

18

CORRELATION AND LINEAR REGRESSION

SYLLABUS

- Concepts of scatter diagram, correlation coefficient and linear regression
 - Understand that bivariate data consists of the values of 2 variables (independent and dependent variables) obtained from the same sample, expressed as ordered pairs;
 - Use a graphic calculator to plot the scatter diagram for a set of bivariate data to determine if there is a linear relationship between the 2 variables;
 - Understand that the correlation coefficient is a measure of the fit of a scatter diagram to a linear model
- Calculation and interpretation of the product moment correlation coefficient and of the equation of the least squares regression line
 - 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;
 - Understand that a high correlation between 2 variables does not necessarily imply one directly causes the other;
 - Understand the concepts of linear regression and 'least squares' with reference to the scatter diagram;
 - Use a graphic calculator to find the equation of the least squares regression line, and interpret its slope and intercept (A different line of regression will be obtained if we interchange the independent and dependent variables)
- Interpolation and extrapolation

- 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
- Use of square, reciprocal or log transformation to achieve linearity.
- Use an appropriate transformation to linearise a set $y = a + bx$ of bivariate data to fit the regression model.

CONTENTS

1	Introduction	3
2	Bivariate Samples	3
3	Independent and Dependent Variables	3
4	Scatter Diagrams	4
5	Linear Correlation	7
5.1	Calculating the value of Product Moment Correlation Coefficient, r	7
5.2	Properties of r	9
6	Linear Regression	14
6.1	Least Squares Regression Line of y on x	15
6.2	Least Squares Regression Line of x on y	16
7	Estimation Using Least Squares Regression Lines	21
8	Linear Transformation	29
9	Miscellaneous Examples	32
Annex A:	Gradient of Least Squares Regression Line	38
Annex B:	Relationship between Correlation Coefficient and Least Squares Regression Lines	39

LECTURE 1

Lesson Outline

- Introduction to bivariate samples with independent and dependent variables
 - Plot the scatter diagram using a graphing calculator
 - Calculate and interpret the linear product moment correlation coefficient
-

1 INTRODUCTION

Previous chapters have focused on developing methods and models for a single random variable. However, there are many data sets that provide information about two variables and we are keen to find out the relationship between these variables, such as petrol consumption of a car and the speed at which it is driven, height of father and height of son when 18, or a student's grade in mock examination and in national examination.

In this chapter, we will study the correlation between two variables and in particular, to investigate whether there is linear relationship between two variables, which will then allow for possible predictions.

2 BIVARIATE SAMPLES

Bivariate samples consist of members which have measurements of two different characteristics available. For example, we measure the heights and the weights of students in a particular class in ACJC. This type of data is called bivariate (as opposed to univariate) data.

3 INDEPENDENT AND DEPENDENT VARIABLES

Independent variables and dependent variables refer to values that change in relationship to each other. The **independent** variables (controlled variables) are those that are deliberately manipulated to invoke a change in the dependent variables. The **dependent** variables (response variables) are those that are observed to change in response to the independent variables.

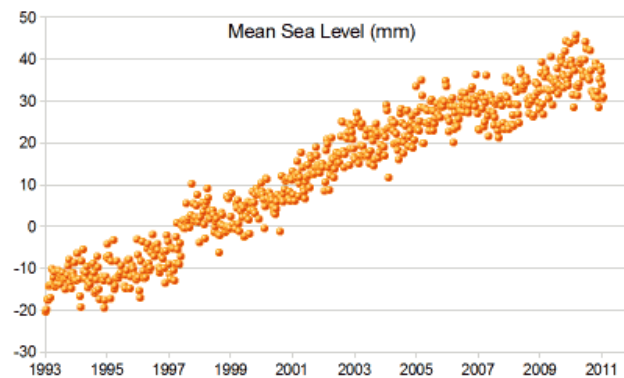
Independent variable, x	Dependent variable, y
Number of people in a lift.	Total weight of people in the lift.
Amount of fertilizer used per unit field area.	Weight of crop yield per unit field area.
Hours of sleep the night before an examination.	Examination score.

4 SCATTER DIAGRAMS

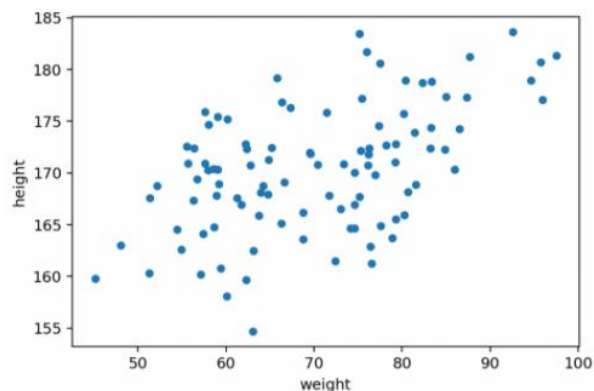
Observations of variables X and Y from a bivariate sample can be considered as ordered pairs or coordinate pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. These can then be plotted against axes that represent the characteristics measured. This diagram is called a **scatter diagram**.

Conventionally, the independent variable is plotted on the horizontal axis while the dependent variable is plotted on the vertical axis.

Real life examples of scatter diagrams:



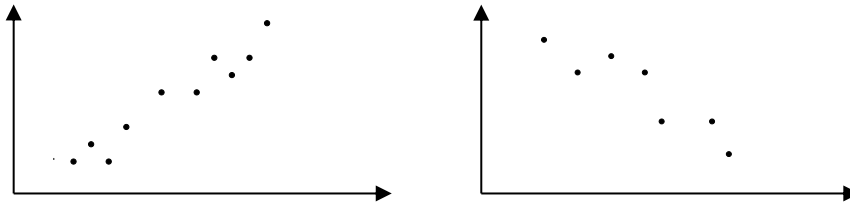
Source: <https://www.mathsisfun.com/data/scatter-xy-plots.html>



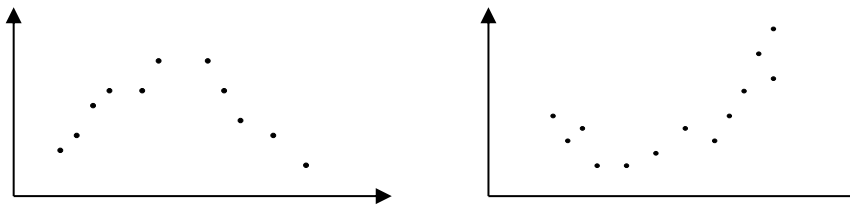
Source: <https://data36.com/scatter-plot-pandas-matplotlib/>

Possible scatter diagrams and relationships

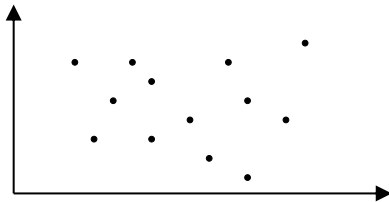
Linear Relationship



Curvilinear Relationship



No Relationship



The scatter diagram provides an overview of the relationship between variables. Often, given a data set, we draw the scatter diagram to observe whether there is linear relationship between the variables.

When drawing a scatter diagram,

- the **axes** must be labelled
- the **extreme values** (minimum and maximum values) of each variable on the horizontal and vertical axes must be indicated on the diagram.

The scatter diagram is also useful in identifying **outliers**, which are data points that appear to be quite different from the trend shown by the other data points in the diagram, and can be removed to provide a more accurate picture of the relationship between the variables.

Example 1

The yield (per hectare) of a crop, c , is believed to depend on the May rainfall, m . Records for nine regions are kept for the average values of c and m , and these are recorded below.

m	14.7	10.4	18.8	13.1	14.9	13.8	16.8	11.8	12.2
c	8.3	10.1	15.2	6.4	11.8	12.2	13.4	11.9	9.9

Sketch a scatter diagram for the data.

[N1990/part]

Solution

- 1) Press **STAT** and select **1: Edit...** to access the lists.

```

EDIT  CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor

```

- 2) Enter the data for variables m and c in the existing list names **L1** and **L2** for the independent and dependent variables respectively.

c is the **dependent** variable (y -axis) and m is the **independent** variable (x -axis).

L1	L2	L3	L4	L5	2
14.7	8.3				
10.4	10.1				
18.8	15.2				
13.1	6.4				
14.9	11.8				
13.8	12.2				
16.8	13.4				
11.8	11.9				
12.2	9.9				
---	---				
L2(10)=					

- 3) To plot, press **[2ND] [STATPLOTS]**, turning on **1:Plot1**.

```

STAT PLOTS
1:Plot1...Off
  L1 L2
2:Plot2...Off
  L1 L2
3:Plot3...Off
  L1 L2
4:PlotsOff
5:PlotsOn

```

- 4) Press **[2ND] [1]**, **[2ND] [2]** to key in **L1**, **L2** for the respective **Xlist** and **Ylist**.

```

Plot1 Plot2 Plot3
On Off
Type: [ ] [ ] [ ] [ ] [ ]
Xlist:L1
Ylist:L2
Mark: [ ] + * .
Color: BLUE

```

- 5) To view the scatter diagram, press **ZOOM** and select **9:ZoomStat**.

```

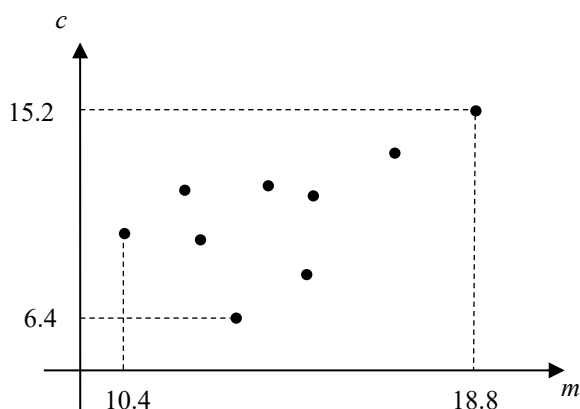
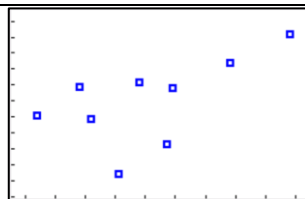
ZOOM MEMORY
1:ZBox
2:Zoom In
3:Zoom Out
4:ZDecimal
5:ZSquare
6:ZStandard
7:ZTrig
8:ZInteger
9:ZoomStat

```

- 6) A scatter diagram with data points is obtained.

To read the coordinates of each point on the scatter diagram, use **TRACE**.

Copy the scatter diagram from the GC, **label both axes**, and the **extreme x and y values**.



■

Note

- From the scatter diagram, there appears to be some relationship between the variables c and m . As m increases, c increases. There seems to be some linear relationship between the variables.

5 LINEAR CORRELATION

Linear correlation is a measure of the linear relationship between two random variables X and Y . The **correlation coefficient** is a numerical measure of the strength of the linear relationship between two variables.

5.1 Calculating the value of the Product Moment Correlation Coefficient

To measure the linear correlation between two variables, we use **Pearson's product moment correlation coefficient**, r , given by (in MF27)

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

Example 1 (cont'd)

The yield (per hectare) of a crop, c , is believed to depend on the May rainfall, m . Records for nine regions are kept for the average values of c and m , and these are recorded below.

m	14.7	10.4	18.8	13.1	14.9	13.8	16.8	11.8	12.2
c	8.3	10.1	15.2	6.4	11.8	12.2	13.4	11.9	9.9

Find the product moment correlation coefficient between c and m .
[N1990/part]

Solution

- 1) Press **STAT** and select **1: Edit...**

```

EDIT CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor

```

- 2) Enter the data for variables m and c in the existing list names **L1** and **L2** for the independent and dependent variables respectively.

L1	L2	L3	L4	L5	2
14.7	8.3				
10.4	10.1				
18.8	15.2				
13.1	6.4				
14.9	11.8				
13.8	12.2				
16.8	13.4				
11.8	11.9				
12.2	9.9				

L2(10)=

- 3) Press **[STAT]**, **[CALC]** and select **8:LinReg(a+bx)**.

```

EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9:LnReg

```

- 4) Enter **L1** for **Xlist** and **L2** for **Ylist**. Press **[ENTER]** to get the value of r .

```

LinReg(a+bx)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:
Calculate

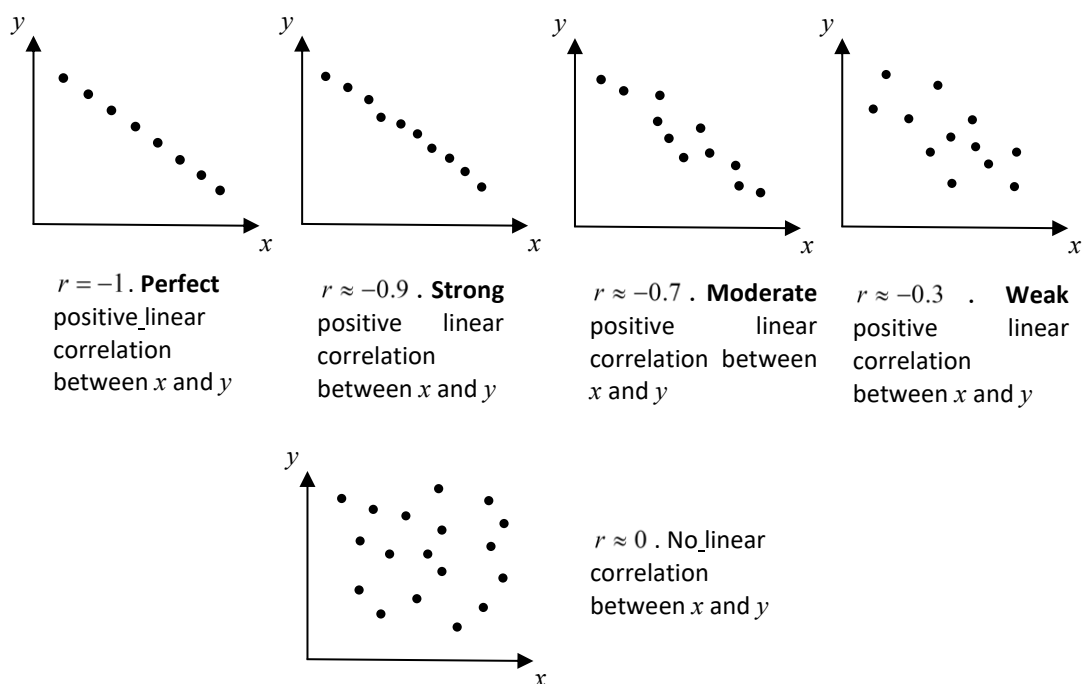
```

If you do not see the r value shown on this screen, press **MODE** and check whether the **Stats Diagnostics On** is turned on.

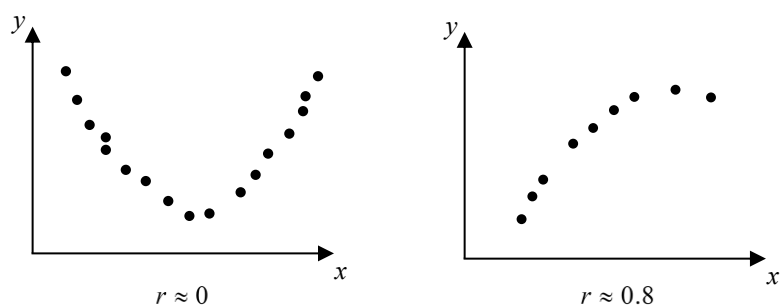
```

LinReg
y=a+bx
a=2.481329413
b=0.6076524528
r^2=0.3515890416
r=0.5929494427

```

- (i) If $r = \pm 1$, this is called a **perfect linear correlation**. All data points lie exactly on a straight line. Note that the gradient of the regression lines may not be 1 or -1 .
- (ii) The closer the value of r to ± 1 , the stronger the linear correlation between the two variables i.e. the closer the points on the scatter diagram are to a straight line.
- (e) A scatter diagram should be drawn to establish the linear relationship between the variables before r is calculated.
- (i) The low value of r (computed from the formula) seems to indicate a weak linear correlation between the variables. However, the scatter diagram (below left) shows a curvilinear relationship between the variables i.e. no linear relationship between X and Y :



- (ii) The value of r is close to 1. However the scatter diagram shows a non-linear relationship between the variables X and Y .

- (f) The existence of correlation between two variables does not imply cause and effect. It is possible that the two variables are each related in turn to another variable which produces the changes in both.

Example 2 (Correlation does not imply Causation)

Consider two variables x and y where x is the height of a person and y is the salary of a person. A scatter diagram of data on these two variables shows a positive association with a positive value for the correlation coefficient.

Does your height cause you to earn a certain salary level?

Suppose you learn that the ages of the people sampled range from 12 years old to 30 years old.

Give a possible explanation of what is going on here.

Solution

A positive relationship between height and salary only indicates that as height of a person increases, the salary increases. However, it does not show a causal relationship between height and salary i.e. it does not imply that an increase in height causes an increase in salary.

The relationship between height and salary could be due to a third variable, age of a person. Older age group is associated with greater height and higher salary, hence the positive relationship between height and salary.

■

Note

We say that the positive association between height and salary is due to age, the common response, i.e. a third variable that is common to both variables. The common response, age, created an association even though there is no direct causal link between height and salary of a person.

In other words, a correlation between two quantitative variables does not imply that the value of one variable causes the other variable to take on a certain value. The apparent association can be due to a third variable through common response. (Interactive Statistics, 3rd Edition)

LECTURE 2

Lesson Outline

- Understand the concept of equations of least squares regression lines.
- Finding regression lines.

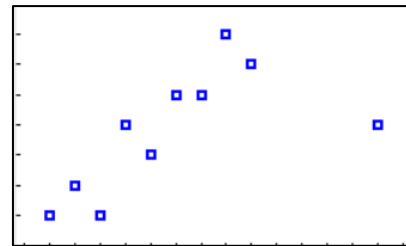
Example 3 (Correlation is sensitive to Outliers)

Consider the following data and its scatter diagram.

x	2	3	4	5	6	7	8	9	10	15
y	3	4	3	6	5	7	7	9	8	6

The point A in the scatterplot ($x = 15, y = 6$) is an observation that stands apart from the remaining points. It is called an **outlier**.

- Calculate the correlation coefficient for these data **with** point A .
- Calculate the correlation coefficient for these data **without** point A .
- Did r change? How did it change?



(Adapted from Interactive Statistics, 3rd Edition)

Solution

- The correlation coefficient for these data with point A is $r = 0.648$.
- The correlation coefficient for these data without point A is $r = 0.927$.
- The correlation coefficient did change from a _____ positive association to a _____ positive association. ■

We have learnt that the correlation coefficient can change dramatically when an outlier is removed.

Example 4 (Possible non-linear relationship)

The rate of growth y of a particular organism is thought to depend in some way on the temperature t . The table shows the results of 10 experiments.

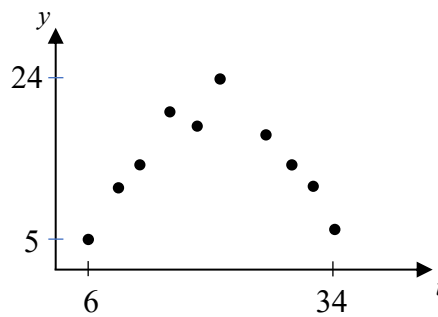
Temperature t (°C)	6	9	11	14	17	22	26	29	32	34
Rate of growth y (arbitrary units)	5	13	15	21	20	24	19	16	10	7

Calculate the value of the linear product moment correlation coefficient. Represent the data on a suitable diagram and comment briefly upon the statement “The linear correlation coefficient does not differ significantly from zero and thus y does not depend in any way upon t ”.

Solution

From GC, $r = -0.0177$ (3 s.f.)

The statement is not necessarily true as $r \approx 0$ only indicates that the rate of growth and temperature are not linearly correlated, but as shown in the scatter diagram, the two variables appear to have a curvilinear relationship.

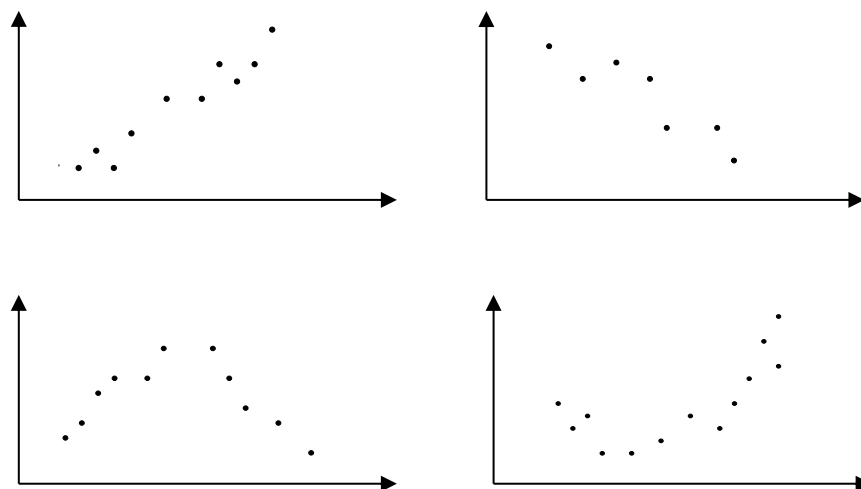


■

6 LINEAR REGRESSION

From a bivariate sample we look for a relationship of the form $y = f(x)$, i.e. given the points on the scatter diagram we work backwards or ‘regress’ to obtain the original relationship. This function is called a **regression function**.

The relationship between the dependent and independent variables would be linear and of the form $y = mx + c$.



While in above 2 graphs, the 2 variables could be related by equations such as $y = ax^2 + bx + c$ or $y = a \sin bx$.

We shall only consider the simplest type of relationship between x and y , the **linear relationship**, in which most or all the points in the scatter diagram seem to lie near or on a straight line.

Linear regression is a way to model the relationship between two variables by fitting a **linear equation** to a set of observed data.

Given a set of observed data, we first draw a scatter diagram to determine whether it is appropriate to model the observed data with a straight line. When the scatter diagram suggests a linear correlation between the two variables, we want to draw a line that best fits the given data such that the line is ‘as close as possible’ to the data points. Such a line is called a regression line.

This regression line can then be used to predict behaviour of one of the variables as the other variable changes in value.

There are two types of regression lines that can be drawn.

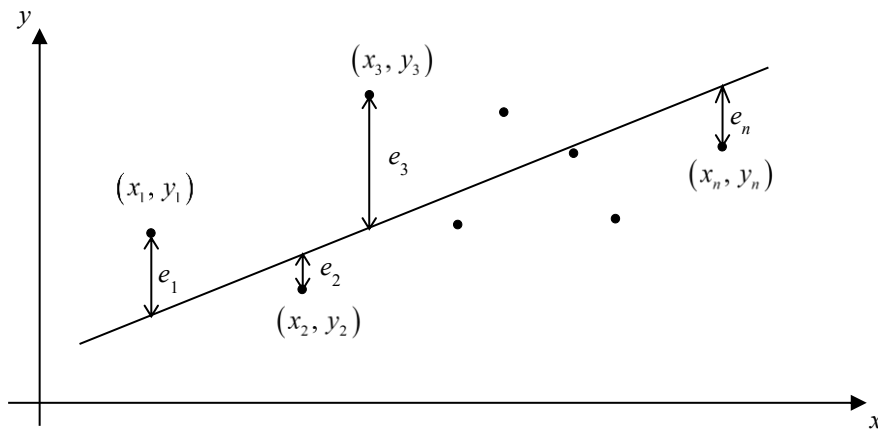
- A **regression line of y on x** is obtained by taking the **x values as accurate** and considering the error in the y values whereby x is the independent variable and y is the dependent variable.
- A **regression line of x on y** is obtained by taking the **y values as accurate** and considering the error in the x values whereby y is the independent variable and x is the dependent variable.

6.1 Least Squares Regression Line of y on x

Consider the equation of a line $y = a + bx$.

We assume the values of x to be **accurate** and consider the deviation of the points (x_i, y_i) from the line in the y direction as **residuals** (or ‘errors’), and represent them as $e_1, e_2, e_3, \dots, e_n$, i.e.

$$\begin{aligned} e_i &= \text{difference between the observed value } y_i \text{ and the predicted value} \\ &\quad \text{given by the line } y = a + bx \text{ at the point } x_i \\ &= |y_i - (a + bx_i)| \end{aligned}$$



The **least squares regression line of y on x** is obtained by minimising

$$\sum_{i=1}^n e_i^2.$$

Estimated least squares regression line of y on x is given by

$$y = a + bx \quad \text{---(1)}$$

It can be shown that this regression line passes through the point $(\bar{x}, \bar{y})$,
so

$$\bar{y} = a + b\bar{x} \quad \text{---(2)}$$

From (1) and (2), the estimated regression line of y on x is

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

where

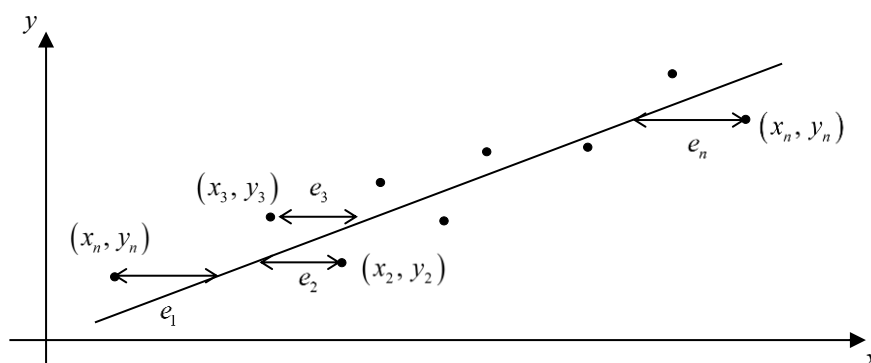
$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\underbrace{\sum (x - \bar{x})^2}_{\text{in MF27}}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}.$$

6.2 Least Squares Regression Line of x on y

Consider the equation of a line $x = c + dy$.

We assume the values of y to be **accurate** and consider the deviation of the points (x_i, y_i) from the line in the x direction as **residuals** (or ‘errors’), and represent them $e_1, e_2, e_3, \dots, e_n$, i.e.

$$\begin{aligned} e_i &= \text{difference between the observed value } x_i \text{ and the predicted value} \\ &\quad \text{given by the line } x = c + dy \text{ at the point } y_i \\ &= |x_i - (c + dy_i)|. \end{aligned}$$



The **least squares regression line of x on y** is obtained by minimising

$$\sum_{i=1}^n e_i^2.$$

Estimated least squares regression line of y on x is given by

$$x = c + dy \quad \text{---(1)}$$

It can be shown that this regression line passes through the point $(\bar{x}, \bar{y})$,
so

$$\bar{x} = c + d\bar{y} \quad \text{---(2)}$$

From (1) and (2), the estimated regression line of y on x is

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

where

$$d = \frac{\underbrace{\sum (x - \bar{x})(y - \bar{y})}_{\text{not in MF27}}}{\sum (y - \bar{y})^2} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}}.$$

Notice that this is just a replacement of x by y and vice versa from the regression line of y on x .

- 3) Press **[2ND] [1]** to enter **L1** (x) for **XList**, and **[2ND] [2]** to enter **L2** (y) for **YList**.
(Optional) You can also store the equation using under **Store RegEQ** by pressing **[ALPHA] [F4]** to select **Y1**.
- 4) Select **Calculate** and press **[ENTER]** to view the equation and the value of r .

```
LinReg(a+bx)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:Y1
Calculate
```

```
LinReg
y=a+bx
a=4.077647059
b=0.1588235294
r^2=0.8437255866
r=0.9185453645
```

The least squares regression line of y on x is

(b) Finding regression line of x on y

- 1) Enter the data if it has not been entered. This was done in (a) in this case.
- 2) Press **[STAT]**, select **CALC** and select **8: LinReg(a+bx)**. You may also select **4: LinReg(ax+b)**.
- 3) Press **[2ND] [2]** to enter **L2** (y) for **XList**, and **[2ND] [1]** to enter **L1** (x) for **YList**.
You cannot store the equation using under **Store RegEQ** for x on y as x is expressed in terms of y .
- 4) Select **Calculate** and press **[ENTER]** to view the equation and the value of r .

```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9:LnReg
```

```
LinReg(a+bx)
Xlist:L2
Ylist:L1
FreqList:
Store RegEQ:
Calculate
```

```
LinReg
y=a+bx
a=-19.31775701
b=5.312346286
r^2=0.8437255866
r=0.9185453645
```

Since the **L2** (y) is used as the **XList** and **L1** (x) is used as **YList**, the equation obtained is actually of the form $x = a + by$.

The least squares regression line of x on y is

You can make y the subject if needed, especially if you are required to plot both regression lines on the same diagram.

(c) The product moment correlation coefficient, $r =$

(d) Finding $\bar{x}$ on $\bar{y}$

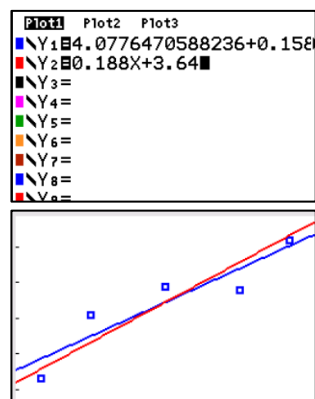
1) Press [STAT] then select CALC and 2:2-Var Stats .	EDIT CALC TESTS 1:1-Var Stats 2:2-Var Stats 3:Med-Med 4:LinReg(ax+b) 5:QuadReg 6:CubicReg 7:QuartReg 8:LinReg(a+bx) 9:LnReg
2) Enter L1 and L2 into XList and YList respectively.	2-Var Stats Xlist:L1 Ylist:L2 FreqList: Calculate
3) Select Calculate and press [ENTER] .	2-Var Stats $\bar{x}=15$ $\Sigma x=75$ $\Sigma x^2=1397$ $Sx=8.246211251$ $\sigma x=7.375635566$ $n=5$ $\bar{y}=6.46$ $\downarrow \Sigma y=32.3$

From GC, $\bar{x} =$ and $\bar{y} =$ ■

Sketching regression lines

When the equation of regression line of y on x is stored in **Y1**, press **ZOOM**, then select **9:ZoomStat** to view the regression line on the scatter diagram.

To sketch the regression line of x on y on the y against x axis, you will need to express the equation by making y the subject and then key into **Y2**.



Example 6

The following data gives the regression line of y on x as $y = 2x - 2$.

x	3	2	3	6	1
y	p	2	6	q	0

- (i) Find $\bar{y}$.
- (ii) Find the value of $\sum xy$.
- (iii) Find the values of p and q .

Solution

- (i) Using GC, $\bar{x} = 3$.

Since the regression line passes through $(\bar{x}, \bar{y})$,

- (ii) Using GC, $\sum x = 15$ and $\sum x^2 = 59$.

Since gradient of regression line of y on x is 2,

- (iii) From (i),

From (ii),

■

LECTURE 3

Lesson Outline

- Finding estimation using least squares regression lines (example 7–9)
 - Understand the use of appropriate transformation to achieve linearity
-

7 ESTIMATION USING LEAST SQUARES REGRESSION LINES

Determine the appropriate regression line

Case 1: One of the variables is an **independent** variable.

- When x is an independent variable, **regression line of y on x** is used for any estimation.
- When y is an independent variable, **regression line of x on y** is used for any estimation.

Case 2: Neither variable can be considered independent.

- **Regression line of y on x** is used to estimate the value of y for a given value of x .
- **Regression line of x on y** is used to estimate the value of x for a given value of y .

If the linear correlation between x and y is very strong ie. r is close to 1 or -1 , and a given x value is within the data range, we can use the least squares regression line of x on y to predict the value of y or the regression line of y on x to predict the value of x for a given y value within the data range.

Reliability of Estimates

Interpolation involves making a prediction *within* the ranges of values for the data.

Extrapolation involves making a prediction *outside* the range of values for the data.

An estimate is **reliable** if both the following conditions are satisfied:

- (a) We are estimating within the data range;
- (b) The value of the product moment correlation coefficient is close to 1 or -1 . i.e. there is a strong linear correlation between the variables (This may also be reflected in the scatter diagram, that the data points indicate a strong linear correlation).

An estimate is **unreliable** as long as either one of the above conditions is not satisfied. For example, the estimated value of y will not be reliable if the x values used are outside the range of the data (i.e. extrapolation done) as the relationship may not continue to be linear.

Presence of Outliers

An outlier is an observation that is very different from the rest of the data. It could also be a point that is far from the regression line.

Outliers may significantly influence the value of r and the equation of the line of regression. You may need to identify and remove outliers before calculating r and finding the equation of the line of regression.

Example 7A (With dependent and independent variables)

The following table shows the data on the number of hours spent playing computer games in a week, x and the test scores for maths, y for six students.

No. of hours spent playing computer games in a week (x)	15	18	13	16	12	22
Test score for Math, max of 20 marks (y)	8	3	9	6	14	2

Given that the least squares regression line of y on x is

$$y = 24.69697 - 1.106060x,$$

and the least squares regression line of x on y is

$$x = 21.32292 - 0.76042y.$$

- (a) Estimate the number of hours spent playing computer games in a week given that the test score for Math is 10.
- (b) Estimate the test score for Math given that the number of hours spent playing computer games in a week is 20.

Solution

Independent variable (Horizontal axis):

Dependent variable (Vertical axis):

A suitable least squares regression line is

When independent and dependent variable can be determined, use the regression line of "Dependent" on "Independent" to do any estimation.

- (a) When $y = 10$,

$$10 = 24.69697 - 1.106060x \Rightarrow x = 13.2877$$

The estimated number of hours spent playing computer games in a week is 13.3 hours.

- (b) When $x = 20$,

$$y = 24.69697 - 1.106060(20) = 2.57577$$

The estimated test score for Math is 2.58. ■

Example 7B (Without dependent and independent variables)

Test score for GP, max of 30 marks (x)	15	18	13	16	12	22
Test score for Math, max of 20 marks (y)	8	3	9	6	14	2

Given that the least squares regression line of y on x is

$$y = 24.69697 - 1.106060x$$

and the least squares regression line of x on y is

$$x = 21.32292 - 0.76042y.$$

- (a) Estimate the test score for GP given that the test score for Math is 10.
- (b) Estimate the test score for Math given that the test score for GP is 20.

Solution

When independent and dependent variables cannot be determined, it is not possible to determine a suitable least squares regression line.

- (a) Since y is given, use the regression line

$$\text{When } y = 10, x = 21.32292 - 0.76042(10) = 13.71872$$

The estimated test score for GP is 13.7.

- (b) Since x is given, use the regression line

$$y = 24.69697 - 1.106060x$$

$$\text{When } x = 20, y = 24.69697 - 1.106060(20) = 2.57577$$

The estimated test score for Math is 2.58. ■

When independent and dependent variable cannot be determined, the variable with the "given value" will be the independent variable.

Example 8

The following data pertains to the demand for a product and its price charged in 6 different market areas.

Price (x)	15	18	13	16	12	22
Demand (y)	8	3	9	6	14	2

- Fit a least squares regression line, which enables us to predict the demand for the product in terms of its price.
- Fit a least squares regression line to estimate the dependence of price on demand
- On a scatter diagram, sketch the two regression lines. Find the point of intersection of these two lines. State the significance of this point.
- Use an appropriate regression line to estimate the integer value of x if the value of y is correctly determined as 20. Comment on this value found.

Solution

To find the least squares regression line of y on x and the least squares regression line of x on y .

Step 1a:

Press **STAT** and select **1:Edit...** to create two new lists.

Enter the data in the existing list names **L1** and **L2** for the independent and dependent variables respectively.

```

EDIT  CALC  TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9:LnReg
  
```

Step 1b:

Press **STAT**, **CALC** and select **8: LinReg(a+bx)**.

[You may also select **4: LinReg(ax+b)**]

**Least squares regression
line of y on x**

**Least squares regression
line of x on y**

Step 2:

Press **2ND 1**, **2ND 2** to select **L1**, **L2** and **ALPHA F4** to select **Y1**.

The regression line of y on x is stored in **Y1**.

```

LinReg(a+bx)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:Y1
Calculate
  
```

```

LinReg(a+bx)
Xlist:L2
Ylist:L1
FreqList:
Store RegEQ:
Calculate
  
```

Note that we should interchange the positions of x and y based on the result given in the GC.

Step 3:

Press **ENTER** to view the equation.

LinReg
 $y=a+bx$
 $a=24.6969697$
 $b=-1.106060606$
 $r^2=0.8410669192$
 $r=-0.9170970064$

LinReg
 $y=a+bx$
 $a=21.32291667$
 $b=-0.7604166667$
 $r^2=0.8410669192$
 $r=-0.9170970064$

- (a) The least squares regression line of y on x :
 $y = 24.697 - 1.1061x = 24.7 - 1.11x$ (3 s.f.)
- (b) The least squares regression line of x on y :
 $x = 21.323 - 0.76042y = 21.3 - 0.760y$ (3 s.f.)
- (c) **Sketch least squares regression lines on scatter diagram**
 The least squares regression line of y on x :
 $y = 24.697 - 1.1061x = 24.7 - 1.11x$ (3 s.f.)

The least squares regression line of x on y :
 $x = 21.323 - 0.76042y = 21.3 - 0.760y$ (3 s.f.)

$$\Rightarrow y = \frac{x - 21.3}{-0.760}$$

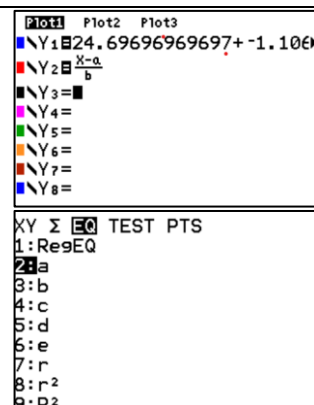
- 1) To sketch the regression line x on y on the same diagram as the regression line y on x , key the equation $\frac{x-a}{b}$ in

Y2.

The values of **a** and **b** can be found in **[VARS]** then select **5: Statistics** followed by **EQ**. These values of **a** and **b** are obtained from the last use of **8: LinReg(a+bx)**.

OR

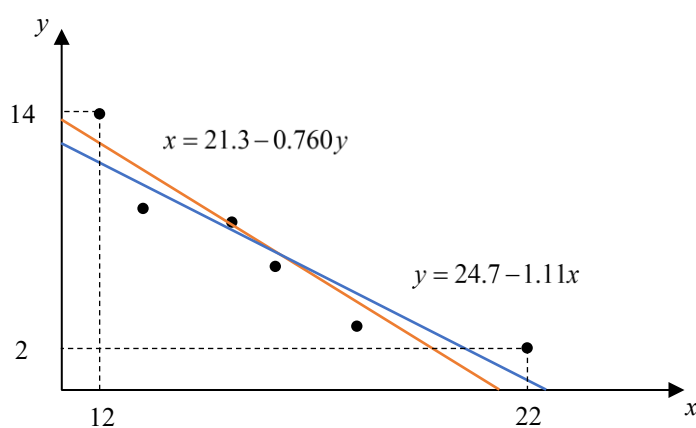
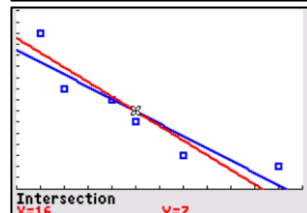
Key in $\frac{x-21.3}{-0.760}$ in **Y2.**



- 2) Turn on **[STAT PLOT]** and use **ZoomStat** to get both regression lines drawn on the scatter diagram.

CALCULATE
 1:value
 2:zero
 3:minimum
 4:maximum
 5:intersect
 6:dy/dx
 7:∫f(x)dx

- 3) Press **[2ND] [CALC]** and select **5: intersect** to find the point of intersection of these two lines as shown.



From GC, the point of intersection is $(16, 7)$. This point is $(\bar{x}, \bar{y})$.

- (d) As y is correctly determined as 20, so y is the independent variable. The regression line of x on y should be used to estimate x .
 Using the regression line of x on y : $x = 21.323 - 0.76042y$
 When $y = 20$, $x = 21.323 - 0.76042(20) = 6.11$ (3 s.f.)

The result should be used with caution as the value $y = 20$ is outside the given range of the sample data, $2 \leq y \leq 14$. The extrapolated value of x may not be reliable, as the relationship may not continue to be linear. ■

Example 9 (Significance of gradient and y-intercept):

An instrument used to measure the levels of glucose in a person's blood is monitored on a random sample of 10 people. The levels are also measured using a very accurate laboratory procedure. The following table gives the instrument measure, y , and the very accurate laboratory procedure measure, x , in suitable units.

x	4.1	5.3	6.3	6.8	7.1	7.5	8.4	8.6	9.5	12.5
y	4.6	5.0	6.6	6.5	7.5	7.8	8.1	9.0	9.2	12.0

- (i) Assuming that there is an approximate linear relationship between x and y , calculate the equation of a suitable least squares regression line, giving a reason for your choice of line.
- (ii) Calculate an estimate for the error in an instrument reading of 5.5.
- (iii) What is the significance of the gradient and y -intercept of the regression line in (i)?

Solution

- (i) We use the regression line of _____ because _____

Regression line of y on x :

- (ii) When $y = 5.5$,
Error in reading = _____

- (iii) Significance of gradient:
From the least squares regression line of y on x , $y = 0.683 + 0.913x$,
we can say that _____

Significance of y -intercept: _____

■

Note

- Treat this value with caution as $x = 0$ is **outside** the range of values of x that was measured (extrapolation).

LECTURE 4

Lesson Outline

- Understand linear law
 - Application of linear law
-

8 LINEAR LAW

In this section, we use a quadratic, reciprocal or logarithmic transformation to achieve linearity. That is, we transform non-linear data points to linear data points. For example,

Equation	Linear Form $Y = mX + c$	Y	X
$y = a + bx$	$y = a + bx$	y	x
$y = a + bx^2$	$y = a + b(x^2)$	y	x^2
$y = a + \frac{b}{x}$	$y = a + b\left(\frac{1}{x}\right)$	y	$\frac{1}{x}$
$y = a + b \ln x$	$y = a + b(\ln x)$	y	$\ln x$
$y = ab^x$	$\ln y = \ln a + x \ln b$	$\ln y$	x

Example 10

Two variables x and y are thought to be related by the equation $y = Ax^B$ where A and B are constants. Eight pairs of observations are given in the following table.

x	2.5	3.1	4.3	5.0	5.9	8.1	7.1	9.4
y	9	13	18	25	35	81	53	132

Determine an appropriate regression model and find the values of A and B .

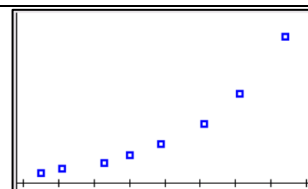
Solution

$$y = Ax^B \Rightarrow \ln y = \ln A + B \ln x$$

Using GC to transform data

- 1) Store given data x in **L1** and y in **L2**.

For illustration purpose - Plot the scatter diagram of the given data. The scatter diagram shows the relationship between y and x to be non-linear. Therefore it is not suitable to find the linear regression line for y on x .



- 2) Move the cursor to the header of the columns to enter the formula.

Key in $\ln(\mathbf{L1})$ into **L3** then press **[ENTER]**.

Key in $\ln(\mathbf{L2})$ into **L4** then press **[ENTER]**.

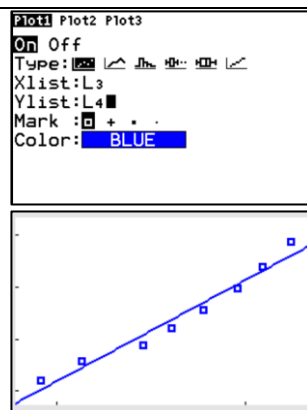
The values of $\ln x$ and $\ln y$ are stored in the lists **L3** and **L4** respectively.

L1	L2	L3	L4	L5	
2.5	9	0.9163	2.1972		
3.1	13	1.1314	2.5649		
4.3	18	1.4586	2.8904		
5	25	1.6094	3.2189		
5.9	35	1.775	3.5553		
8.1	81	2.0919	4.3944		
7.1	53	1.9601	3.9703		
9.4	132	2.2407	4.8828		
-----	-----	-----	-----		
L4=ln(L2)					

- 3) Find the least squares regression line of $\ln y$ on $\ln x$. Press **[ENTER]** to obtain the values of the y -intercept and the gradient of the regression line of $\ln y$ on $\ln x$.

LinReg(a+bx)
Xlist:L3
Ylist:L4
FreqList:
Store RegEQ:Y1
Calculate
LinReg
y=a+bx
a=0.240217873
b=1.95341374
r ² =0.9636608958
r=0.9816623125

- 4) For illustration purpose - Plot the scatter diagram of the new data and note that a linear model of $\ln y$ against $\ln x$ is a good fit for the data.



Example 11

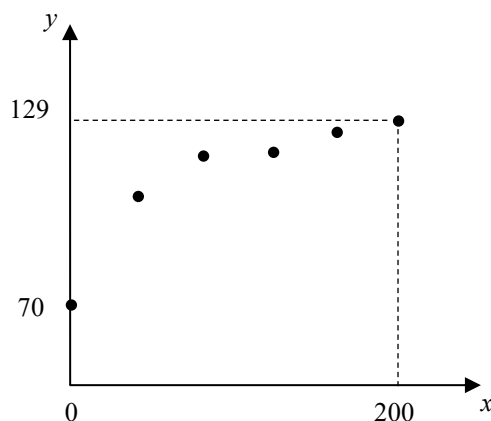
An investigation into the effect of a fertilizer on yields of corn found that the amount of fertilizer applied, x , resulted in the average yields of corn, y , given below, where x and y are measured in suitable units.

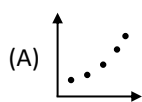
x	0	40	80	120	160	200
y	70	104	118	119	126	129

- (i) Draw a scatter diagram for these values. Explain which of the following equations, where a and b are positive constants, provides the most accurate model of the relationship between x and y .
- (A) $y = ax^2 + b$ (B) $y = \frac{a}{x^2} + b$
- (C) $y = a \ln 2x + b$ (D) $y = a\sqrt{x} + b$
- (ii) Using the model you chose in part (i), write down the equation for the relationship between x and y , giving the numerical values of the coefficients. State the product moment correlation coefficient for this model.
- (iii) Give two reasons why it would be reasonable to use your model to estimate the value of y when $x = 189$. [9758/2017/2/8b]

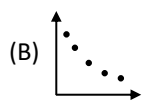
Solution

- (i)

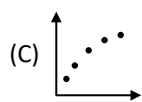




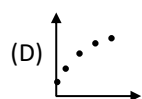
Model A ($y = ax^2 + b$) is



Model B ($y = \frac{a}{x^2} + b$) is



Model C ($y = a \ln 2x + b$)



Model D ($y = a\sqrt{x} + b$) is

(ii) Using GC,

Product moment correlation coefficient for this model: $r =$

(iii)

9 MISCELLANEOUS EXAMPLES

Example 12

A random sample of eight drivers insured with a company and having similar auto policies was selected. The table lists their driving experiences (in t years) and monthly auto insurance premiums (x).

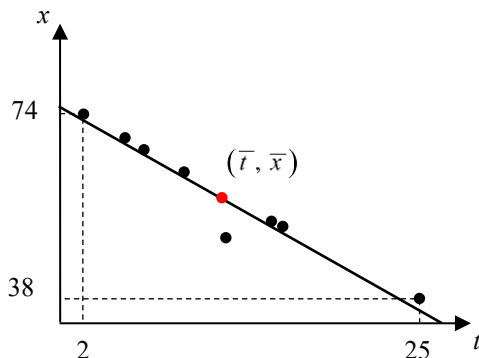
Driving Experience (t years)	5	2	12	9	15	6	25	16
Monthly Auto Insurance Premium (\$ x)	69	74	50	63	53	67	38	52

- (i) Give a sketch of the scatter diagram for the data, as shown on your calculator. [1]
- (ii) Find the product moment correlation coefficient and comment on its value in the context of the data. [2]
- (iii) Find the equation of the line of regression in the form $x = mt + c$, giving the values of m and c correct to 3 decimal places. Sketch this line in your scatter diagram. [2]
- (iv) Find $\bar{x}$ and $\bar{t}$ and mark the point $(\bar{t}, \bar{x})$ in your scatter diagram. [2]

- (v) Calculate an estimate of the monthly auto insurance premium for a driver who has 18 years of driving experience. State why you would expect this to be a reliable estimate. [2]
- (vi) Calculate an estimate of the number of years of driving experience of a driver if the monthly auto insurance premium were to be \$12. Comment on the suitability of the linear model in this context. [2]
- (vii) Explain the meaning of the values of m and c in the context of the data and comment on the reliability of this value of c . [2]

Solution

(i)



(ii) $r = -0.971$

This shows a strong negative linear correlation between driving experience and monthly auto insurance premium. The more the years of driving experience, the lower the monthly premium.

(iii) $x = 75.954 - 1.574t$

(iv) $\bar{t} = 11.25$ and $\bar{x} = 58.25$

(v) When $t = 18$, monthly premium,

$$x = 75.954 - 1.5737(18) = 47.6 \text{ (3 s.f.)}$$

This estimate is reliable as r is close to -1 and t is within the data range (interpolation).

(vi) When $x = 12$, $t = \frac{75.954 - 12}{1.5737} = 40.6 \text{ (3 s.f.)}$

Linear model is not suitable. It is not appropriate to expect a driver with 40.6 years of driving experience to pay a low premium as the driver would be very old. It is reasonable to expect older drivers to be more accident-prone so they should pay a higher premium.

(vii) $m = -1.574$ means that for every unit increase in driving experience in 1 year, we will expect a decrease of \$1.57 in monthly auto insurance premium.

$c = 75.954$ means that the monthly auto insurance premium of a new driver without experience is \$75.95.

This value of c may not be reliable as $t = 0$ is outside the data range of values of t that was measured. The extrapolated value of c may not be reliable since the relationship may not continue to be linear.

■

Example 13

Investigation of the relationship between the number of questions (x) completed in an examination and the percentage marks (m) for 12 candidates resulted in the following data:

$$\begin{aligned}\sum x &= 76 & \sum m &= 913 & \sum m^2 &= 75153 \\ \sum xm &= 6425 & \sum (x - \bar{x})^2 &= \frac{236}{3}\end{aligned}$$

- (i) Calculate the linear product moment coefficient of correlation between x and m for this sample. Comment on this value found.
- (ii) The teacher of the class added 5 marks to each student's mark to obtain the new mark (y). Explain why the linear product moment correlation coefficient remains unchanged.
- (iii) Find the equation of the appropriate regression line.
- (iv) Estimate the number of questions completed by a candidate if his percentage marks is 80. Comment on the reliability of this estimate.

Solution

$$\begin{aligned}\text{(i)} \quad r &= \frac{\sum xm - \frac{\sum x \sum m}{12}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{12}\right)\left(\sum m^2 - \frac{(\sum m)^2}{12}\right)}} \\ &= \frac{6425 - \frac{(76)(913)}{12}}{\sqrt{\left(\frac{236}{3}\right)\left(75153 - \frac{913^2}{12}\right)}} = 0.961\end{aligned}$$

Since the r value is close to 1, it indicates a strong positive linear correlation between the number of questions completed and the percentage marks. As the number of questions completed increases, the percentage marks increases.

$$\begin{aligned}\text{(ii)} \quad y &= m + 5 \Rightarrow \bar{y} = \bar{m} + 5 \\ \Rightarrow y - \bar{y} &= (m + 5) - (\bar{m} + 5) = m - \bar{m} \\ \therefore r_{xy} &= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum (x - \bar{x})(m - \bar{m})}{\sqrt{\sum (x - \bar{x})^2 \sum (m - \bar{m})^2}} = r_{xm}\end{aligned}$$

Hence r is not affected by a translation of data of m .

- (iii) You need to find $\bar{x}$ and $\bar{m}$:

$$\bar{x} = \frac{\sum x}{12} = \frac{76}{12} = \frac{19}{3} \quad \text{and} \quad \bar{m} = \frac{\sum m}{12} = \frac{913}{12}$$

Gradient of the line:

$$\begin{aligned}\text{regression coefficient} &= \frac{\sum (x - \bar{x})(m - \bar{m})}{\sum (x - \bar{x})^2} = \frac{\sum xm - \frac{\sum x \sum m}{12}}{\sum x^2 - \frac{(\sum x)^2}{12}} \\ &= \frac{6425 - \frac{76 \times 913}{12}}{\left(\frac{236}{3}\right)} = \frac{482}{59}\end{aligned}$$

$\therefore$ Regression line of m on x :

$$m - \frac{913}{12} = \frac{482}{59} \left(x - \frac{19}{3} \right) \Rightarrow m = \frac{482}{59}x + \frac{5745}{236} \approx 8.17x + 24.3$$

(iv) When $m = 80$, $80 = \frac{482}{59}x + \frac{5745}{236} \Rightarrow x \approx 6.81$

Although the r value is close to 1 which indicates a strong positive linear correlation between the percentage marks and number of questions completed, however, it is not known if $m = 80$ is within the given data range and thus we are not able to determine the reliability of this estimate. ■

Example 14

An aptitude test, designed to predict the potential productivity of new employees, was given to a random sample of nine employees. Based on the original sample data, a student calculated the least squares regression line of y on x to be $y = 1.309x + 7.059$. However, he was rather careless and accidentally erased the aptitude score for Employee A . The table shows the test results together with a measure of the actual productivity of each employee:

Employee	A	B	C	D	E	F	G	H	J
Aptitude score, x	k	17	20	19	20	23	16	24	22
Productivity, y	20	25	29	33	43	32	24	33	47

- (i) Based on the regression line $y = 1.309x + 7.059$, find the value of k , correct to the nearest integer. Explain the significance of the slope of this regression line.

- (ii) Let $e_i = y'_i - y_i$ where y'_i is the predicted productivity score obtained by a regression line and y_i is the corresponding observed score. Find $\sum_{i=1}^n e_i^2$, correct to 2 decimal places, for
- the least squares regression line $y = 1.309x + 7.059$,
 - a regression line $y = 1.5x + 8$.
- Comment on your results.
- (iii) Calculate an estimate of the productivity score of a person who scored 30 for his aptitude test. Comment on the reliability of this estimate.

Solution

Note that k cannot be found by substituting into the regression line equation. Instead we need to use the fact that $(\bar{x}, \bar{y})$ lies on the regression line.

(i) $\bar{y} = \frac{\sum y}{9} = \frac{286}{9}$ and $\bar{x} = \frac{\sum x}{9} = \frac{k+161}{9}$

Since $(\bar{x}, \bar{y})$ lies on the regression line,

$$\frac{286}{9} = 1.309 \left(\frac{k+161}{9} \right) + 7.059 \Rightarrow k \approx 8.95 \approx 9$$

Significance of the slope:

For every unit increase in aptitude score, productivity increased by 1.309 units.

(ii)(a) Finding $\sum_{i=1}^n e_i^2$ using GC

Since $e_i = y'_i - y_i$, we need to create a list for y'_i and a list for e_i^2 (or e_i).

- 1) Key in **L3** as **1.309L1+7.059**. This will be the list for y'_i .

Key in **L4** as **(L3-L2)²**. This will be e_i^2 .

L1	L2	L3	L4	L5	3
9	20	18.84	-----	-----	
17	25	29.312			
20	29	33.239			
19	33	31.93			
20	43	33.239			
23	32	37.166			
16	24	28.003			
24	33	38.475			
22	47	35.857			
-----	-----	-----			
L3=1.309L1+7.059					

L1	L2	L3	L4	L5	4
9	20	18.84	1.3456	-----	
17	25	29.312	18.593		
20	29	33.239	17.969		
19	33	31.93	1.1449		
20	43	33.239	95.277		
23	32	37.166	26.688		
16	24	28.003	16.024		
24	33	38.475	29.976		
22	47	35.857	124.17		
-----	-----	-----	-----		
L4=(L3-L2) ²					

- 2) Press **[STATS]**, select **CALC** and then **1:1-Var Stats** for **L4**. $\sum_{i=1}^n e_i^2$ is found by looking at **Σx** .

1-Var Stats
List:L4
FreqList:
Calculate

1-Var Stats
$\bar{x}=36.79819167$
$\Sigma x=331.183725$
$\Sigma x^2=27034.29263$
$Sx=43.08034709$
$\sigma x=40.61654075$
$n=9$
$\min X=1.1449$
$\downarrow Q_1=8.6848045$

$$\sum_{i=1}^n e_i^2 = 331.18$$

- (ii)(b) Repeat the same process as in (ii)(a), except for **L3**, use **1.5L1+8**, we

get $\sum_{i=1}^n e_i^2 = 524$.

Since the result from (a) is calculated based on the least square regression line, the sum should be lower than the sum of errors for any other regression line used to predict the relationship between aptitude and productivity.

- (iii) When $x = 30$ $y = 1.309(30) + 7.059 = 46.3$

Since the aptitude score of 30 is outside the given data range, $19 \leq x \leq 24$, the extrapolated value of y may not be reliable since the relationship may not continue to be linear. ■

ANNEX A**Gradient of Least Squares Regression Line**

$$\begin{aligned}
b &= \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \\
&= \frac{\sum (xy - \bar{x}y - x\bar{y} + \bar{x} \cdot \bar{y})}{\sum (x^2 - 2\bar{x}x + \bar{x}^2)} \\
&= \frac{\sum xy - \bar{x} \sum y - \bar{y} \sum x + n\bar{x} \cdot \bar{y}}{\sum x^2 - 2\bar{x} \sum x + n\bar{x}^2} \\
&= \frac{\sum xy - \frac{\sum x}{n} \sum y - \frac{\sum y}{n} \sum x + n \frac{\sum x}{n} \frac{\sum y}{n}}{\sum x^2 - 2 \frac{\sum x}{n} \sum x + n \left(\frac{\sum x}{n} \right)^2} \\
&= \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}
\end{aligned}$$

Hence, it can be observed that

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n} \text{ and } \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

and replacing y by x ,

$$\sum (x - \bar{x})(x - \bar{x}) = \sum x^2 - \frac{\sum x \sum x}{n} = \sum x^2 - \frac{(\sum x)^2}{n}$$

or replacing x by y ,

$$\sum (y - \bar{y})(y - \bar{y}) = \sum y^2 - \frac{\sum y \sum y}{n} = \sum y^2 - \frac{(\sum y)^2}{n}$$

ANNEX B

Relationship between Correlation Coefficient and Least Squares Regression Lines

Equation of the regression line of y on x :

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

Where $b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$.

Equation of the regression line of x on y :

$$x - \bar{x} = d(y - \bar{y}) \text{ or } y - \bar{y} = \frac{1}{d}(x - \bar{x})$$

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

Now $r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$, therefore

$$r^2 = \frac{(\sum (x - \bar{x})(y - \bar{y}))^2}{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2} = bd = \frac{b}{\frac{1}{d}},$$

which is the ratio of the gradients of the two regression lines.

Consequently, we have the following:

1. If the linear correlation is perfect (the sample points lie on a straight line), $r = \pm 1$, then $b = \frac{1}{d}$. Hence, the least squares regression line of y on x and the least squares regression line of x on y will **coincide**. In general, the closer the value of r to 1 or -1 , the nearer the two regression lines are to each other.
2. Since $r^2 \geq 0 \Rightarrow bd \geq 0$, therefore b and d must have the same sign and the sign of b or d follows that of r .

If b and d are both positive, then $r = \sqrt{bd}$, the regression lines have positive gradients.

If b and d are both negative, then $r = -\sqrt{bd}$, the regression lines have negative gradients.

3. Since $r^2 \leq 1 \Rightarrow bd \leq 1 \Rightarrow |b| \leq \frac{1}{|d|}$, therefore the gradient of the line of y on x is “less steep” than the gradient of the line of x on y when plotted on the y against x axes.

Diagrams of Regression Lines

