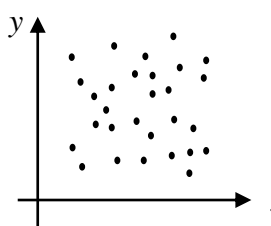
Solution:



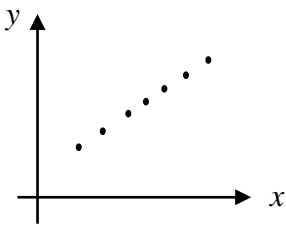
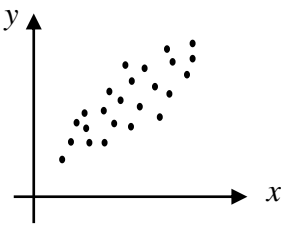
7.2.2 Interpreting Scatter Diagrams

Plotting a scatter diagram gives us an overview of how the variables are related, and hence indicates the strength of correlation between the variables. To interpret scatter diagrams, we can comment on 3 things:

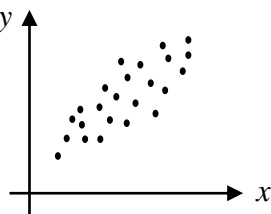
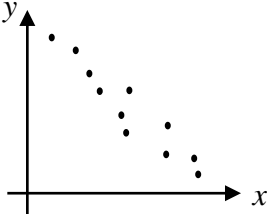
1. Type of relationship: The relationship might be linear, curved or no clear relationship.

		
<p>If all the points in a scatter diagram seem to lie near a straight line, we say there is linear correlation between the variables.</p>	<p>If most points lie close to some curve, the correlation is curvilinear.</p>	<p>If all the points are randomly scattered, with no discernible pattern, we say that the variables are uncorrelated or there is no correlation.</p>

2. Strength

	
<p>If the data points are close to the line of best fit, there is a strong correlation between x and y.</p>	<p>If the data are not clustered near the line of best fit, then the correlation between x and y may not be strong.</p>

3. Direction: The two variables x and y can be positively correlated, negatively correlated or not correlated at all.

	
<p>If y generally increases as x increases, then x and y are positively correlated.</p>	<p>If y generally decreases as x increases, then x and y are negatively correlated.</p>

Note:

A scatter diagram can also give us visual evidence of **outliers** (i.e. points that are significantly different from the rest of data) or suspicious observations. Removing such points gives a truer picture of the relationship between the variables.

Exercise 1

- With the aid of a graphic calculator, sketch a scatter diagram for each of the following sets of data.

(a)

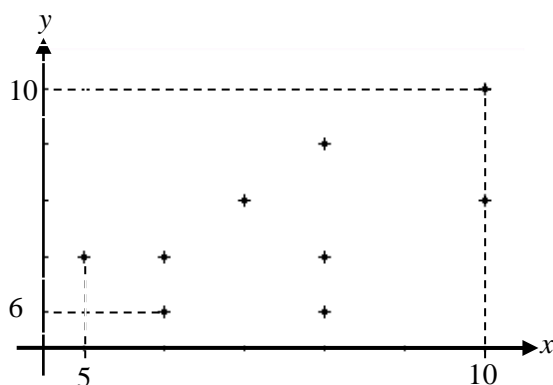
x	8	6	10	8	5	6	8	10	7
y	6	7	8	7	7	6	9	10	8

(b)

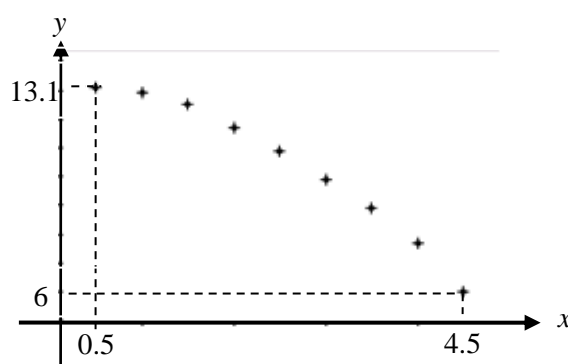
x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
y	13.1	12.9	12.5	11.7	10.9	9.9	8.9	7.7	6.0

Solution:

1(a)



1(b)



2. For each of the scatter diagrams in Question 1, describe the correlation between the values of x and y .

Solution:

- (a) Moderately strong positive linear correlation.
 (b) Negative curvilinear correlation.

7.3 Product Moment Correlation Coefficient, r

Interpreting the strength of correlation between two variables based solely on the scatter diagram is subjective. It can even be deceiving when different scales are used for the axes. To avoid such situations, we will use a **numerical analysis of the data** to determine the type of relation that exist between two variables.

7.3.1 What is Product Moment Correlation Coefficient, r ?

The product-moment correlation coefficient, denoted by r , where $-1 \leq r \leq 1$, is a numerical value which indicates the **strength** and **direction** of a linear relationship between two variables x and y .

The product moment correlation coefficient is also known as sample correlation coefficient, linear correlation coefficient or Pearson Product Moment Correlation Coefficient.

7.3.2 Calculation of r using formula

For a sample of n measurements on x and y , the product-moment correlation coefficient r is given by

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

(formula given by MF 26)

Note:

1. $\bar{x} = \frac{\sum x}{n}$ and $\bar{y} = \frac{\sum y}{n}$
2. The derivation of these two formulae is beyond the scope of the A Level syllabus.

Example 3 (Use Formula when data are given in summary form)

10 sets of lengths (x) and breadths (y) in mm have been taken and the data has been summarized as:

$$\sum x = 1782, \sum y = 1483, \sum x^2 = 318086, \sum y^2 = 220257, \sum xy = 264582.$$

Find the product moment correlation coefficient for the sample and comment on your answer.

Solution:

$$\begin{aligned} r_{xy} &= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}} \\ &= \frac{264582 - \frac{(1782)(1483)}{10}}{\sqrt{\left(318086 - \frac{1782^2}{10}\right)\left(220257 - \frac{1483^2}{10}\right)}} = 0.744 \end{aligned}$$

Lengths and breadths have positive and moderately strong linear correlation.

7.3.3 Calculating r using the Graphic Calculator


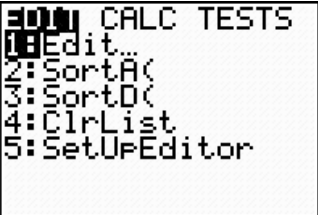
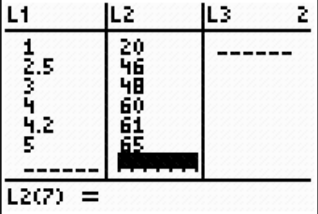
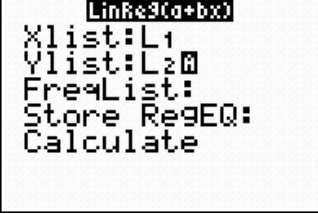
Your graphic calculator can be used to calculate the value of r for a given set of data.

Example 4

The table shows the number of hours (x hours) that 6 average students spent studying for a test and the marks obtained (y marks out of 100 marks). Find the linear product-moment correlation coefficient between x and y .

x	1	2.5	3	4	4.2	5
y	20	46	48	60	61	65

To find the value of r , we may use the following GC keystrokes:

Steps	Screenshot	Remarks
Press [MODE] then [ARROW DOWN]. Select [On] under Stat Diagnostic.		Alternatively, you may also Press [2ND] [0] to select [CATALOG]. Select [DiagnosticOn] from the menu and press [ENTER] to turn on GC's diagnostic mode Note: If DiagnosticOff is used, r and r^2 will not be displayed.
Select [STAT], followed by [EDIT]. Select [1]: [Edit]		
Enter the values of x in list L_1 and values of y in L_2 .		
Select [STAT], followed by [CALC]. Select [8]: [LinReg (a+bx)] followed by [ENTER] Select [L_1] as X list and [L_2] as Y list followed by [ENTER].		In this example, the values of the independent variable (x -values) are stored in L_1 and the values of the dependent variable (y -values) are stored in L_2 . Alternatively, press [4] for

	<pre> LinReg y=a+bx a=12.94791837 b=11.28489796 r²=.9516985687 r=.9755503927 </pre>	LinReg ($ax+b$)
--	--	-------------------

Solution:

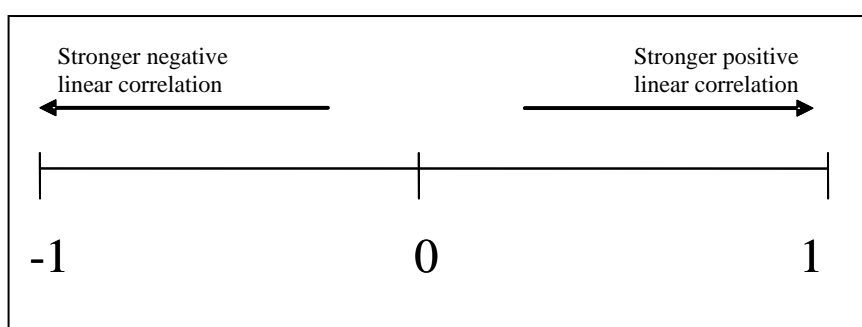
From GC, linear product-moment correlation coefficient between x and y , $r = 0.976$ (3 sf).

7.3.4 Properties of Product Moment Correlation Coefficient, r

- The **sign** of the correlation coefficient indicates the **direction** of linear correlation.
 - When $r > 0$, there is a **positive linear correlation** between the variables i.e. as x increases, y also increases.
 - When $r < 0$, there is a **negative linear correlation** between the variables i.e. as x increases, y decreases.
- The **magnitude** of r indicates the **strength** of the linear correlation, where $-1 \leq r \leq 1$. The closer the value of r is to 1 (or -1), the closer the points on a scatter diagram are to a positively (or negatively) sloped straight line and the stronger the linear relationship.
 - When $r = 1$, there is a **perfect** positive linear correlation. All the points lie exactly on a straight line with positive gradient.
 - When $r = -1$, there is a **perfect** negative linear correlation. All the points lie exactly on a straight line with negative gradient.
 - When $r = 0$, there is **no linear** correlation. However it does not mean there is no relationship between the two variables.

As a general guide,

- $0.8 \leq |r| < 1$ indicates **strong** linear correlation between the two variables.
- $0.5 \leq |r| < 0.8$ indicates **moderately strong** linear correlation between the two variables.
- $0 < |r| < 0.5$ indicates **weak** linear correlation between the two variables.



3. r is a measure of the degree of scatter, it is independent of the units used to measure x and y .

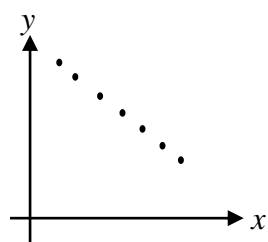
For example, if X is the weight in pounds and Y is the height in inches of a person, then the correlation between X and Y would be the same as if X is measured in kilograms and Y in metres.

4. The value of product moment correlation coefficient, r has no units.

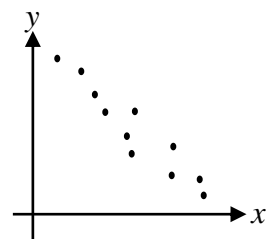
Example 5 Interpreting scatter diagrams and r

Describe the type of correlation between the two variables for each of the scatter diagrams and suggest an appropriate value for r .

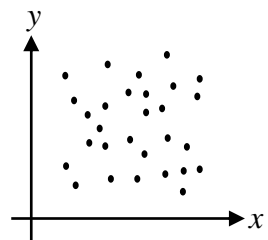
Solution:



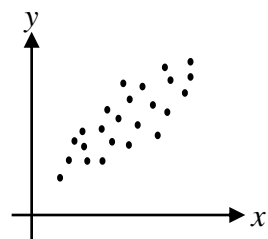
(a)
Perfect negative linear correlation
 $r = -1$



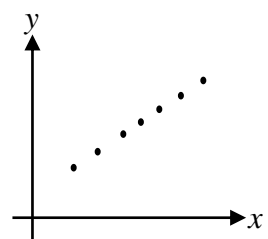
(b)
Strong negative linear correlation
 $r = -0.9$



(c)
Little or no linear correlation
 $r = 0.1$



(d)
Moderately strong positive linear correlation
 $r = 0.7$



(e)
Perfect positive linear correlation
 $r = 1$

7.3.5 Limitations of Product Moment Correlation Coefficient, r

1. The value of r can only provide information on the **degree of linear correlation** but **not causality**. In other words, an increase in one variable **does not necessarily cause** an increase or decrease in the other variable.

For example, consider two variables, the height and weight of a sample of people. If $r = 0.8$, it indicates a positive linear correlation between height and weight. However, this does not mean that being heavy causes one to grow taller, nor does it necessarily mean that being tall causes one to become heavier.

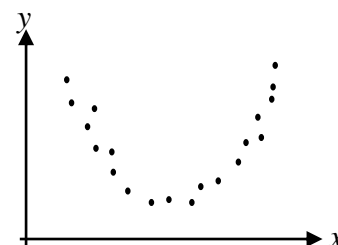
2. The value of r should be considered in conjunction with a **scatter diagram**. Note that regardless of whether you plot x against y , or y against x , the value of r is the same.

A scatter diagram is useful in showing correlation as r does not show any kind of relation except linear correlation.

	Scatter Plot	Product Moment Correlation Coefficient, r .
Check the strength and direction of linear correlation between X and Y .	✓	✓
Check if there's any curved/linear relationship between X and Y .	✓	×

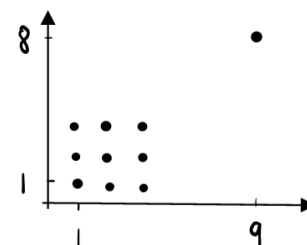
EXAMPLE

- (a) The scatter diagram on the right gives a value of r that is close to 0, but there is an obvious curvilinear relationship.



- (b) The scatter diagram on the right has $r = 0.862$. Based on this value of r , we would conclude that there is strong positive linear correlation between the variables.

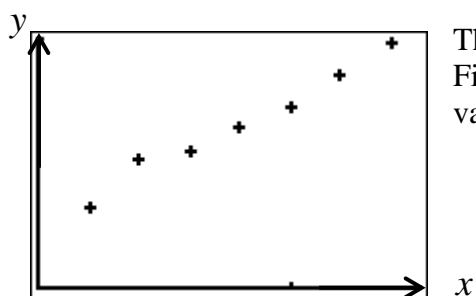
However, the scatter diagram shows that there is an outlier at $(9, 8)$. If we remove this outlier, then there is actually no clear relation between the variables.



Example 6

- (a) We are interested in the relationship between the amount of food supplement fed to hens (in grams/day) and the hardness of the shells of eggs laid (on a scale of 1 to 10).

Food supplement, x (g/day)	2	4	6	8	10	12	14
Hardness of shells, y	3.2	5.2	5.5	6.4	7.2	8.5	9.8



The scatter diagram is given on the left. Find r and comment on the relation between the variables.

Solution:

$$r = 0.987$$

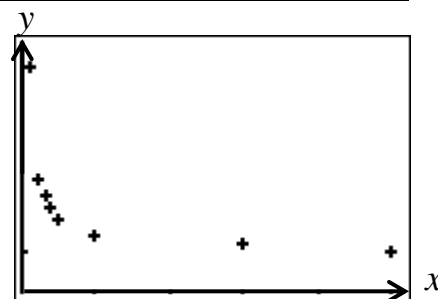
Since the value of r is positive and close to 1, and the scatter diagram reveals a strong linear relationship, there is a strong positive linear correlation between the amount of food supplement fed to hens and the hardness of the shells of eggs laid.

- (b) The daily rate charged by a car-hire firm varies with the length of the hire period, as shown in the table below.

Rental period, x days	1	2	3	4	5	10	30	50
Daily rate, \$ y	149	119	115	112	109	105	103	101

The scatter diagram is given on the right.

Find r and comment on the relation between the variables.



Solution:

$$r = -0.563$$

Since the value of r is negative but not close to -1, and the scatter diagram suggests a curved relationship, there is a negative correlation between the length of the hire period and the daily rate.

7.3.6 The Effect of Transformation on r

Example 7

The table shows the number of hours (x hours) that 6 average students spent studying for a test and the marks obtained (y marks out of 100 marks).

x	1	2.5	3	4	4.2	5
y	20	46	48	60	61	65

The teacher decided to scale the test marks by multiplying each mark by 2 and then add 10 marks to obtain the new mark z . Find the linear product-moment correlation coefficient between x and z and comment on the value obtained.

What do you notice about the value of r compared to the value found in Example 4?

Solution:

To find the value of r , we may use the following GC keystrokes:

New mark, $z = 2y + 10$

Store the values for z in L_3 by keying in the formula

$$L_3 = 2L_2 + 10.$$

Select [STAT], followed by [CALC].

Select [8]: [LinReg (a+bx)] followed by [ENTER]

Select [L_1] as X list and [L_3] as Y list followed by [ENTER].

L1	L2	L3	3
1	20	-----	
2.5	46	-----	
3	48	-----	
4	60	-----	
4.2	61	-----	
5	65	-----	

L3 = 2L2 + 10			
L1	L2	L3	3
1	20	60	
2.5	46	102	
3	48	106	
4	60	130	
4.2	61	132	
5	65	140	

L3(1)=50			
LinReg(a+bx)			
Xlist:L1			
Ylist:L3			
FreqList:			
Store RegEQ:			
Calculate			
LinReg			
y=a+bx			
a=35.89583673			
b=22.56979592			
r ² =.9516985687			
r=.9755503927			

From GC, linear product-moment correlation coefficient between x and z is $r = 0.976$.

The value of r suggests a strong positive linear correlation between the amount of time spent studying and the marks obtained on the test i.e. a student who spends more time studying is likely to obtain higher marks on the test.

The values of r calculated are exactly the same whether we consider x and y , or x and $(2y + 10)$, because the value of r is **NOT** affected by a linear transformation on x or y .

Exercise 2

1. Explain why it is advisable to plot a scatter diagram before interpreting a correlation coefficient calculated for a sample drawn from a bivariate distribution.

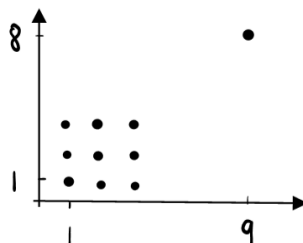
Solution:

A scatter diagram is useful in showing correlation as the correlation coefficient does not show any kind of relation except linear correlation. By plotting a scatter diagram, we are able to have an overview of how the variables are related and hence give an indication of the strength of correlation (which may not be linear) between the variables.

2. A sample linear correlation coefficient between two random variables X and Y gave a value close to 1. Explain why this need not indicate a linear relation between X and Y .

Solution:

A sample linear correlation coefficient between two random variables X and Y is a measure of the strength of linear correlation between the two variables, but it does not provide further information about the actual relation between the variables.



3. With the aid of a graphic calculator, find the value of the product moment correlation coefficient for each of the following sets of data.

(a)

x	1	2	4	5	7	8
y	6.5	5.5	4	4	7	2.5

(b)

x	98	84	75	66	50	43	31
y	96	91	86	85	73	66	54

Solution:(a) -0.436 (b) 0.975

4. For each of the data set in Question 3, use the product moment correlation coefficient to suggest the correlation between the values of x and y .

Solution:

(a) Weak negative linear correlation

(b) Strong positive linear correlation

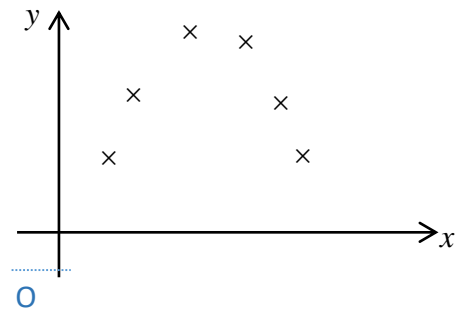
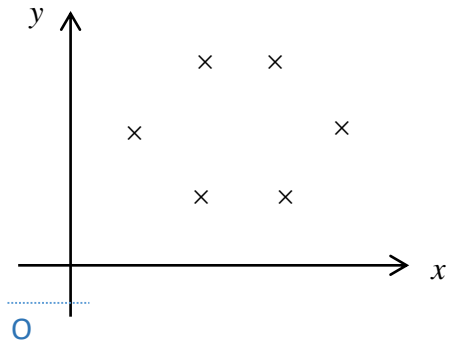
5. Six pairs of values of variables x and y are measured where x and y are positive values.

Draw a sketch of a possible scatter diagram of the data for each of the following cases:

- (i) The product moment correlation coefficient is approximately zero.
- (ii) The product moment correlation coefficient is approximately -0.9 .

Solution:

- (i) Possible solutions:



- (ii)



7.4 Linear Regression

7.4.1 Types of regression lines

Linear Regression attempts to model the relationship between two variables by fitting a linear equation to a set of observed data.

Before attempting to fit a linear model to the observed data, we should first determine whether there is a linear relationship between the two variables. **A scatter diagram and the product moment correlation coefficient, r , between the two variables will give us a good indication of whether it is meaningful to model the observed data with a straight line.**

To calculate the equation of the linear regression line, the method most commonly used is called the least squares method. The resulting line is called the line of best fit or the **least squares regression line** because we have to work backwards or regress, using the given points, to find the original linear equation.

In general, given a set of bivariate data, we can find two types of regression lines:

1. the **regression line of y on x** , and
2. the **regression line of x on y** .

To decide which one is the appropriate line to construct, we can consider the context of the problem and the purpose of the regression line.

x	y	Purpose	Equation of Regression Line	Remarks
Independent	Dependent	Predict y given x	y on x	
		Predict x given y	y on x	
Random	Random	Predict y given x	y on x	Treat x as the independent variable
		Predict x given y	x on y i.e. $x = dy + e$	Treat y as the independent variable

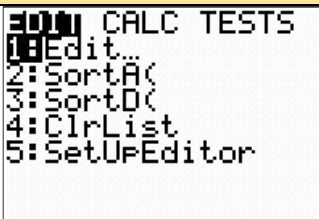
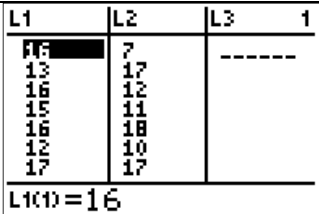
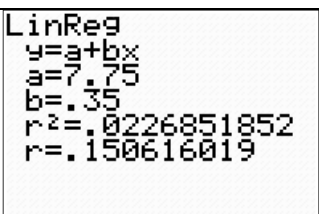
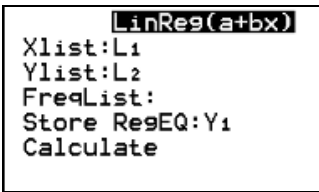
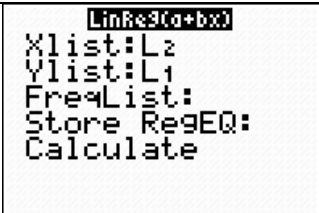
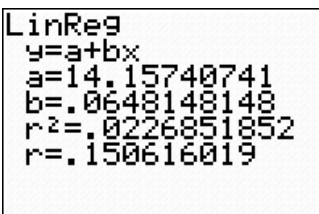
7.4.2 Using Graphic Calculator to find Equations of Regression Lines

To begin, let us try to find regression lines for the following set of data, which contains eight pairs of observations of the variables x and y .

x	16	13	16	15	16	12	17	15
y	7	17	12	11	18	10	17	12

The GC can be used to find the equations of both types of regression lines: the regression line of y on x , and the regression line of x on y .

GC Keystrokes

Steps	Screenshot	Remarks
Select [STAT], followed by [EDIT]. Select [1]: [Edit]		
Enter the values of x in list L_1 , and values of y in L_2		
Select [STAT], followed by [CALC]. Select [8]: [LinReg (a+bx)] followed by [ENTER] Select [L_1] for the XList, then select [L_2] for the YList, followed by [ENTER] If the question asked to sketch the regression line on the scatter diagram. You can insert the regression line equation in “Store RegEQ”. Select [Alpha F4] to key in Y_1	 	Regression line of y on x: The regression line of y on x is $y = 0.35x + 7.75$ We can use this equation to find predicted values of y for given values of x .
Select [STAT], followed by [CALC]. Select [8]: [LinReg (a+bx)] followed by [ENTER] Select [L_1] for XList, then select [L_2] for the YList, followed by [ENTER]	 	Regression line of x on y: The regression line of x on y is $x = 0.0648y + 14.157$ Note: The “ y ” on GC screen is now x , and the “ x ” on GC screen is now y . We can use this equation to find predicted values of x for given values of y .

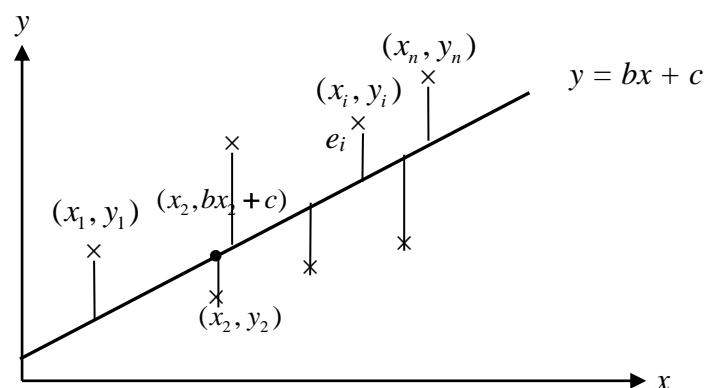
With the GC, it is easy to find the equations of both types of regression lines.

Question: How does the GC compute the equations of regression lines, and how was this method derived?

7.4.3 Least Squares regression line of y on x

In this section, we will explain the least squares method for deriving the equation of the regression line of y on x .

Given a set of n pairs of bivariate data with x as the independent variable and y as the dependent variable, our aim is to find a line with equation $y = bx + c$, where b is the gradient and c is the y -intercept.



Suppose the line $y = bx + c$ in the above figure is the line that provides a minimum vertical deviation of predicted values, $bx_i + c$, from the observed values, y_i . The deviation of a predicted value from an observed value is called a residual, denoted by e_i .

$$\begin{aligned} e_i &= \text{observed } y \text{ value} - \text{predicted } y \text{ value} \\ &= y_i - (bx_i + c) \end{aligned}$$

Hence,

the least squares regression line of y on x is the line (compared to all other lines drawn) that produces the least sum of the square of the residuals, $\sum_{i=1}^n e_i^2$.

It can be written as $y = bx + c$.

It has been mathematically proven that the equation of the line $y = bx + c$ where $\sum_{i=1}^n e_i^2$ is minimised is given by the formula:

$$y - \bar{y} = b(x - \bar{x}), \text{ where } b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \quad (\text{as in MF 26})$$

Note:

1. b is also known as the **coefficient of regression** of y on x .
2. The point (\bar{x}, \bar{y}) lies on the regression line $y = bx + c$.

Interpretation of the Slope and Intercept of a linear regression line

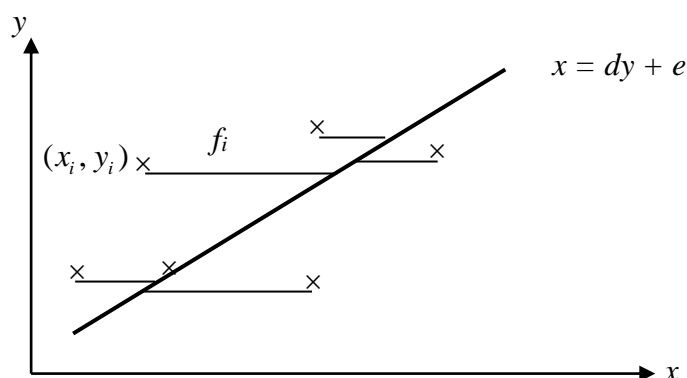
For the regression equation $y = bx + c$, b is the **gradient** (slope) of the line and c is the **y-intercept** of the line. It is necessary to interpret the meaning of the values of b and c in the context of the question.

c is the value of y when $x = 0$. Note that this value of y is meaningful only if the range of values of x in the bivariate data includes 0 or has values near 0.

b is the amount by which y increases for every unit of increase in x i.e. b is the rate of change of y with respect of x .

7.4.4 Least Squares regression line of x on y

Similarly, the least squares regression line of x on y can be written as $x = dy + e$, and is the line that produces the least sum of the residuals, $\sum f_i^2$.



The least squares regression line of x on y can be written as $x = dy + e$.

$$\text{or } x - \bar{x} = d(y - \bar{y}), \text{ where } d = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2}$$

(NOT in MF 26)

Note:

- d is also known as the **coefficient of regression** of x on y .
- The point (\bar{x}, \bar{y}) lies on the regression line $x = dy + e$. Hence the point (\bar{x}, \bar{y}) lies on both types of regression lines.
- The equation of the regression line of x on y cannot be found by making x the subject in the equation of the regression line of y on x . In general, the two lines are not identical unless there is perfect linear correlation.
- Equation $x = dy + e$ may be written as $y = \frac{x - e}{d} = \left(\frac{1}{d}\right)x - \frac{e}{d}$.

Therefore the gradient of the line is $\frac{1}{d}$.

Example 8 [A Comprehensive Guide: H2 Mathematics for A Level Vol. 2 (Modified)]

Eight pairs of observations on the variables x and y are given below.

x	16	13	16	15	16	12	17	15
y	7	17	12	11	18	10	17	12

- (i) Having found earlier that the regression line of y on x is $y = 0.35x + 7.75$, and the regression line of x on y is $x = 0.0648y + 14.157$, sketch the two regression lines on a scatter diagram. Find the point of intersection of the two lines.
- (ii) Use an appropriate regression line to estimate the
- value of x if the value of y is correctly determined as 16. Explain why this is not a good estimation.
 - value of x when y is 13.
 - value of y when x is 15.

Solution:

y on x : $y = 0.35x + 7.75$
 x on y : $x = 0.0648y + 14.157$

1. For **regression line of x on y** , we have to **rearrange the variables and express y in terms of x** before sketching.

$$x = 0.0648y + 14.157$$

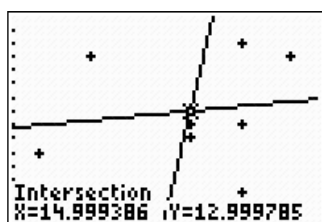
$$y = \frac{x - 14.157}{0.0648}$$

We need to enter this equation in Y_2 by pressing [Y=].

2. To find intersection point, press [2ND] [CALC], select [5]:[intersect].

(i)

$$y = 0.35x + 7.75$$



$$y = \frac{x - 14.157}{0.0648}$$

The point of intersection is (15, 13).

Note:

Notice that the intersection point of the two regression lines is at $\bar{x} = 15$, $\bar{y} = 13$.

```

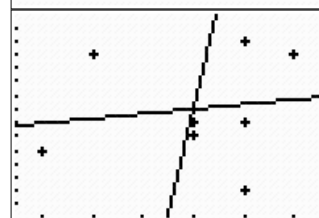
[2ND] Plot2 Plot3
Y1=7.75+.35X
Y2=(X-14.157)/.0648
Y3=
Y4=
Y5=
Y6=

```

```

[2ND] Plot2 Plot3
Off
Type: [ ] [ ] [ ]
Xlist:L1
Ylist:L2
Mark: [ ] [ ]

```



```

[2ND] [CALC]
1:value
2:zero
3:minimum
4:maximum
5:intersect
6:dy/dx
7:ff(x)dx

```

```

Y1=7.75+.35X
First curve?
X=15.010638 Y=13.003723

```



- (ii) (a) As y is correctly determined as 16, y is controlled. Thus y is the independent variable. Hence the regression line of x on y should be used to estimate the value of x .

$$\text{So } x = 14.157 + 0.0648(16) = 15.2 \text{ (3 s.f.)}$$

This is not a good estimation because the scatter diagram does not show a clear linear relationship between x and y .

- (b) Since there is no clear indication of the independent variable, the regression line of x on y should be used to estimate the value of x given $y = 13$. Hence $x = 14.157 + 0.0648(13) = 15.0$

- (c) Since there is no clear indication of the independent variable, the regression line of y on x should be used to estimate the value of y given $x = 15$. Hence $y = 0.35(15) + 7.75 = 13$

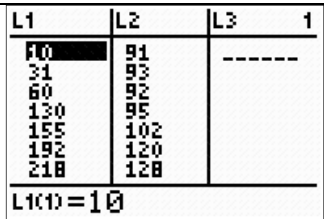
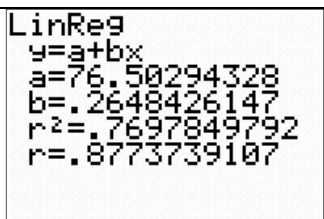
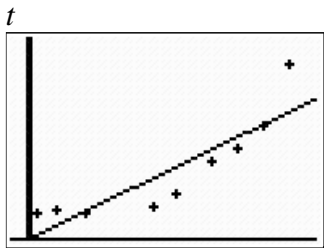
Example 9 [CJC 2008 Prelim (Modified)]

Singapore Stock Exchange (SGX) tracked the price of crude oil per barrel from 1st January 2008. The observations are presented in the table below.

Days in 2008 (t days) from 1 st Jan 2008	10	31	60	130	155	192	218	245	271
Price of Crude Oil per Barrel (US\$ x)	91	93	92	95	102	120	128	140	175

- Find the equation of the regression line of x on t in the form $x = at + b$. Explain the meaning of the slope and intercept in the context of this question.
- Find the linear product-moment correlation coefficient between x and t . Comment on what its value implies about the regression line found in (i).
- On the same diagram, plot a scatter diagram and sketch the regression line of x on t . Comment on the suitability of using a linear model in this case.

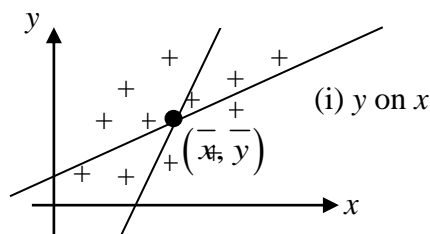
Solution:

<p>(i) Equation of regression line of x on t is $x = 0.26484t + 76.503$</p> <p>Slope: The price of the crude oil per barrel increases by US\$0.265 for every increase of 1 day in 2008.</p> <p>Intercept: The price of the crude oil per barrel is \$76.50 on 1st January 2008.</p>	
<p>(ii) $r = 0.877$ The value of r shows that there is a strong positive linear correlation between x and t. Hence all the data points lie close to the regression line.</p>	
<p>(iii) From the scatter diagram, the relationship between x and t is more curved than linear, hence the linear model is not very suitable in this case.</p>	

7.4.5 Properties of Regression Lines

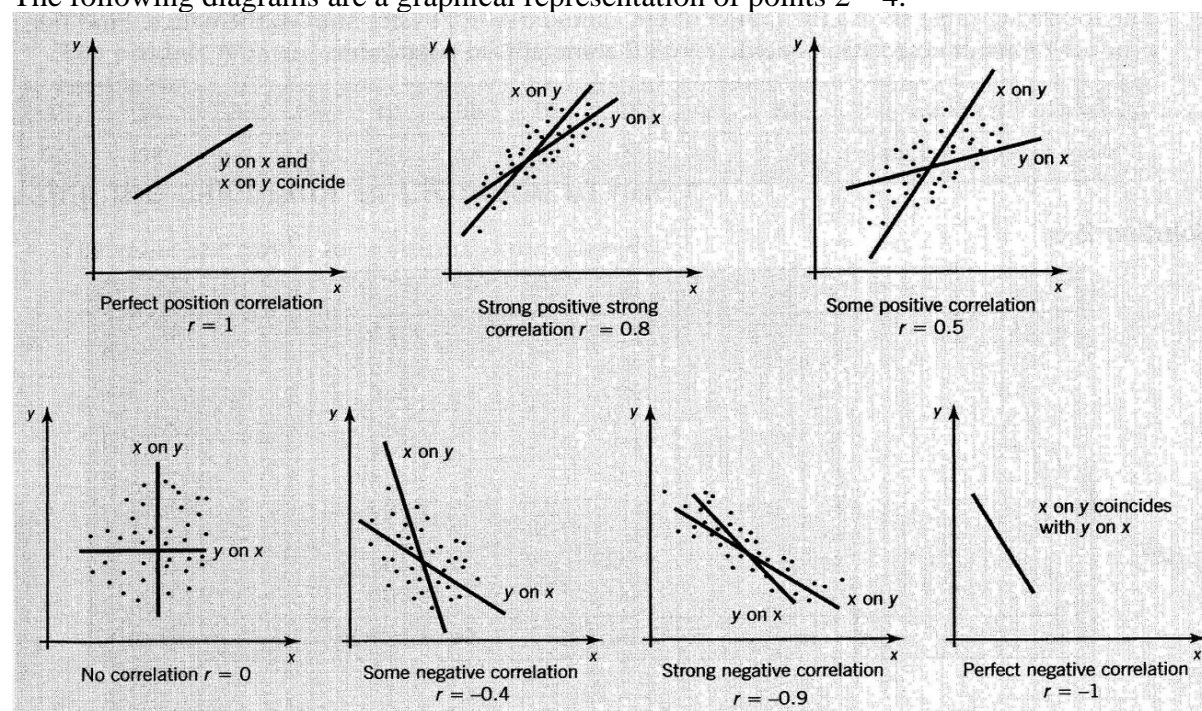
- The point (\bar{x}, \bar{y}) lies on both the regression lines of y on x , and x on y . Hence the two regression lines intersect at (\bar{x}, \bar{y}) . (as observed in Example 7)

(ii) x on y



- The bigger the value of $|r|$, the closer the regression lines of y on x is to the regression line of x on y .
- If there is perfect linear correlation i.e. $r = \pm 1$, the regression lines of y on x and x on y are identical.
- If there is no linear correlation i.e. $r = 0$, the regression lines of y on x and x on y are at right angles.

The following diagrams are a graphical representation of points 2 – 4.



Example 10 [SRJC 2007 Prelim]

The table below gives the observed values of bivariate x and y .

x	20	30	34	35	36	40	42
y	32	25	a	22	26	18	19

It is given that the equation of the regression line of y on x is $y = 43.5 - 0.602x$.
Find the value of a correct to the nearest integer.

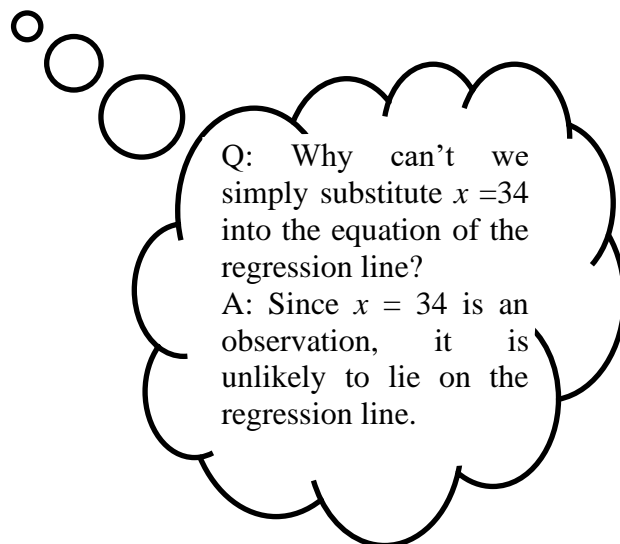
Solution:

$$\bar{x} = \frac{20 + 30 + 34 + 35 + 36 + 40 + 42}{7} = \frac{237}{7}$$

$$\bar{y} = \frac{32 + 25 + a + 22 + 26 + 18 + 19}{7} = \frac{142 + a}{7}$$

The point (\bar{x}, \bar{y}) lies on the regression line of y on x , $y = 43.5 - 0.602x$.

$$\Rightarrow \frac{142 + a}{7} = 43.5 - 0.602\left(\frac{237}{7}\right) \Rightarrow a = 19.826 \approx 20.$$



7.5 Interpolation and Extrapolation

Interpolation	Extrapolation
<p>Interpolation refers to the process of predicting an unknown value that is within the range of the sample data.</p>	<p>Extrapolation refers to the process of predicting an unknown value that is outside the range of the sample data.</p>

EXAMPLE

Consider the table of data which shows the age (x years) of 8 female athletes and their time (y seconds) taken to complete a 100m race.

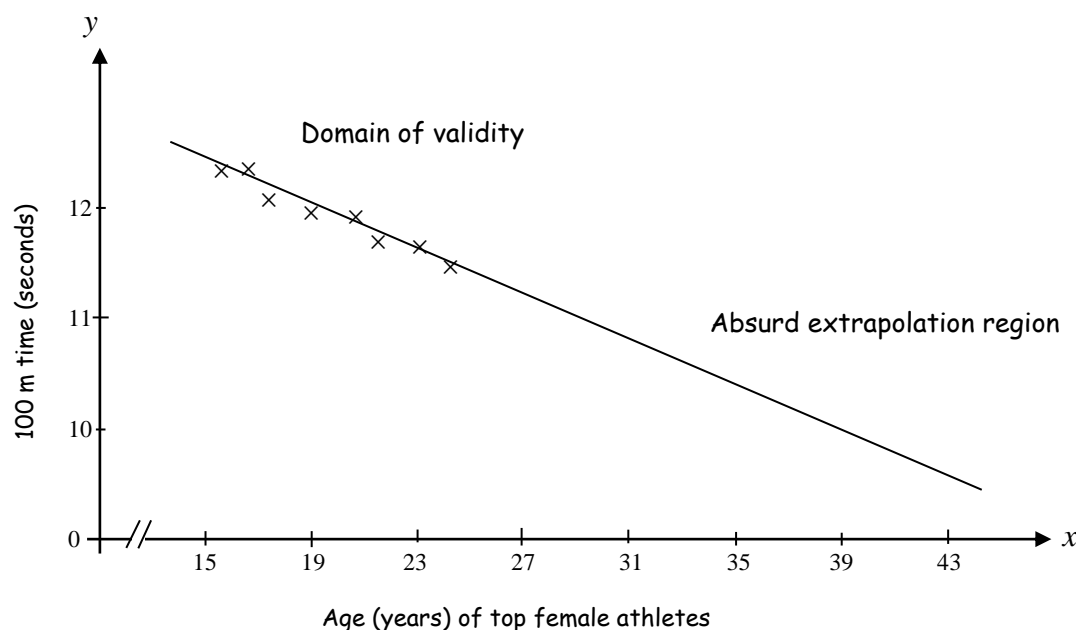
x years	15.5	16.5	17.2	19.0	20.5	21.4	23.0	24.3
y seconds	12.4	12.4	12.1	12.0	11.9	11.7	11.5	11.4

Observe that the least and greatest values of x are 15.5 and 24.3 years respectively.

Using the least squares regression line of y on x to estimate values of y when x is between 15.5 and 24.3 years is known as **interpolation**.

Using the line to estimate values of y when x is smaller than 15.5 or greater than 24.3 is known as **extrapolation**. Note that **extra care** must be taken when **extrapolation** is used. The more remote the prediction is from the range of values used to fit the model, the riskier the prediction becomes since there is no way to check that the relationship continues to be linear.

For instance, there is a strong correlation between the age in years and the 100 m race times of female athletes between the ages of 15.5 and 24.3 years. To extend the connection, the following figure would suggest that veteran athletes are quicker than athletes who are in their prime, and if they live long enough, they can complete the 100 m race in 0 seconds!



Example 11 [SRJC 2008 Prelim (Modified)]

An engineer collected data on how the temperature, T °C, generated in the shoulder of a car tyre varies with its speed, V km/h, when the car is driven under specified conditions. He found that the two variables follow a linear model with regression line of T on V as

$t = 0.8667v + 26.5833$ with product-moment correlation coefficient, $r = 0.9939$. As a backup, the engineer's assistant was also on site to record the temperature readings when the experiment was being carried out. The assistant submitted his observations as shown in the table below:

v	20	30	40	50	60	70	80	90
t	45	52	64	66	66	89	98	104

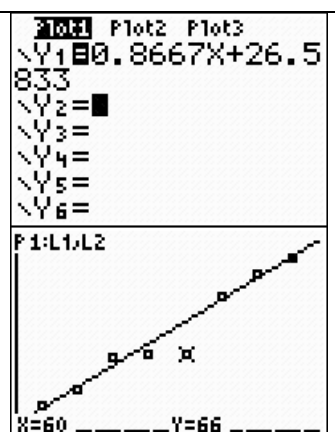
It came to the engineer's attention that his assistant's temperature readings contained an error. Having misplaced his original records, the engineer decided to work out the correct data. He is certain that the wrong t value (the corresponding v value remains correct) occurred in one of these three data points: (60, 66), (70, 89) or (80, 98).

- Which one of the above three data pairs is most likely to contain the wrongly entered value of t ? Show clearly how you are able to identify the error.
- Find the correct value of t recorded by the engineer giving your answer correct to 3 significant figures.
- With the use of a suitable regression line, estimate the value of
 - v when $t = 70$.
 - t when $v = 300$
 Give a reason for the choice of line and comment on the reliability of your estimation.

Solution:

- (i) By plotting the values recorded by the assistant on a scatter diagram, together with the regression line calculated by the engineer.

Of the 3 points suspected to be wrong, (60, 66) is the data point with the largest vertical distance from the line. Thus, it is most likely the point containing the wrong value of t is 66.



- (ii) $t = 0.8667v + 26.5833$

Since (\bar{v}, \bar{t}) lies on the regression line,

$$\bar{t} = 0.8667\bar{v} + 26.5833 \quad \text{and} \quad \bar{v} = \frac{\sum v}{n} = \frac{440}{8} = 55$$

$$\bar{t} = 0.8667(55) + 26.5833 = 74.2518$$

$$\sum t = 8(74.2518)$$

$45 + 52 + \dots + t' = 594.0144$ $518 + t' = 594.0144$ $t' = 76.014 \approx 76.0$ (3 s.f.)	
<p>(iii)</p> <p>(a) Since v is the independent variable (predetermined and controlled), use t on v. $t = 0.8667v + 26.5833$ When $t = 70$, $70 = 0.8667v + 26.5833$ $v = 50.09 \approx 50.1$ The answer obtained by interpolation is reliable as the value of r suggests a strong positive linear correlation between the two variables for this given range of data.</p>	
<p>(b) Since v is the independent variable (predetermined and controlled), use t on v. $t = 0.8667v + 26.5833$ when $v = 300$, $t = 0.8667(300) + 26.5833$ $t = 0.8667(300) + 26.5833$ $= 286.593 \approx 287$ The answer obtained by extrapolation is not reliable.</p>	

Exercise 3

1. For the given set of data,

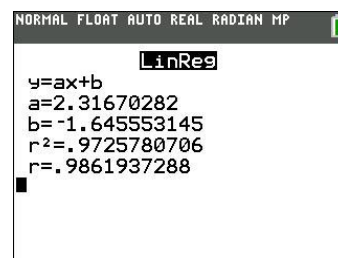
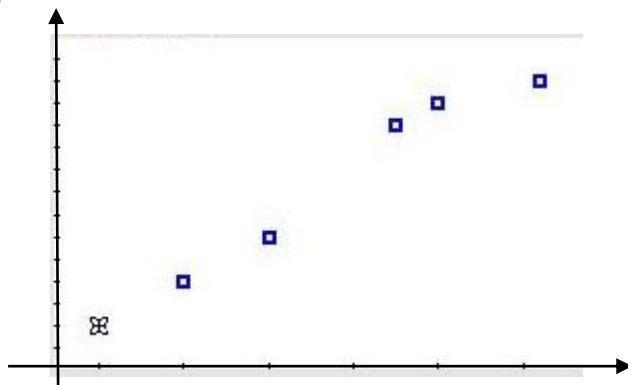
x	5	6	7	8.5	9	10.2
y	10	12	14	19	20	21

- (i) Sketch the scatter diagram of y against x .
(ii) Find the correlation coefficient between x and y .
(iii) Find the equation of the regression line y on x .
(iv) Find the equation of the regression line x on y .
(v) Using the line in (iii), estimate the value of y when $x=6.5$

[Ans: (ii) 0.986 (iii) $y = 2.32x - 1.65$ (iv) $x = 0.420y + 0.900$ (v) 13.4]

Solution:

- (i)



- (ii) $r = 0.986$
(iii) $y = 2.32x - 1.65$
(iv) $x = 0.420y + 0.900$
(v) When $x=6.5$, $y = (2.3217)(6.5) - 1.6456 = 13.4$

2. For a given set of data, it is known that
- $\bar{x} = 10$
- and
- $\bar{y} = 4$
- . The gradient of the regression line
- y
- on
- x
- is 0.6. Find the equation of this regression line and estimate
- y
- when
- $x = 12$
- .

Solution:

Equation of the regression line y on x is $y = a + 0.6x$.

Substitute $x = 10$ and $y = 4$, we have $a = -2$.

Equation of the regression line y on x is $y = -2 + 0.6x$.

When $x = 12$, $y = -2 + 0.6(12) = 5.2$.

3. [IJC Prelims 08/9]

A manufacturer of soft drinks launched a new drink. The table shows the weekly advertising expenditure (x) and weekly sales (y) during the launch period.

Advertising, \$ x (in thousands)	1.60	3.05	0.45	2.00	0.75	2.65
Weekly sales, \$ y (in thousands)	240	450	130	324	200	365

- (a) Find the equation of the regression line of y on x in the form $y = a + bx$, giving the values of a and b correct to 1 decimal place.
Interpret the values of a and b in terms of the amount spent on advertising and the weekly sales.
- (b) Find the product moment correlation coefficient of the data, and say what it leads you to expect about the scatter diagram for the data.

[Ans: (a) $y = 89.3 + 111.8x$ (b) $r = 0.981$]

Solution:

(ai) $y = 89.2646 + 111.754x$

$$y = 89.3 + 111.8x$$

- (aii) The weekly sale of the soft drink is \$89300 when no money is spent on advertising.

There is an increase of \$111 800 in weekly sales for every thousand dollars invested in advertising.

- (b) $r = 0.981$. Since $r = 0.981$ is close to 1, the points representing the data on a scatter diagram would be close to a line with a positive gradient.

4. [FM N2006/2/10 (Modified)]

For a random sample of 12 observations of pairs of values (x, y) , the equation of the regression line of y on x is $y = 4.82 - 2.25x$. The sum of the 12 values of x is 20.64 and the product-moment correlation coefficient for the sample is -0.3 .

- (i) Find the sum of the 12 values of y .
- (ii) Find the estimated value of y when $x = 2.8$ and comment on the reliability of this estimate.

[Ans: (i) 11.4 (ii) -1.48]

Solution:

(i) $\bar{x} = \frac{20.64}{12} = 1.72$

Since (\bar{x}, \bar{y}) lies on the regression line of y on x , substitute $\bar{x} = 1.72$ into the regression line.

$$\text{We have } \bar{y} = 4.82 - 2.25(1.72) = 0.95.$$

Therefore, the sum of the 12 values of y is $0.95(12) = 11.4$.

- (ii) When $x = 2.8$, $y = 4.82 - 2.25(2.8) = -1.48$

The estimate is not very reliable as $r = -0.3$, which is close to 0, indicating a weak negative linear correlation.

5. [H1 Math A Level Nov 2007/I/8]

Seven cities in a certain country are linked by rail to the capital city. The table below shows the distance of each city from the capital and the rail fare from the city to the capital.

City	A	B	C	D	E	F	G
Distance, x km	124	44	76	148	16	180	104
Rail fare, \$ y	156	53	99	169	23	177	138

- (i) Give a sketch of the scatter diagram for the data, as shown on your calculator.
 (ii) Calculate the product moment correlation coefficient.

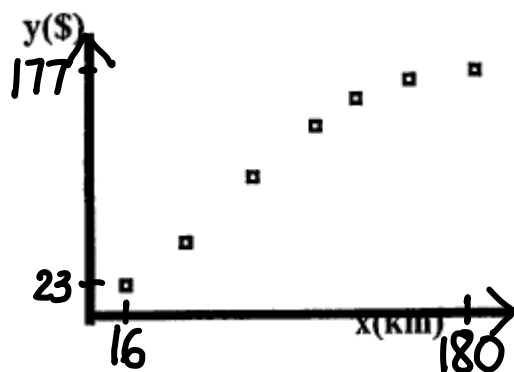
You are given that the regression line of y on x has equation $y = 16.7 + 1.01x$, where the coefficients are given correct to 3 significant figures.

- (iii) Calculate the equation of the regression line of x on y , giving your answer in the form $x = a + by$.
 (iv) Use the appropriate regression line to estimate
 (a) the rail fare from a city that is 28km from the capital,
 (b) the distance of a city from the capital if the rail fare is \$198.
 (v) Comment briefly on the reliability of the estimates in part (iv).

[Ans: (ii) $r = 0.973$ (iii) $x = 0.940y - 10.6$ (iv)(a) \$45.00 (b) 176 km]

Solution:

(i)



- (ii) From GC, $r \approx 0.973479 = 0.973$ (to 3 s.f.)
 (iii) From GC,
 $x = 0.9398y - 10.558$
 i.e. $x = 0.940y - 10.6$
 (iv) (a) When $x = 28$, using the regression line of y on x ,
 $y = 16.7 + 1.01(28)$
 $= \$45.00$ (to 2 d.p.)
 (iv) (b) When $y = 198$, using the regression line of x on y ,
 $x = 0.940(198) - 10.6$
 $= 176$ km (to 3 s.f.)
 (v) The estimate for part (a) is reliable because 28 km falls within the range of x -

values. However, the estimate for part (b) is not reliable because it is an extrapolation, since \$198 falls outside of the range of y -values.

6. [AJC/2009/II/12 (Modified)]

One end A of an elastic string was attached to a horizontal bar and a mass, m grams, was attached to the other end B . The mass was suspended freely and allowed to settle vertically below A . The length AB , l mm, was recorded, for various masses as follows:

m	100	200	300	400	500	600
l	228	236	256	a	285	301

The equation of the regression line of l on m is $l = 0.15257m + 210.6$.

Find the value of a corrected to the nearest integer, and the product moment correlation coefficient of l and m .

[Ans: (i) 0.991]

Solution:

$$(i) \quad \bar{m} = 350, \quad \bar{l} = \frac{1306 + a}{6}$$

Substitute them into regression line of l on m ,

$$\frac{1306 + a}{6} = 0.15257(350) + 210.6$$

$$\Rightarrow a = 277.997$$

$$ie a = 278 \text{ (shown)}$$

From GC, $r = 0.991$

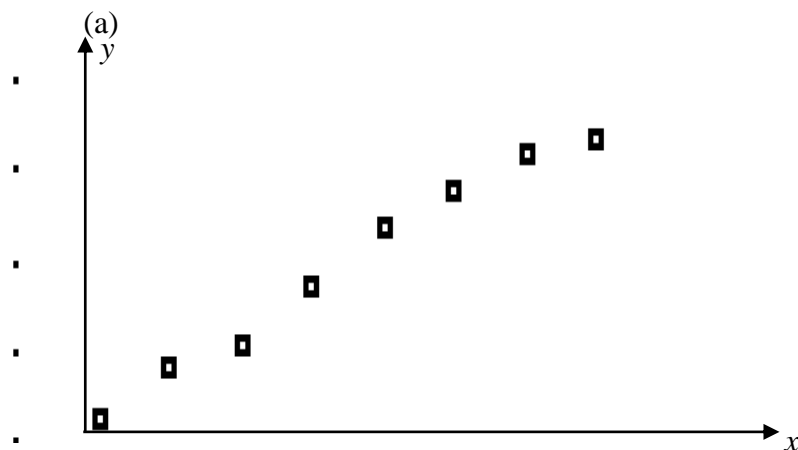
Practice Questions (To go through Q2 and Q3)

1 [AJC Prelims 08/13]

Cucumbers are stored in brine before being processed into pickles. Data were collected on x , the percentage of sodium chloride in the salt used to make brine, and y , a measure of the firmness of the pickles produced. The data are shown below:

x	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5
Y	15.3	15.8	16.1	16.7	17.4	17.8	18.2	18.3

- Give a sketch of the scatter diagram for the data.
- Calculate the product moment correlation coefficient and comment on the value obtained.
- Calculate the equation of the regression line of y on x .
- Interpret in context the value of the gradient and the value of the intercept with the y -axis of the regression line respectively.
- Explain why in this case it would be inappropriate to calculate the equation of the regression line of x on y .
- Use the equation of the line in (c) to predict the value of y when x is
 - 6.7
 - 10.7
 Comment briefly on these predictions.

[Ans: (b) 0.990 (c) $y = 9.79 + 0.924x$ (fi) 16.0 (fii) 19.7]**Solution:**

- $r = 0.990$. Very strong positive linear correlation between x and y
- $y = 9.79 + 0.924x$
- Gradient: Increase in firmness of cucumber per unit increase in the percentage of sodium chloride
y-intercept: The initial firmness of the cucumber before being processed.
- The concentration of brine is the independent variable. Hence the regression line y on x should be used.
- (f i) 16.0.
The prediction should be reasonably reliable as x is within range of observed values and r is close to 1.
- (f ii) 19.7.

The prediction may not be reliable, as x is outside range of observed values.

- 2*. After a major renovation at the computer-mart Sim Lum Square, the manager Keith carried out a study to investigate how their sales figures in the first six months after re-opening depended on their renovation cost. The data below were collected from eight randomly chosen shops, in thousands of dollars.

Shop No.	1	2	3	4	5	6	7	8
Renovation cost, x (in thousands of dollars)	19.5	23	28.1	31.6	35	38.1	46.5	49.2
Sales figure, y (in thousands of dollars)	55.5	a	77.2	78.7	79	88	73.4	80

- Keith found the equation of the regression line of y on x for the above data is $y = 52.244 + bx$. It was known that this line and the regression line of x on y both pass through the point (33.9, 74.3). Find the value of b to 5 significant figures and show that a is 62.6.
- Explain why it is not appropriate to use the regression line of x on y to predict the renovation cost when given the sales figures.
- Give a sketch of the scatter diagram for the data. Calculate the product moment correlation coefficient and comment on the value obtained.

[Ans: (i) 0.65062 (iii) 0.663]

Solution:

- At (33.9, 74.3), $74.3 = 52.244 + b(33.9) \Rightarrow b = 0.65062$

Since line passes through (33.9, 74.3),

$$\bar{y} = 74.3$$

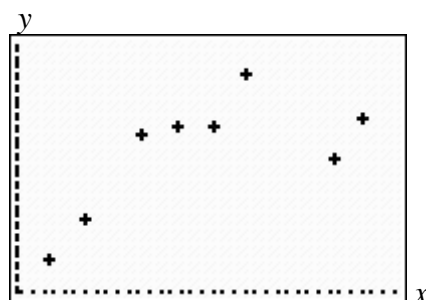
$$\Rightarrow \frac{\sum y + a}{8} = 74.3$$

$$\Rightarrow \frac{531.8 + a}{8} = 74.3$$

$$\Rightarrow a = 62.6$$

- Since the renovation cost is the independent variable, the line x on y is inappropriate for prediction purposes as the line y on x should be used.

-



$r = 0.663$. It suggests that there is a moderately strong positive linear relation between renovation cost and sales.

3*. [NJC 2009/II/11]

At a clinical laboratory, a machine is used to measure the growth of a certain bacteria at fixed time intervals and the results are tabulated as follow:

Time (days), x	5	10	24	35	48	55	67
Number of bacteria (thousands), y	1.3	42.0	14.8	30.1	60.8	81.3	98.6

It was discovered that one of the results may be wrong. Identify the result that is most likely to be incorrect. Justify your answer.

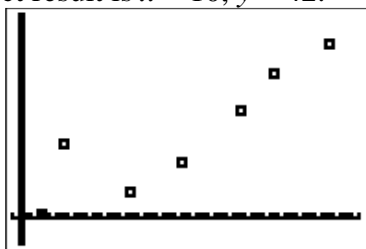
The incorrect result which you identified above is rectified. The data with the correct result yields the following regression lines y on x and x on y respectively,

$$y = -14.16830 + 1.61302x \quad \text{and} \quad x = 9.97265 + 0.59168y.$$

- Determine the value of the correct result, correct to 1 decimal place.
- State, giving a reason, which of the least squares regression lines, y on x or x on y , should be used to express a possible linear relation between y and x .
- Using the regression line you chosen in (ii), an estimate for the number of bacteria in 40 days is obtained. Comment on its reliability.
- For each of the seven sample values of x , Y' is given by $Y' = a + bx$, where a and b are any real constants. Explain why $\sum (y - Y')^2 \geq c$ where c is a constant to be determined.

Solution:

Incorrect result is $x = 10, y = 42$.



From the scatter diagram, $x = 10, y = 42$ is an outlier.

- $y = -14.16830 + 1.61302x$ and $x = 9.97265 + 0.59168y$

Using GC to solve the above simultaneous equations,

$$\bar{x} = 34.8522 \quad \bar{y} = 42.04899$$

Let k be the correct value when $x = 10$,

$$\begin{aligned} k &= 42.04899(7) - (1.3 + 14.8 + 30.1 + 60.8 + 81.3 + 98.6) \\ &= 7.4 \text{ (to 1 decimal place)} \end{aligned}$$

- As x is the independent variable, y on x should be used.
- When $x = 40$, $y = 50.3525$.

The estimate is reliable since $r = 0.977$ is close to 1, indicates a strong positive linear correlation between x and y and $x = 40$ is within data range.

- (iv) c is the sum of least square deviation between the observed value y and the predicted value on the regression line y on x .
Using GC, $c = 401.541488 = 402$ (to 3 sig figs)

7.6 Self-Reading Examples

Example 12 [N2002/2/6 (Modified)]

A random sample of five students is taken from those sitting examinations in English and History, and their marks, x and y , each out of 100, are given in the table.

English mark (x)	56	41	75	88	84
History mark (y)	32	24	70	65	47

Find, in any form, the equation of the regression line of (i) y on x (ii) x on y

Hence, find the point of intersection of the two lines.

A sixth student scored 55 in the History examination, but missed the English examination. Use the appropriate regression line to estimate what his English mark would have been.

Solution:

From GC, (i) $y = 0.838x - 10.1$, (ii) $x = 0.822y + 29.7$

Solving (i) and (ii) simultaneously using GC, $x = 68.8$ and $y = 47.5$ (to 3 s.f.).

(Note: $\bar{x} = 68.8$ and $\bar{y} = 47.6$)

Since both x and y are random, and History marks y is given, we shall use the equation of x on y . Hence, the English marks of the 6th student is given by $x = 0.822(55) + 29.7 = 74.9$ (Ans)

Example 13 [HCI 2009/II/Q10]

The table below shows the Certificates of Entitlement Quota Premiums for eight bidding exercises in 2009:

Small cars premium x (in thousands of dollars)	4.890	5.116	7.090	7.589	8.489	9.889	11.690	12.899
Big cars premium y (in thousands of dollars)	5.101	5.001	7.501	7.490	7.552	9.180	11.889	14.840

- Obtain the value of the linear (product moment) correlation coefficient r for the data. Explain whether we can conclude that the rise in the big car premiums is due to the rise in the small car premiums.
- Plot a scatter diagram for the data and explain how its shape is related to the value of r obtained in (i).

Solution:

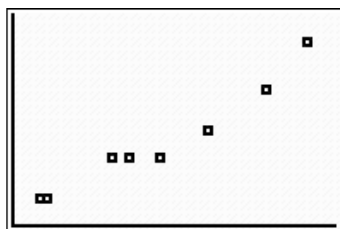
- $r = 0.972$

L1	L2	L3	3
4.89	5.101		
5.116	5.001		
7.09	7.501		
7.589	7.49		
8.489	7.552		
9.889	9.18		
11.69	11.889		
L3(1)=			

```
LinReg
y=ax+b
a=1.125565908
b=-.9490981029
r2=.944222009
r=.971710867
```

Even though r is close to 1, it does not mean that there is a cause and effect relationship between x and y .

-



The points on the scatter diagram lie close to a straight line with positive gradient. This agrees with the value of r obtained in part (i).

Practice Questions**1. RI Prelim 8865/2018/Q9**

Clean Air Company develops and manufactures air purifiers for indoor use. To increase sales, it bought a fixed amount of advertisement time on Channel R each month. Its sales revenues, y thousand dollars in month x , are as follows.

Month x	1	2	3	4	5	6
Sales revenue y	20	23	37	29	33	38

- (i) Give a sketch of the scatter diagram of the data.

[2]

- (ii) Suggest a possible reason why one of the observed sales revenue does not seem to follow the trend.

[1]

The observed data in part (ii) is now omitted.

- (iii) Find the product moment correlation coefficient and comment on its value in the context of the data.

[2]

- (iv) Find the equation of the regression line of y on x in the form $y = mx + c$.

[1]

- (v) State, in context, the meaning of c .

[1]

- (vi) Use the equation of your regression line to calculate estimates of the sales revenue in month 3 and in month 9. Comment on the reliability of your estimates.

[3]

Answers

(iii) 0.994 (iv) $y = 3.5x + 16$ (vi) 26.5, 47.5

Solution

1(i)	
1(ii)	The observed sales revenue in month 3 does not seem to follow the trend. This could be due to a seasonal hazard (such as haze) which increases the demand for air purifiers.
1(iii)	<p>The product moment correlation coefficient, $r = 0.994$</p> <p>Since $r = 0.994$ is close to 1, there is a strong positive linear correlation between the amount of time advertisement was shown on Channel R and the sales revenue in a month.</p>
1(iv)	The regression line of y on x is $y = 3.5x + 16$
1(v)	The value $c = 16$ means that when there is no advertisement time the expected sales revenue is 16 thousand dollars in a month.
1(vi)	<p>When $x = 3$, $y = 26.5$, when $x = 9$, $y = 47.5$.</p> <p>The estimated sales revenue of \$26500 ($y = 26.5$) for month $x = 3$ is reliable as $x = 3$ lies inside the data range of $1 \leq x \leq 6$. Furthermore $r = 0.994$ is close to 1, showing a strong positive linear correlation between x and y.</p> <p>The estimated sales revenue of \$47500 ($y = 47.5$) for month $x = 9$ is not reliable as $x = 9$ lies outside the data range of $1 \leq x \leq 6$.</p>

2. PJC Prelim 8865/2018/Q10

A car manufacturer, investigating the efficiency of experimental brakes on a new model of car, obtains the stopping distances, d metres for different car speeds, s kilometres per hour. The results of 9 sets of data are summarised in the table below.

Data Set	1	2	3	4	5	6	7	8	9
s	20	25	30	35	40	45	50	55	60
d	9.7	14.3	18.6	20.4	9.3	26.8	28.4	32.7	38.6

(i) Give a sketch of the scatter diagram of the data, as shown on your calculator.
[2]

(ii) One of the data set appears incorrect. State which of the set of data is incorrect.
[1]

Omit the incorrect data for the following parts of the question.

(iii) Find the product moment correlation coefficient and comment on its value in the context of the data.

[2]

(iv) Find the equation of the regression line of d on s , in the form $d = as + b$, giving the values of a and b correct to 2 decimal places.

[1]

(v) Estimate the stopping distance of a car travelling at 70 kilometres per hour. Comment on the reliability of your estimate.

[2]

(vi) Given that 1 kilometres per hour = 0.621 miles per hour, re-write your equation from (iv) so that it can be used to estimate the stopping distance, in metres when speed is given miles per hour.

[1]

Answers

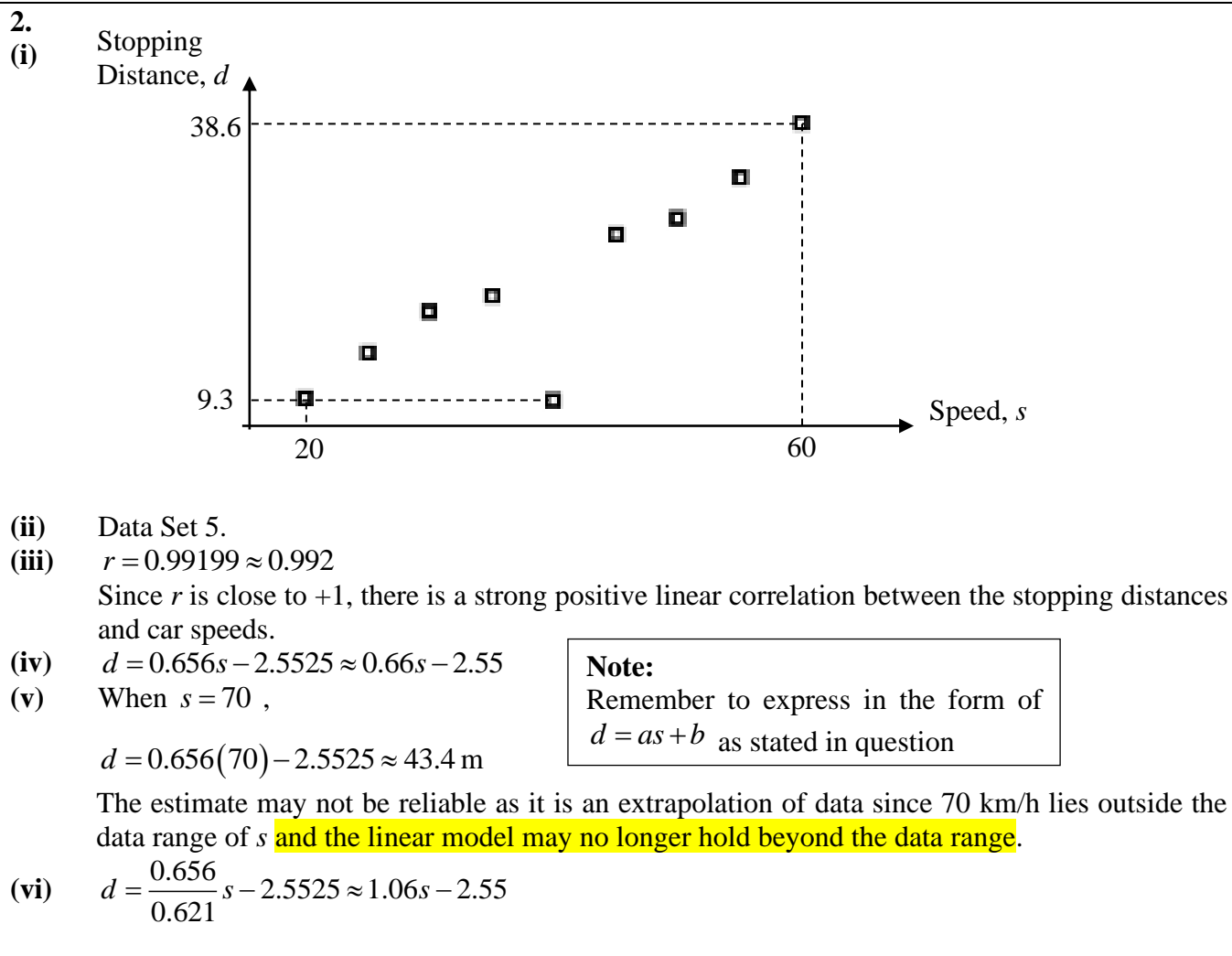
(ii) Data Set 5

(iii) $r \approx 0.992$

(iv) $d = 0.66s - 2.55$

(v) $d = 43.4$ m

(vi) $d = 1.06s - 2.55$

Solution**3. RVHS Prelim 8865/2018/Q7**

VR Secondary School Mathematics department wants to investigate the relationship between the Elementary Mathematics and Additional Mathematics scores obtained by their students. A random sample of 7 students are chosen and their respective scores are given below:

Elementary Mathematics (x)	83	83	80	51	64	74	65
Additional Mathematics (y)	69	67	67	35	50	27	47

- (i) Sketch the scatter diagram for the given data and find the product moment correlation coefficient. [2]
- (ii) On the scatter diagram, **circle** the point that is most likely an outlier of the data. Find the product moment correlation coefficient

when the outlier is omitted from the data.

[2]

- (iii) Explain why it is not appropriate to comment on the relationship between x and y based on just the r -value without reference to the scatter diagram [1]

For the following part, exclude the outlier identified in part (ii) in all calculations.

- (iv) Using the regression line of y on x , estimate a student's Elementary Mathematics score when his Additional Mathematics score is 30, rounding it to the nearest whole number. Comment on the reliability of the estimate. [3]

Answers

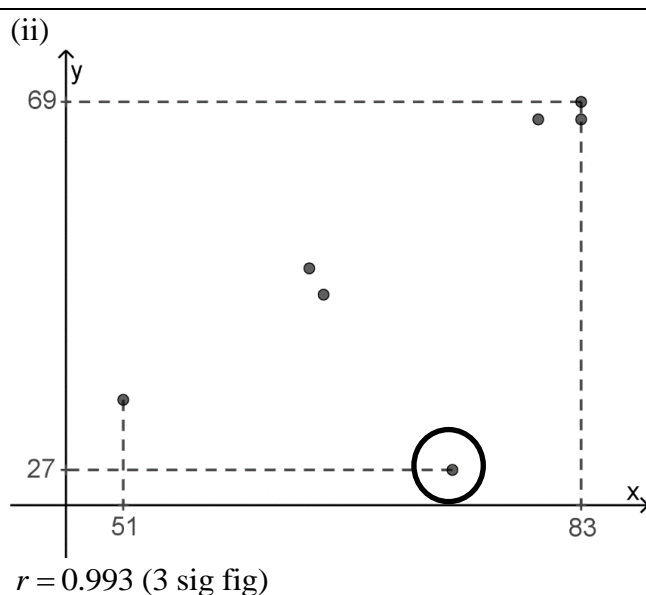
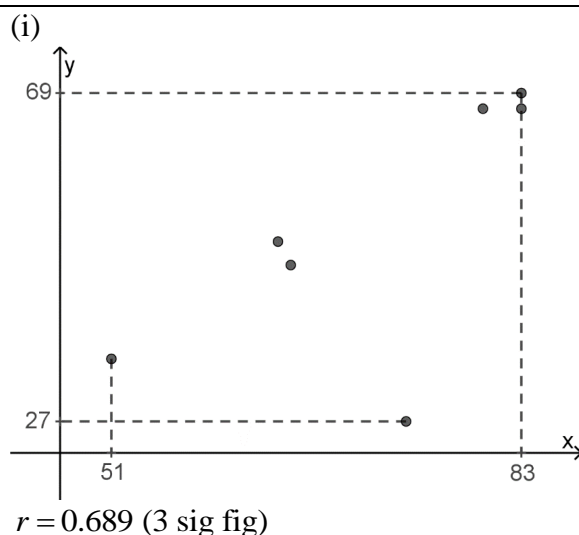
(i) $r = 0.689$ (3 sig fig)

(ii) $r = 0.993$ (3 sig fig)

(iv) 47

Solution

3



	<p>(iii)</p> <p>The r value of 0.689 which is not close to 1 suggests that the 2 scores do not have a strong linear correlation but the scatter diagram shows that other than an outlier point, the rest of the points do exhibit a strong linear correlation.</p>
	<p>(iv)</p> <p>Regression line of y on x is</p> $y = -19.24043716 + 1.057377049x$ <p>When $y = 30$,</p> $30 = -19.24043716 + 1.057377049x$ $x = 46.56847546$ ≈ 47 <p>The estimate is not reliable as the given value of y is outside the range of the data for y after excluding the outlier.</p>

4. SAJC Prelim 8865/2018/Q10

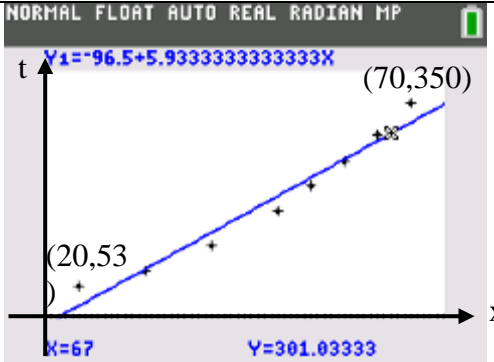
A physical instructor conducted a test to assess the physical fitness level of his student, Mary. To do this, he timed Mary when she carried out a predetermined number of sit-ups on different occasions. The table below shows the number of sit-ups, x , and the time taken, t seconds, Mary took to complete the sit-ups.

x	20	30	40	50	55	60	65	70
t	53	77	120	175	215	254	298	350

- (i) Calculate the product moment correlation coefficient between x and t . Draw a scatter diagram for the above data.
[3]
- (ii) Calculate the equation for the regression line of t on x . Give, in context, an interpretation for the gradient of the regression line of t on x . Sketch also the line t on x on your scatter diagram.
[3]
- (iii) Estimate the time that Mary would take to complete 67 sit-ups. Comment also on the reliability of your answer.
[3]
- (iv) State the value of product moment correlation coefficient if the physical instructor records the time in minutes. Justify your answer.
[2]
- (v) Another physical instructor suggests that, instead of carrying out a predetermined number of sit-ups, Mary should complete as many sit-ups as possible in predetermined periods of time. How, if at all, would your method of the regression equation for data generated in this way differ from your method of calculation in part (ii)? Explain your reason clearly.

	[2]
	Answers
	(i) 0.979
	(ii) $t = 5.93x - 96.5$
	(iii) 301s
	(iv) 0.979

Solution

10 (i)	 <p>The product moment correlation coefficient, r, is 0.979 (3sf).</p>
(ii)	<p>From GC, equation for the regression line of t on x is $t = 5.9333x - 96.5$ $t = 5.93x - 96.5$</p> <p>There is an increase of 5.93 seconds for every situp Mary takes. (see graph for the regression line $t = 5.93x - 96.5$)</p>
(iii)	<p>When $x = 67$, $t = 5.9333(67) - 96.5$ $= 301s$</p> <p>Since $x = 67$, is within data range and the product moment correlation coefficient is close to 1 suggesting a strong positive linear correlation, the predicted value is reliable.</p>
(iv)	The product moment correlation coefficient remains at 0.979 as it is not affected by linear transformation of the variables.
(v)	An appropriate regression equation would be $x = c + dt$ since the time is now the independent variable.

5. SRJC Prelim 8865/2018/Q9

The total distance run per week, x kilometres, and amount of weight loss, y kilograms, of 8 men undergoing a particular special training programme after a period of time are given in the following table.								
x	0.6	1.2	1.5	2.4	2.5	3.2	3.6	3.4
y	0.5	2.5	2.9	5.5	5	8.1	9	10

(i)

Give a sketch of the scatter diagram of the data, as shown on your calculator.

(ii)

Find the product moment correlation coefficient and comment on its value in the

[2]

	context of the data.	[2]
(iii)	Find the equation of the regression line of y on x in the form $y = ax + b$, giving the values of a and b correct to 4 significant figures. Explain the meaning of value of a in the context of the question. Sketch this line on your scatter diagram.	[3]
(iv)	Use a suitable regression line to calculate an estimate of the weight loss for a man who runs 700 metres daily. Comment on the reliability of this estimate.	[2]
(v)	It is decided to record the distance run per week for person in metres instead of kilometres. Without any further calculations, state any change you would expect in the value of the product moment correlation coefficient.	[1]

Answers

(ii) $r = 0.981$. Since r is close to 1, there is a strong positive linear correlation between the total distance run per week and the amount of weight loss. As the total distance run per week increases, the amount of weight loss increases.

(iii) $y = 3.0141x - 1.49497$
 $a = 3.014, b = -1.495$
The meaning of a
For every 1 km increase in the total distance run per week, there is a weight of loss of 3.014 kg.

(iv) 13.3 kg. Although r is close to 1, $x = 4.9$ km is outside the data range $[0.6 - 3.6]$, the linear correlation may not hold, the estimate obtained by extrapolation is not reliable.

(v) There is no change in the value of product moment correlation coefficient.

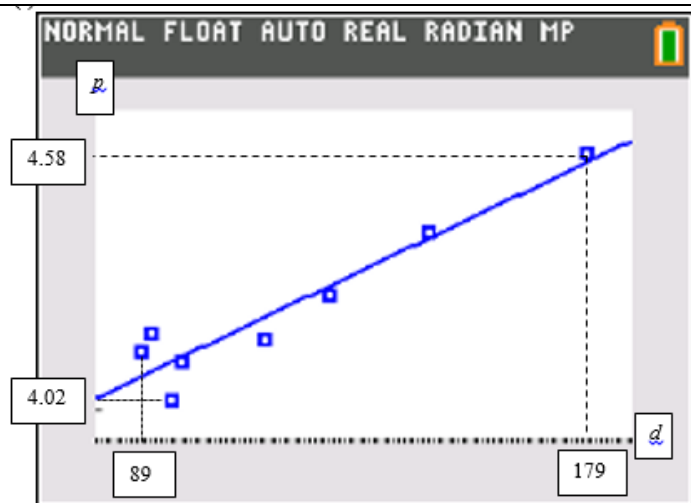
Solution

5	<p>(i)</p> <p>1 mark – the scale is evenly spaced and the axes are clearly labelled</p>
---	---

	<p>1 mark – the correct number of points labelled and the endpoints indicated</p> <p>1 mark – correct regression line drawn within the data range.</p> <p>(ii) $r = 0.98102 = 0.981$. Since r is close to 1, there is a strong positive linear correlation between the total distance run per week and the amount of weight loss. As the total distance run per week increases, the amount of weight loss increases.</p> <p>(iii) $y = 3.0141x - 1.49497$ $a = 3.014, b = -1.495$ <u>The meaning of a</u> For every 1 km increase in the total distance run per week, there is a weight of loss of 3.014 kg.</p> <p>(iv) When $x = 0.7(7) = 4.9$ km per week, $y = 3.0141(4.9) - 1.49497 = 13.3$ kg. There is an estimate of 13.3 kg of weight loss. Although r is close to 1, $x = 4.9$ km is outside the data range $[0.6 - 3.6]$, the linear correlation may not hold, the estimate obtained by extrapolation is not reliable.</p> <p>(v) There is no change in the value of product moment correlation coefficient.</p>
--	--

6. CJC H1 Prelim 2017/Q10

	A researcher investigates the relationship between the Gross Domestic Product (GDP), \$ d , in billions of dollars, and population, p , in millions. The historical data is shown in the table below.								



(iii) $p = 0.00538d + 3.60$

(iv) $r = 0.953$

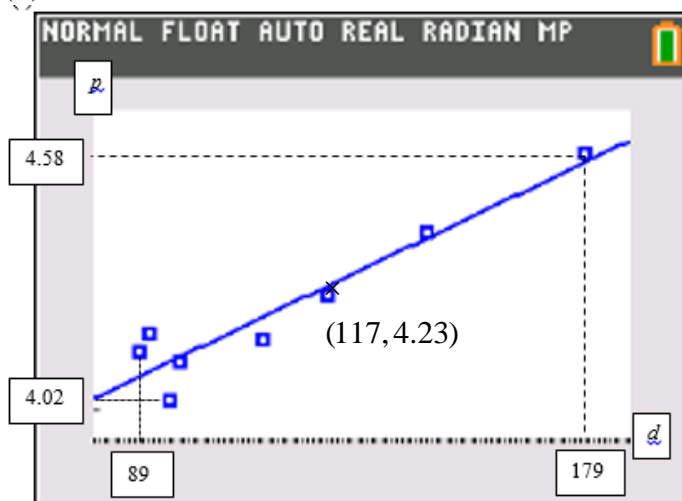
(v) 5.17 million

(vi) \$129 billion

Solution:

(i) $\bar{d} = 117.375$, $\bar{p} = 4.22875$ (exact)

(ii)



(iii)

$$p = 0.00538d + 3.60$$

Sketch on scatter diagram

(iv)

$$r = 0.953$$

Since r is close to 1, there is a strong and positive linear relationship between the GDP and population. As GDP increases, population increases.

(v)

$$d = 292, p = 5.17 \text{ million (3 s.f.)}$$

The estimate is not reliable since it is an extrapolation.

(vi)	Regression line of d on p : $d = 168.7564739p - 596.2539388$ When $p = 4.3$, $d = \$129$ billions (3 s.f.) The estimate is reliable since it is an interpolation and the r -value is close to 1.
------	---

7. ACJC H1 Prelim 2017/Q12

Seven primary school boys took the standing board jump test. The weight, w kg, of each boy and the distance he jumped on the test, x cm, are given in the table below.

w (kg)	42	30	34	37	40	55	45
x (cm)	138	157	158	152	148	126	136

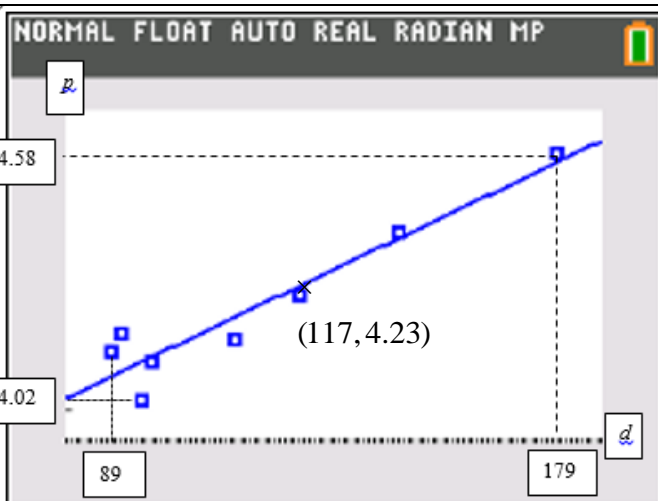
- (i) Give a sketch of the scatter diagram for the data, as shown on your calculator. [2]
(ii) Find the product moment correlation coefficient and comment on its value in the context of the question. [2]
(iii) Find the equation of the least squares regression line of x on w in the form $x = a + bw$, leaving the values of a and b correct to 5 significant figures.
Give an interpretation of the value of b in the context of the question. [2]
(iv) Use the equation of your least squares regression line to calculate an estimate for the standing board jump distance of a boy who weighs
(a) 35 kg,
(b) 15 kg.
Comment on the reliability of your answers. [3]
(v) Aaron also took the test, but it was found that his standing broad jump result was not recorded. After including his weight and the distance he jumped on the test, a new least squares regression line of x on w is calculated to be $x = 202.98 - 1.4249w$.
Given that Aaron weighs 39 kg, find the distance he jumped on the test. [3]

Answers

- (ii) -0.962; (iii) $x = 202.23 - 1.4156w$;
(iv) 153 cm; 181 cm

Solution:

(i) $\bar{d} = 117.375$, $\bar{p} = 4.22875$ (exact)
(ii)



(iii)

$$p = 0.00538d + 3.60$$

Sketch on scatter diagram

(iv)

$$r = 0.953$$

Since r is close to 1, there is a strong and positive linear relationship between the GDP and population. As GDP increases, population increases.

(v)

$$d = 292, p = 5.17 \text{ million (3 s.f.)}$$

The estimate is not reliable since it is an extrapolation.

(vi)

$$\text{Regression line of } d \text{ on } p: d = 168.7564739p - 596.2539388$$

$$\text{When } p = 4.3, d = \$129 \text{ billions (3 s.f.)}$$

The estimate is reliable since it is an interpolation and the r -value is close to 1.

8. IJC H1 Prelim 2017/Q10

Mr Lee recorded the length of time, t minutes, taken to travel to work when leaving home x minutes after 7 am on 10 mornings over two weeks. The results are as follows.

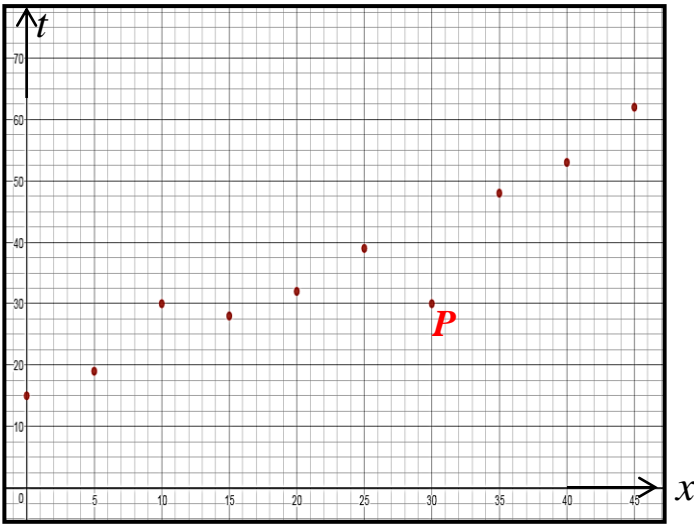
x	0	5	10	15	20	25	30	35	40	45
t	15	19	30	28	32	39	30	48	53	62

- (i) Plot a scatter diagram on graph paper for this data, labelling the axes, using a scale of 2 cm to represent 10 minutes on the t -axis and an appropriate scale for the x -axis. [2]
- (ii) Suggest a reason why one of the data points does not seem to follow the trend and indicate the corresponding point on your diagram by labelling it P . [2]

Omit the point P .

	<p>(iii) Calculate the product moment correlation coefficient and comment on this value. [2]</p> <p>(iv) Find the equation of the least squares regression line of t on x, writing your answer in the form $t = ax + b$. [1]</p> <p>(v) Sketch the regression line on your scatter diagram and interpret the meaning of the value of a in the context of the question. [2]</p> <p>(vi) Mr Lee needs to arrive at work no later than 8.30 am. Estimate, to the nearest minute, the latest time that he has to leave home without arriving late at work. [3]</p>
	<p style="text-align: right;">Answers</p> <p style="text-align: right;">(iii) 0.987</p> <p style="text-align: right;">(iv) $t = 0.978x + 15.025$</p> <p>(v) $a = 0.978$ means that for every additional minute that Mr Lee delays in leaving home after 7am, his travelling time will increase by 0.978 minutes.</p> <p style="text-align: right;">(vi) 7.37 am</p>

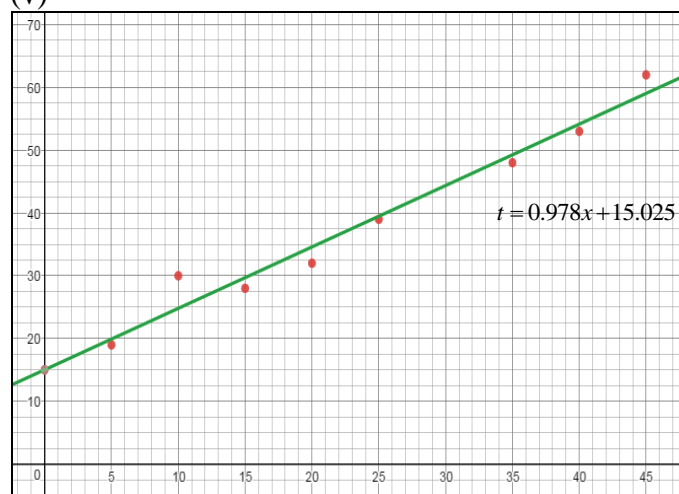
Solution:

10	<p>(i)</p>  <p>(ii)</p> <p>Acceptable reasons:</p> <p>The traffic condition on the road was good (Lesser cars on the road, no traffic jam) and thus he required much shorter travelling time though he left home only at 7.30am.</p> <p>It was a public holiday/school holiday/Sunday and yet Mr Lee has to work.</p> <p>(iii)</p> <p>$r \approx 0.987$</p> <p>The pmcc is close to 1, indicating a strong positive linear correlation between x and t. I.e. the later Mr Lee leaves home</p> <div data-bbox="1129 1861 1364 2040" style="border: 1px solid black; padding: 5px; margin-top: 10px;"> <p style="text-align: right; font-size: small;">NORMAL FLOAT AUTO REAL RA</p> <p style="text-align: right; font-weight: bold; font-size: small;">LinReg</p> <p>$y = ax + b$</p> <p>$a = .9783333333$</p> <p>$b = 15.025$</p> <p>$r^2 = .9749009733$</p> <p>$r = .9873707375$</p> </div>
----	--

after 7 am, the longer the travelling time would take.

(iv) $t = 0.978x + 15.025$

(v)



$a = 0.978$ means that for every additional minute that Mr Lee delays in leaving home after 7am, his travelling time will increase by 0.978 minutes.

(vi)

Method 1:

There are 90 minutes from 7 am to 8.30 am.

$$x + t \leq 90$$

$$x + (0.97833x + 15.025) \leq 90$$

$$1.97833x \leq 74.975$$

$$x \leq 37.898$$

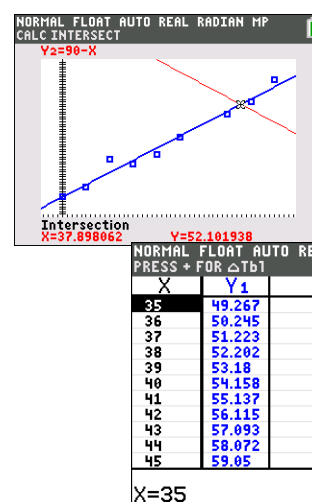
The largest possible value of x is 37 (correct to the nearest minute)

The latest time Mr Lee could leave home without being late for work is 7.37 am.

Method 2:

Sketch the line $x + t = 90$ and find x -coordinate of the point of intersection with the regression line.

If $x = 38$, Mr Lee will arrive late for work. Thus the latest time he needs to leave home is 7.37 am.



Method 3:

By Trial & Error, using GC

From part (vi),

if $x = 40, t = 54.158, x + t > 90$

	$x = 39, t = 53.18, x + t > 90$ $x = 38, t = 52.202, x + t > 90$ $x = 37, t = 51.223, x + t < 90$ Thus the latest time he needs to leave home is 7.37am.
--	---

Summary

1. A **scatter diagram** is useful in showing the general patterns and relationship (which may be linear or non-linear) between two variables. It is also useful for identifying outliers in the data set.
To interpret scatter diagrams, we can comment on three things: type of relationship, strength of relationship, and direction of correlation.
2. The (Linear) Product-moment correlation coefficient, r , indicates the **strength of linear correlation** between 2 variables.
 - If $0 < r < 1$, there is a **positive** linear correlation, i.e. as x increases, y also increases.
 - If $-1 < r < 0$, there is a **negative** linear correlation, i.e. as x increases, y decreases.
 - If $r = 1$, there is a **perfect positive** linear correlation.
 - If $r = -1$, there is a **perfect negative** linear correlation.
 - If $r \approx 1$, there is a **strong** positive linear correlation.
 - If $r \approx -1$, there is a **strong** negative linear correlation.
 - $r = 0$ indicates **no** linear correlation. But you can plot a **scatter diagram** to see whether there is a non-linear relationship.
 - r does NOT provide information about causal relationship (or cause and effect) between variables. For example, if $r \approx 1$, we cannot say that an increase in x causes y to increase.
 - r is NOT affected by a linear transformation of x or y and it has no units.
 - The value of r should be considered in conjunction with a **scatter diagram**.
3. Use of Least Squares Regression Lines
 - **Interpolation**, i.e. estimating one value when given the other **within** the range of the given x and y values.
 - **Extrapolation**, i.e. estimating one value when given the other **outside** the range of the given x and y values.

Note: For extrapolation, the estimated value of x or y may not be reliable as the given value of y or x is outside the range of the given y or x values.
4. (Estimated) Least squares regression lines

- The **least squares regression line of y on x** (x is the independent variable) is given by $y = a + bx$
- The **least squares regression line of x on y** (y is the independent variable) is given by $x = c + dy$

Common point is (\bar{x}, \bar{y}) . Note: $\bar{x} = \frac{\sum x}{n}$ and $\bar{y} = \frac{\sum y}{n}$ for n pairs of sample data.
By solving the 2 equations simultaneously, you can obtain (\bar{x}, \bar{y}) .

5. Which type of regression line to use?

x	y	Purpose	Equation of Regression Line	Remarks
Independent	Dependent	Predict y given x	y on x	
		Predict x given y	x on y	
Random	Random	Predict y given x	y on x	Treat x as the independent variable
		Predict x given y	x on y i.e. $x = dy + e$	Treat y as the independent variable

Checklist

I am able to:

- ☐ understand that the study of correlation is the study of the relation between two variables.
- ☐ understand that regression analysis is a method by which we predict or estimate unknown values given a finite set of data.
- ☐ use the graphic calculator for a given data set
 - ☐ to draw the scatter diagram
 - ☐ to compute the value of the product-moment correlation coefficient
 - ☐ to compute the equation of the least squares regression line of y on x
 - ☐ to compute the equation of the least squares regression line of x on y
- ☐ use the appropriate least squares regression line to predict or estimate unknown values

Appendix 1: Supplementary Notes

1. Valid conclusion on the relationship between two variables can only be derived based on both value of r and scatter diagram (see Anscombe's quartet).
2. Without a scatter plot, when asked to describe the relationship between two variables, avoid using definite terms. Instead we should use "there may be or may not be linear correlation..."
3. While deciding whether a variable is controlled, one must study carefully the method of data collection as described in the question.

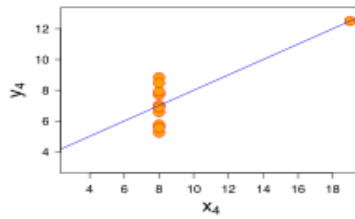
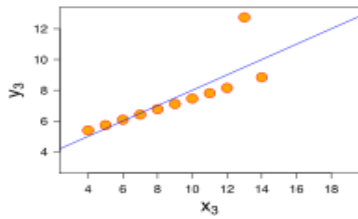
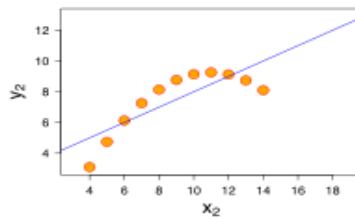
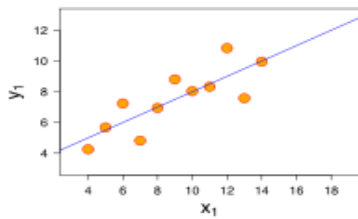
Context	x	y	Purpose	Equation of Regression Line	Remarks
Non-experimental	Independent	Dependent	Predict y given x	y on x	
			Predict x given y	y on x	
	Random	Random	Predict y given x	y on x	Treat x as the independent variable
			Predict x given y	x on y i.e. $x = dy + e$	Treat y as the independent variable

Anscombe's quartet comprises four [datasets](#) that have identical simple statistical properties, yet appear very different when the scatter plot is being graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the [statistician Francis Anscombe](#) to demonstrate both the importance of graphing data before analysing it and the effect of [outliers](#) on statistical properties. Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

For all four datasets:

Property	Value
<u>Mean</u> of x in each case	9 (exact)
<u>Variance</u> of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Variance of y in each case	4.122 or 4.127 (to 3 decimal places)
<u>Correlation</u> between x and y in each case	0.816 (to 3 decimal places)
<u>Linear regression</u> line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)



Appendix 2: Relationship between Regression Coefficient and r

The regression line y on x is $y - \bar{y} = b(x - \bar{x})$

b = the regression coefficient of y on x (note: the gradient of the regression line y on x)

The regression line x on y is $x - \bar{x} = d(y - \bar{y})$

d = the regression coefficient of x on y ,

(note: $\frac{1}{d}$ is the gradient of the regression line x on y)

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum (x - \bar{x})^2\right] \left[\sum (y - \bar{y})^2\right]}} \text{ and } b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}, d = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2}$$

$$bd = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \times \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{\left(\sum (x - \bar{x})(y - \bar{y})\right)^2}{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2} = \left(\frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \right)^2 = r^2$$

Therefore, $\boxed{r^2 = bd \Leftrightarrow r = \pm \sqrt{bd}}$.

As $\sum (x - \bar{x})^2 \geq 0$ and $\sum (y - \bar{y})^2 \geq 0$, r take the sign of $\sum (x - \bar{x})(y - \bar{y})$.

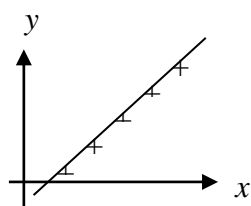
Hence, $r = \sqrt{bd}$ if both b and d are positive, and
 $r = -\sqrt{bd}$ if both b and d are negative.

In particular,

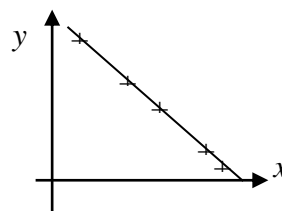
(i) If $r = \pm 1$ (ie. $r^2 = 1$) $\Rightarrow bd = 1 \Rightarrow b = \frac{1}{d}$

Thus the two regression lines have the same gradient.

Hence the two regression lines coincide since they both pass through the point (\bar{x}, \bar{y}) .

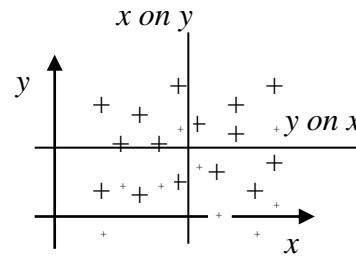


Perfect positive correlation,
the **two regressions line**
coincide . $r = 1$



Perfect negative correlation, the
two regressions line coincide .
 $r = -1$

- (ii) If $r = 0$
 $\Rightarrow \sum (x - \bar{x})(y - \bar{y}) = 0$
 $\Rightarrow b = 0$ and $\frac{1}{d} \rightarrow \infty$
 Hence the two regression lines are perpendicular to each other.



No correlation $r = 0$

Example:

For a set of bivariate data, the regression line of y on x is $y = -1.96x + 15$, and the regression line of x on y is $y = -2.22x + 15.91$. Find the product moment correlation coefficient.

Solution:

The regression line of y on x is $y = -1.96x + 15 \Rightarrow b = -1.96$

The regression line of x on y is $y = -2.22x + 15.91 \Rightarrow d = -\frac{1}{2.22}$

Therefore $r = -\sqrt{bd} = -0.94(2d.p.)$