



## **Chapter S6: Correlation and Regression**

### **Syllabus will include**

- Use of scatter diagram to judge if there is a plausible linear relationship between the two variables
- Correlation coefficient as a measure of the fit of a linear model to the scatter diagram
- Interpreting the product moment correlation coefficient (in particular, values close to  $-1$ ,  $0$  and  $1$ )
- Concepts of linear regression and method of least squares to find the equation of the regression line
- Concepts of interpolation and extrapolation
- Use of the appropriate regression line to make prediction or estimate a value in practical situations, including explaining how well the situation is modelled by the linear regression model
- Use of a square, reciprocal or logarithmic transformation to achieve linearity

### **CONTENT**

#### **1 Terminology**

- 1.1 Describing Variables
- 1.2 Scatter Diagram
  - 1.2.1 Producing a Scatter Diagram using the Graphing calculator
  - 1.2.2 Interpreting Scatter Diagrams

#### **2 Product Moment Correlation Coefficient**

- 2.1 Calculating the Value of the Product Moment Correlation Coefficient
- 2.2 Properties of the Product Moment Correlation Coefficient
- 2.3 Interpreting Correlation
  - 2.3.1 Importance of Scatter Diagram
  - 2.3.2 Correlation does not imply Causation

#### **3 Linear Regression**

- 3.1 Least Squares Regression Line of  $y$  on  $x$
- 3.2 Least Squares Regression Line of  $x$  on  $y$
- 3.3 Calculating the Least Squares Regression Lines and Product Moment Correlation Coefficient using the Graphing calculator

#### **4 Application and Interpretation**

#### **5 Linearisation of Data**

Appendix: Equivalence of the 2 Formulae for the Product Moment Correlation Coefficient

## **INTRODUCTION**

Suppose that you are helping a proud mother to keep track of the weight of her newborn. After monitoring the infant's weight ( $y$  kg) for a period of time ( $t$  days), you obtain the following data:

|     |      |      |      |      |       |       |       |
|-----|------|------|------|------|-------|-------|-------|
| $t$ | 7    | 32   | 71   | 97   | 188   | 273   | 409   |
| $y$ | 4.43 | 4.88 | 6.31 | 7.18 | 10.63 | 13.60 | 17.95 |

Can you then tell the mother how much her baby will weigh by the next month? What about the next year? How certain would you be of your prediction? Can you further use your data to check if the infant's growth is normal?

In this chapter, we will deal with the skills required to answer the above questions. Statistical inference entails the study of **correlation** between variables and **regression analysis**, the art and science of finding these relationships between variables. Empirical data rarely corroborate exactly with theoretical mathematical models, hence the need for statistical analysis to accommodate such deviations.

## **1 TERMINOLOGY**

### **1.1 Describing Variables**

- The data in the above example are a set of pairs of values for two variables. This is an example of **bivariate data**, where each observation requires the values of two variables.
- Since an infant's weight is dependent upon his/her age,  $t$  is the **independent variable** and  $y$  is the **dependent variable** in the above example.
- Sometimes the independent variable is **controlled** so that the variable only assumes a set of predetermined values.

### **Example 1**

Identify the dependent and independent variables in each of the scenarios below.

- (a) An object is heated up and its volume measured at different temperatures. The following table shows eight of these measurements.

|                                         |       |       |       |       |       |       |       |       |
|-----------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|
| Temperature ( $^{\circ}\text{C}$ ), $t$ | 30    | 40    | 50    | 60    | 70    | 80    | 90    | 100   |
| Volume ( $\text{m}^3$ ), $v$            | 1.000 | 1.256 | 1.395 | 1.523 | 1.648 | 1.699 | 1.712 | 1.735 |

**Ans:** The independent variable is  $t$ , and the dependent variable is  $v$ .

**Note:** A controlled variable is an independent variable.

- (b) At a restaurant, the waiting time is defined as the time between sitting down at a table and a waiter first arriving at your table. This waiting time is dependent upon the number of other customers already seated in the restaurant. John is a customer who visited the restaurant on 10 different days. The table shows, for each of these days, the number,  $x$ , of customers already seated and his waiting time, in  $y$  minutes.

|     |    |   |   |    |   |    |   |    |   |   |
|-----|----|---|---|----|---|----|---|----|---|---|
| $x$ | 9  | 3 | 4 | 10 | 8 | 12 | 7 | 11 | 2 | 6 |
| $y$ | 11 | 6 | 5 | 11 | 9 | 13 | 9 | 12 | 4 | 7 |

(Taken from GCE Jan 2006 ASE)

**Ans:** The independent variable is  $x$ , and the dependent variable is  $y$ .

- (c) The following table shows the marks scored by 10 randomly selected students in a French test and in an English test.

|                    |    |    |    |    |    |    |    |    |    |    |
|--------------------|----|----|----|----|----|----|----|----|----|----|
| French score, $x$  | 20 | 43 | 33 | 56 | 50 | 67 | 73 | 68 | 77 | 43 |
| English score, $y$ | 19 | 42 | 44 | 52 | 51 | 53 | 66 | 56 | 60 | 37 |

**Ans:** From the above information, we are not able to determine which is dependent/independent variable.

## 1.2 Scatter Diagram

A picture speaks a thousand words – what better way to obtain an overview of the relationship between variables than observing it from a diagram.

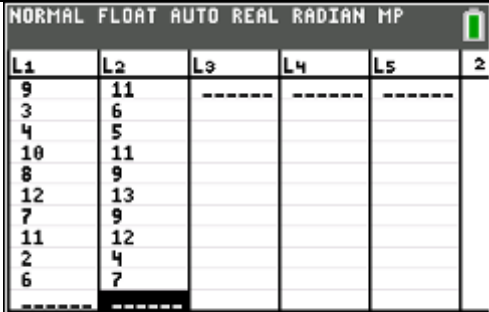
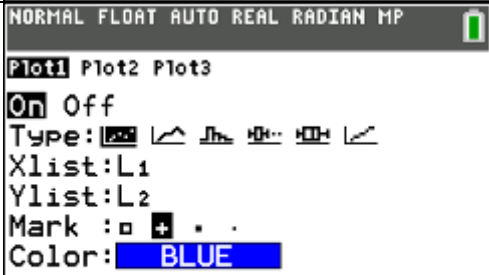
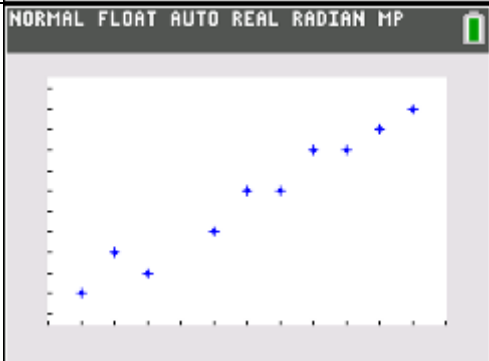
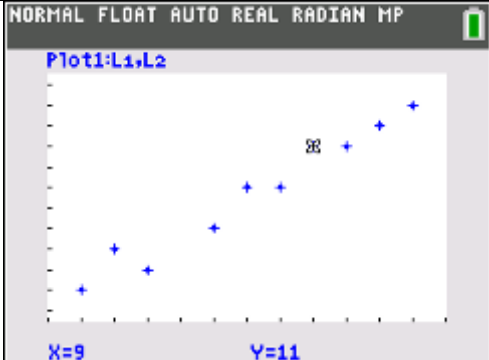
A **scatter diagram** is a two-dimensional plot, with the values of one variable plotted along each axis. By convention, we **plot the independent variable along the horizontal axis**.

### 1.2.1 Producing a Scatter Diagram using the Graphing calculator

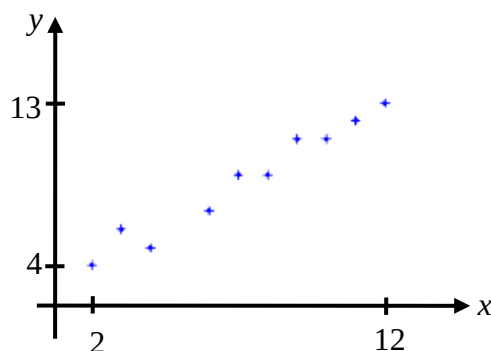
Consider the following set of data, with  $x$  as the independent variable and  $y$  as the dependent variable.

|     |    |   |   |    |   |    |   |    |   |   |
|-----|----|---|---|----|---|----|---|----|---|---|
| $x$ | 9  | 3 | 4 | 10 | 8 | 12 | 7 | 11 | 2 | 6 |
| $y$ | 11 | 6 | 5 | 11 | 9 | 13 | 9 | 12 | 4 | 7 |

Using the Graphing Calculator to produce a scatter diagram:

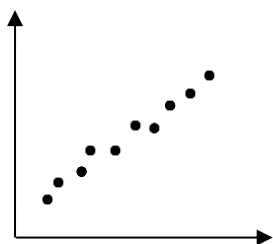
|                                                                                                                                                                                                                                                                                                                |                                                                                                   |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| <p>1. Press <b>[stat]</b><b>[enter]</b> to go to the List editor.</p> <p>2. Key the data into L<sub>1</sub> and L<sub>2</sub>.</p>                                                                                                                                                                             |  <p>L2(11)=</p> |
| <p>3. Press <b>[2nd]</b><b>[stat plot]</b> to turn on the Stat Plot function.</p> <p>4. Xlist and Ylist refer to the list of variables on the horizontal and vertical axes respectively. In this example, ensure that the axes are set to correspond to the correct lists L<sub>1</sub> and L<sub>2</sub>.</p> |                |
| <p>5. Press <b>[zoom]</b><b>[9]</b> to view the scatter diagram with all statistical data points displayed.</p>                                                                                                                                                                                                |               |
| <p>6. Press <b>[trace]</b> and the left and right arrows to see the coordinates of each data point.</p>                                                                                                                                                                                                        |               |

Note: When drawing a scatter diagram, ensure the relative position of the points are correct, and label the axes and range of the data:



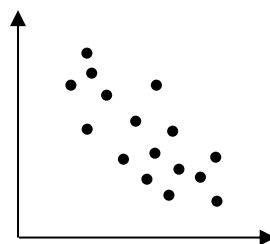
### 1.2.2 Interpreting Scatter Diagrams

If all the points in a scatter diagram seem to lie close to a straight line, we say there is a **linear correlation** between the variables.



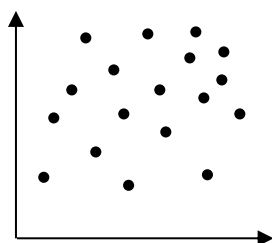
Both the variables increase together. The points lie close to a straight line.

We say that the variables have a strong positive linear correlation.

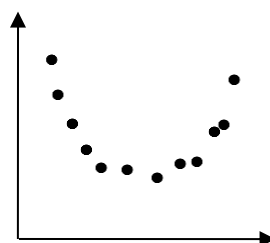


As one variable increases, the other decreases. There is a negative linear correlation between the variables.

The points are further away from a straight line as compared to the first case. So the linear correlation here is not as strong as the first case.



There appears to be no clear relation between the variables.



There is a non-linear relation between the variables.

## 2 PRODUCT MOMENT CORRELATION COEFFICIENT

A scatter diagram shows visually the relation between two variables. In this section, we measure the strength of the linear relation between the variables by quantifying it.

The **product moment correlation coefficient**, denoted by  $r$ , is a numerical measure of **the linear relation** between two variables.

### 2.1 Calculating the Value of the Product Moment Correlation Coefficient

There are 2 equivalent formulae for the product moment correlation coefficient that can be found in the formula list, should one be required to compute it by hand when given a data set of  $n$  observations  $(x, y)$ .

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

How these two formulae came about is beyond the scope of the A-Level syllabus. But their equivalence is easy [Refer to Appendix].

#### Example 2

A data set of 10 observations  $(x, y)$  are summarised below:

$$\sum x = 72, \sum x^2 = 624, \sum y = 87, \sum y^2 = 843, \sum xy = 720$$

Calculate the value of the product moment correlation coefficient.

#### Solution:

The value of the product moment correlation coefficient is

$$r = \frac{720 - \frac{(72)(87)}{10}}{\sqrt{\left( (624) - \frac{(72)^2}{10} \right) \left( (843) - \frac{(87)^2}{10} \right)}} = 0.982 \text{ (3 s.f.)}$$

#### Remark:

In the A-Level examination, raw data is usually given. If you are required to use the formula, you can refer to the formula list.

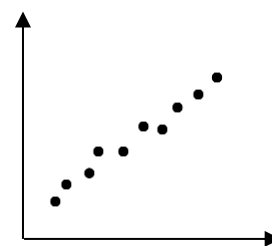
## 2.2 Properties of the Product Moment Correlation Coefficient

- (a) The value of  $r$  is independent of the units of the variables.
- (b) The correlation coefficient is a property of the bivariate data set itself and is not determined by which is the independent variable.
- (c)  $-1 \leq r \leq 1$
- (d) Relationship between the value of  $r$  and the appearance of the scatter diagram

### Positive Linear Correlation, i.e. $r > 0$

**In a scatter plot, when an increase in one variable is generally associated with an increase in the second variable, we say that the two variables are positively correlated.**

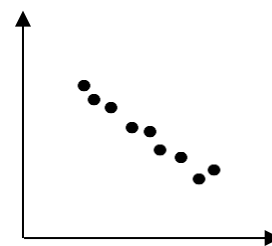
The closer the value of  $r$  is to 1 the stronger the positive linear correlation between the two variables, i.e. the closer the points on the scatter diagram are to a straight line of positive gradient.



### Negative Linear Correlation, i.e. $r < 0$

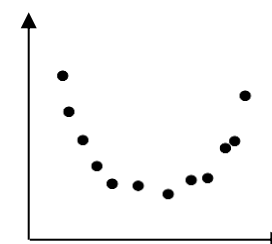
**On the other hand, if an increase in one variable is associated with a decrease in the second variable, we then say that the two variables are negatively correlated.**

The closer the value of  $r$  is to  $-1$  the stronger the negative linear correlation, i.e. the closer the points on the scatter diagram are to a straight line of negative gradient.



### No Linear Correlation, i.e. $r \approx 0$

**If the variables have a non-linear correlation, we will look at how to deal with such relationships in Section 5.**



Special values of  $r$

- (i)  $r = 1$ : Perfect positive linear correlation between variables.  
All data points lie exactly on a straight line of positive gradient.
- (ii)  $r = -1$ : Perfect negative linear correlation between variables.  
All data points lie exactly on a straight line of negative gradient.
- (iii)  $r = 0$ : **No linear correlation.** Note that this does not imply no relationship; the relation between variables may be modelled by another type of function.

## 2.3 Interpreting Correlation

As much as the correlation coefficient can tell us about the relation between two variables, there are also pitfalls that one should avoid falling into when interpreting its value.

### 2.3.1 Importance of Scatter Diagram

The correlation coefficient alone is insufficient for interpretation of relation between variables. The scatter diagram gives a picture of the correlation. It also helps to identify any results that are well away from the other points and so do not fit the general pattern of any correlation shown up by the diagram. Such a result is known as an **outlier**. Removing them gives a truer picture of the relation between the variables.

#### Example 3

Two sets of bivariate data have correlation coefficients that are close to zero. The scatter diagrams for the two sets of data are shown below.



The scatter diagrams show that the variables in the first set of data have no clear relation, while those in the second set exhibit a non-linear relation.



**Example 4**

Consider the following set of data:

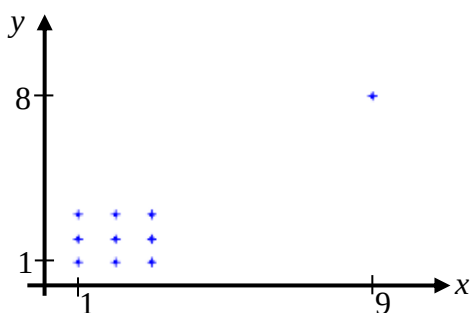
|     |   |   |   |   |   |   |   |   |   |   |
|-----|---|---|---|---|---|---|---|---|---|---|
| $x$ | 1 | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 3 | 9 |
| $y$ | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 8 |

It is given that the product moment correlation coefficient is 0.862 (3s.f.).

- (a) Assuming  $x$  is the independent variable, draw a scatter diagram for the set of data.  
 (b) Is there a strong linear relation between the variables? Justify your answer.

**Answer:**

(a)



- (b) Although  $r = 0.862$  is close to 1, which suggests a strong positive linear correlation between  $x$  and  $y$ , the scatter diagram shows otherwise.

Note: The above inconsistency is most likely due to the presence of the outlier (9,8).  
 If we remove the outlier, then there is no linear relationship between the variables.

**Example 5 MJC Prelims 2007/02/Q9**

In an investigation of a shrub, research workers measured the average width  $y$  (in cm) and stem density  $x$  (number of stems per  $\text{m}^2$ ) at ten sites with the following results:

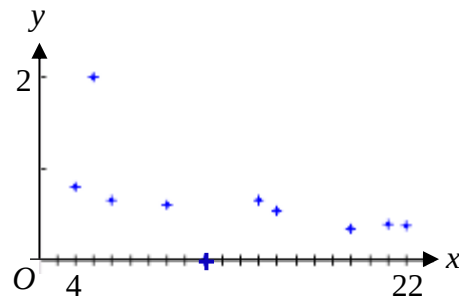
|     |     |   |      |     |    |      |      |      |      |      |
|-----|-----|---|------|-----|----|------|------|------|------|------|
| $x$ | 4   | 5 | 6    | 9   | 11 | 14   | 15   | 19   | 21   | 22   |
| $y$ | 0.8 | 2 | 0.65 | 0.6 | 0  | 0.65 | 0.55 | 0.35 | 0.40 | 0.38 |

- (i) Calculate the correlation coefficient for the data.  
 (ii) Give a sketch of the scatter diagram for the data, as shown on your calculator, and comment on your answer in (i).  
 (iii) Identify the two data pairs which should be removed.

**Solutions**

(i)  $r = -0.542$

(ii)



The scatter diagram indicates a strong negative linear correlation between  $x$  and  $y$ . This is not consistent with the value of  $r$  in (i) due to the presence of the outlier.

(iii) The data pairs which should be removed are (5,2) and (11,0).

\* Note that (11,0) is obviously erroneous as the average width of the shrub must be positive.

### 2.3.2 Correlation does not imply Causation

The existence of a linear correlation between two variables does not mean that the change in one variable causes the change in another.

For example, there may be a positive correlation between the average life expectancy of Singaporeans and the memory capacity of desktop computers, but this does not mean one is the cause of another. Both increases may be due to scientific development over time.

The product moment correlation coefficient merely gives an idea of the linear relationship between the variables. It does not imply any cause-and-effect relationship between the variables. There may be intermediate variables involved in the relationship which we do not know about, or there may even be more than one explanation to the linear relation.

### 3 LINEAR REGRESSION

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

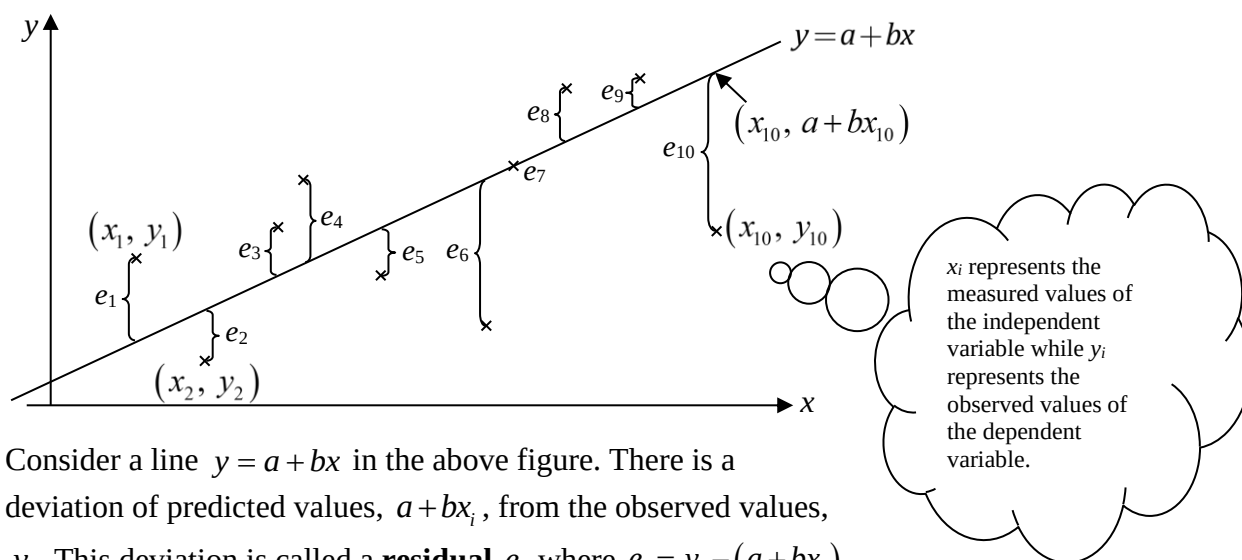
Before attempting to fit a linear model to the observed data, we should first determine whether there is a relationship between the two variables of interest. A scatter diagram will give us a good indication whether it is meaningful to model the observed data with a straight line.

One method that you may have previously employed in drawing a line that best fits the given data is simply to draw a line such that it passes through or as close as possible to as many points as possible. The drawback of such a method is its subjectivity – two people may draw different lines fitting the same scatter plot with neither line having a basis for claiming superior predictive power.

For greater accuracy and uniformity, it is preferable to calculate the equation of the line. One way is as follows.

#### 3.1 The Least Squares Regression Line of $y$ on $x$

Given a data set with  $x$  as the **independent variable** and  $y$  as the **dependent variable**, our goal is to find a line with equation  $y = a + bx$  to model the linear relation between the variables.



Consider a line  $y = a + bx$  in the above figure. There is a deviation of predicted values,  $a + bx_i$ , from the observed values,  $y_i$ . This deviation is called a **residual**  $e_i$  where  $e_i = y_i - (a + bx_i)$ .

We would like a best fit line which makes the residuals as small as possible, and there are different ways this could be done. The **least squares regression line** is the line which minimizes the **sum of the squares of the residuals**, i.e.

$$\sum_{i=1}^n e_i^2 \text{ is a minimum.}$$

The least squares regression line of  $y$  on  $x$  can be shown to be

$$y - \bar{y} = b(x - \bar{x}), \text{ where } b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}. \quad [\text{in Formula List}]$$

### 3.2 The Least Squares Regression Line of $x$ on $y$

When  $y$  is the independent variable and  $x$  the dependent variable, we will use the least squares regression line of  $x$  on  $y$  which takes on the form  $x = c + dy$ .

Similarly, the least squares regression line of  $x$  on  $y$  can be shown to be

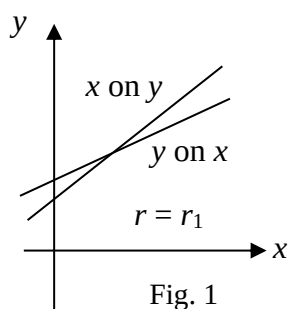
$$x - \bar{x} = d(y - \bar{y}), \text{ where } d = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2}. \quad [\text{not in Formula List}]$$

**Note:** Both the least squares regression lines of  $y$  on  $x$  and  $x$  on  $y$  pass through the point  $(\bar{x}, \bar{y})$ .

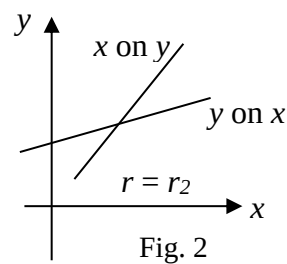
In general, the least squares regression line of  $y$  on  $x$  is **not** the same line as the least square regression line of  $x$  on  $y$ .

The regression lines of  $y$  on  $x$  and  $x$  on  $y$  are the same line if and only if the product moment correlation coefficient between  $x$  and  $y$  is 1 or  $-1$ .

The closer the coefficient is to either of these values, the closer are the lines to each other.



$$r_1 > r_2 > 0$$



### 3.3 Calculating the Least Squares Regression Lines and Product Moment Correlation Coefficient using the Graphing Calculator

The GC is capable of computing the equation of the least squares regression line. We shall illustrate the procedure with a worked example.

#### Example 6

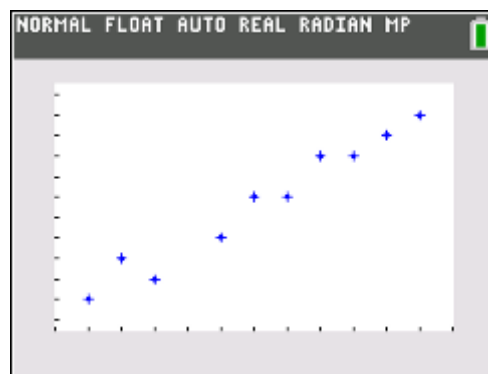
For the following set of data, find the equation of

|     |    |   |   |    |   |    |   |    |   |   |
|-----|----|---|---|----|---|----|---|----|---|---|
| $x$ | 9  | 3 | 4 | 10 | 8 | 12 | 7 | 11 | 2 | 6 |
| $y$ | 11 | 6 | 5 | 11 | 9 | 13 | 9 | 12 | 4 | 7 |

- (i) the least squares regression line of  $y$  on  $x$ ,
- (ii) the least squares regression line of  $x$  on  $y$ .

Find also

- (iii) the product moment correlation coefficient for this set of data,
- (iv) the values of  $\bar{x}$  and  $\bar{y}$ .



#### Solution:

- Key the values of  $x$  and  $y$  into  $L_1$  and  $L_2$  respectively. (The data set is the same as that in Section 1.2.1.)

| L1 | L2 | L3  | L4  | L5  |   |
|----|----|-----|-----|-----|---|
| 9  | 11 | --- | --- | --- | 2 |
| 3  | 6  |     |     |     |   |
| 4  | 5  |     |     |     |   |
| 10 | 11 |     |     |     |   |
| 8  | 9  |     |     |     |   |
| 12 | 13 |     |     |     |   |
| 7  | 9  |     |     |     |   |
| 11 | 12 |     |     |     |   |
| 2  | 4  |     |     |     |   |
| 6  | 7  |     |     |     |   |

L2(L1)=

- Press  $\boxed{\text{stat}}$ , then choose **CALC** and  
**4:LinReg(ax+b)** or  
**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    |        |

3. Enter the independent variable in the Xlist and the dependent variable in the Ylist.

The Store RegEQ: is optional. If the equation of the regression line is stored in  $Y_1$ , the GC can draw the least squares regression line on the scatter diagram, as well as use the equation to predict values.

To obtain  $L_1$ , press  $\text{2nd}[L_1]$

For  $Y_1$ , press  $\text{alpha}[f_4]$ , then choose  $Y_1$ .

Select **Calculate**.

If your GC does not show the values of  $r^2$  and  $r$ , press  $\text{mode}$  and select **On** for **Stat Diagnostics**. Then repeat steps 2 and 3.

NORMAL FLOAT AUTO REAL RADIAN MP

**LinReg(ax+b)**

Xlist:L1  
Ylist:L2  
FreqList:  
Store RegEQ:Y1  
Calculate

NORMAL FLOAT AUTO REAL RADIAN MP

**LinReg**

$y=ax+b$   
 $a=0.8863636364$   
 $b=2.318181818$   
 $r^2=0.9635730124$   
 $r=0.9816175489$

NORMAL FLOAT AUTO REAL RADIAN MP

MATHPRINT CLASSIC  
NORMAL SCI ENG  
FLOAT 0 1 2 3 4 5 6 7 8 9  
RADIAN DEGREE  
FUNCTION PARAMETRIC POLAR SEQ  
THICK DOT-THICK THIN DOT-THIN  
SEQUENTIAL SIMUL  
REAL a+bi re^(θi)  
FULL HORIZONTAL GRAPH-TABLE  
FRACTIONTYPE: n/d Un/d  
ANSWERS: AUTO DEC  
STATDIAGNOSTICS: OFF ON  
STATWIZARDS: ON OFF  
SETCLOCK 01/01/15 12:00 AM  
LANGUAGE: ENGLISH

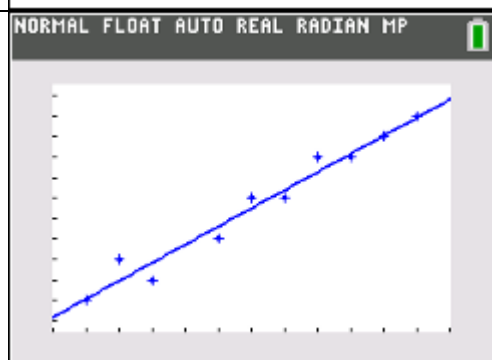
4. The equation of the least squares regression line of  $y$  on  $x$  is now stored in  $Y_1$ .

NORMAL FLOAT AUTO REAL RADIAN MP

Plot1 Plot2 Plot3

$\text{Y}_1=0.8863636364X+2.318181818$   
 $\text{Y}_2=$   
 $\text{Y}_3=$   
 $\text{Y}_4=$   
 $\text{Y}_5=$   
 $\text{Y}_6=$   
 $\text{Y}_7=$   
 $\text{Y}_8=$   
 $\text{Y}_9=$

5. Screenshot of how the regression line on the scatter plot will look like (not required in the solution for this example).



- (i) The equation of the least squares regression line of  $y$  on  $x$  is

$$y = 2.32 + 0.886x,$$

with coefficients correct to 3 significant figures.

To find the least squares regression line of  $x$  on  $y$ , we will enter the  $y$ -values in  $L_2$  as values of the independent variable in the Xlist, while  $x$ -values in  $L_1$  as values of the dependent variable in the Ylist.

```
LinReg(ax+b)
Xlist:L2
Ylist:L1
FreqList:
Store RegEQ:
Calculate
```

[Note that it will not be useful to store the equation in this case as GC stores the variable entered in Xlist as  $x$  and the other entered in Ylist as  $y$ .]

- (ii) The equation of the least squares regression line of  $x$  on  $y$  is

$$x = 1.09y - 2.26,$$

with coefficients correct to 3 significant figures.

- (iii) The product moment correlation coefficient is 0.982 (3 s.f.).

- (iv) To find the values of  $\bar{x}$  and  $\bar{y}$ , press **[stat]**, then choose **CALC** and **2:2-Var Stats**.

Enter Xlist and Ylist.  
Select **Calculate**.

From GC,  $\bar{x} = 7.2$  and  $\bar{y} = 8.7$ .

```
NORMAL FLOAT AUTO REAL RADIAN MP
2-Var Stats
Xlist:L1
Ylist:L2
FreqList:
Calculate
```

```
NORMAL FLOAT AUTO REAL RADIAN MP
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
```

```
NORMAL FLOAT AUTO REAL RADIAN MP
2-Var Stats
x̄=7.2
Σx=72
Σx²=624
Sx=3.425395354
σx=3.249615362
n=10
ȳ=8.7
↓Σy=87
```

**Remark:** Since  $r$  is close to 1, if we solve for  $y$  in terms of  $x$ ,  $y = \frac{1}{1.09}x + \frac{2.26}{1.09}$  is close to the regression line of  $y$  on  $x$ . The 2 regression lines intersect at  $(7.2, 8.7)$  which is  $(\bar{x}, \bar{y})$ .

## 4 APPLICATION AND INTERPRETATION

Regression lines can be used to make predictions.

Consider the following data set:

|     |             |      |      |      |      |      |      |      |      |      |      |             |
|-----|-------------|------|------|------|------|------|------|------|------|------|------|-------------|
| $x$ | <b>4</b>    | 9    | 17   | 24   | 32   | 40   | 46   | 57   | 63   | 69   | 72   | <b>78</b>   |
| $y$ | <b>6.85</b> | 6.75 | 6.74 | 6.73 | 6.68 | 6.52 | 6.54 | 6.48 | 6.36 | 6.33 | 6.35 | <b>6.29</b> |

Observe that the least and greatest values of  $x$  are 4 and 78 respectively, and that of  $y$  are 6.29 and 6.85 respectively.

Using the least squares regression line to predict or estimate missing values within the range of data (e.g. estimate  $y$  when  $x = 5$  or estimate  $x$  when  $y = 6.5$ ) is known as **interpolation**. Using the line to estimate values outside the range of data is known as **extrapolation**.

Note that we do not know what happens outside the given range of experimental values. Extrapolating using values beyond the given data range is dangerous as the regression model (linear, quadratic, reciprocal, etc.) obtained may not be applicable outside the range of experimental data given.

The **least squares regression line of  $y$  on  $x$**  is used when

- $x$  is the independent variable and  $y$  is the dependent variable; or
- the independent/dependent variable cannot be determined from the given information and you want to estimate  $y$  for a given value of  $x$ .

The **least squares regression line of  $x$  on  $y$**  is used when

- $y$  is the independent variable and  $x$  is the dependent variable; or
- the independent/dependent variable cannot be determined and you want to estimate  $x$  for a given value of  $y$ .

When looking at the regression equation  $y = a + bx$ ,  $a$  is the  $y$ -intercept of the line, and  $b$  is the gradient of the line.  $a$  is the value of  $y$  when  $x = 0$  and  $b$  is the amount by which  $y$  increases for every unit of increase in  $x$ , so  $b$  is the rate of change of  $y$  with respect to  $x$ . It is necessary to interpret the meaning of these values in the context provided.

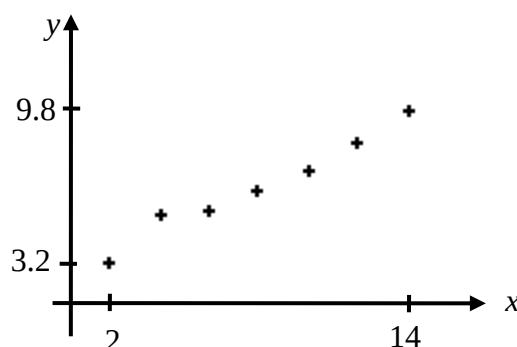


**Example 7**

A scientist working in agricultural research believes that there is a linear relationship between the amount of a certain food supplement given to hens and the hardness of the shells of the eggs they lay. As an experiment, controlled quantities of the supplement were added to the hens' normal diet and the hardness of the shells of the eggs was then measured on a scale from 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 |

- Draw a scatter diagram for the data.
- Find the product moment correlation coefficient. Find also the equation of the regression line of  $y$  on  $x$ . Give your answer in the form  $y = a + bx$ , with  $a$  and  $b$  corrected to 3 decimal places.
- Estimate the shell hardness (correct to 1d.p.) for a food supplement of 11g per day, and comment on the reliability of your estimate.
- Explain what the values of  $a$  and  $b$  tell you and why you should not try to estimate the shell hardness for a food supplement of 20g per day.

**Solution:****(i)**

- (ii)** The product moment correlation coefficient is  $r = 0.987$  (3s.f.).

The regression line of  $y$  on  $x$  is  $y = 2.529 + 0.502x$  (to 3 d.p.)

- (iii)** When  $x = 11$ ,  $y = 8.0$  (1d.p.).

The value  $x = 11$  is within the given range of  $x$ -values and the variables have a strong linear correlation indicating the linear model is a suitable model.  
So the estimate is reliable.

- (iv)**  $a = 2.529$  means the hardness of the eggshells is estimated to be 2.529 when no supplement is given.

$b = 0.502$  means the hardness of the eggshells is estimated to increase by 0.502 for every increase of 1g/day of food supplement.

The value  $x = 20$  is well outside the given range of  $x$ -values and so the estimate would be unreliable as extrapolation is used.

**Example 8**

The following data were collected during a study under experimental conditions, of the effect of temperature,  $x$  °C, on the pH,  $y$ , of skimmed milk.

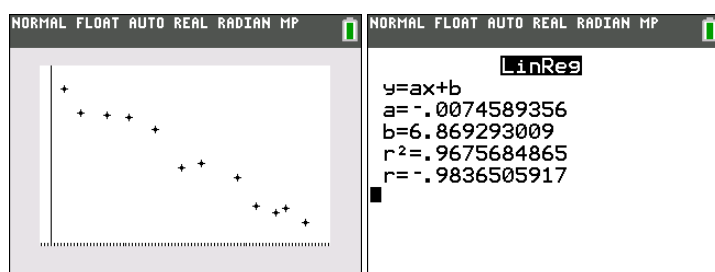
|     |      |      |      |      |      |      |      |      |      |      |      |      |
|-----|------|------|------|------|------|------|------|------|------|------|------|------|
| $x$ | 4    | 9    | 17   | 24   | 32   | 40   | 46   | 57   | 63   | 69   | 72   | 78   |
| $y$ | 6.85 | 6.75 | 6.74 | 6.73 | 6.68 | 6.52 | 6.54 | 6.48 | 6.36 | 6.33 | 6.35 | 6.29 |

Use the appropriate least squares regression line(s) to estimate

- the pH of the skimmed milk when the temperature is 27 °C,
- the temperature (to the nearest °C) of the skimmed milk when the pH is measured to be 6.64.

**Solution:**

Scatter Plot for own reference:



From GC, product moment correlation coefficient =  $-0.984$  (3sf)

Since this is a study of the effect of  $x$  on  $y$ , so  $x$  is independent and  $y$  dependent. The least squares regression line of  $y$  on  $x$  with the following equation is used for both estimates:

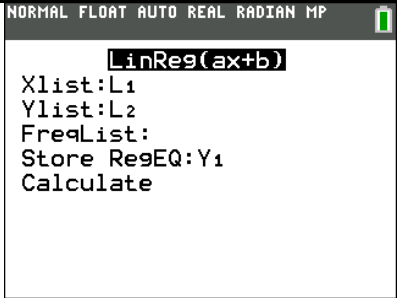
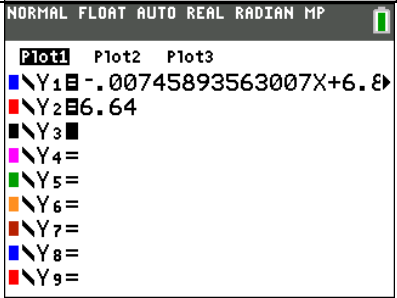
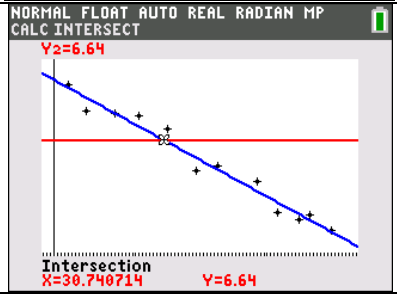
$$y = 6.8693 - 0.0074589x \text{ (5s.f.)}$$

- When  $x = 27$ ,  $y = 6.67$  (3 s.f.).

So the pH of the skimmed milk at 27 °C is estimated to be 6.67.

- When  $y = 6.64$ , we have  $x = 31$  (nearest integer).

Thus the temperature of the skimmed milk is estimated to be 31 °C.

|                                                                                                                                                                                   |                                                                                     |  |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--|
| <p>1. One way to solve for <math>x</math> when <math>y = 6.64</math> is to plot the graph of the regression line in the GC. Store the regression equation as <math>Y_1</math></p> |   |  |
| <p>2. Enter <math>Y_2</math> as 6.64</p>                                                                                                                                          |   |  |
| <p>3. Find the intersection of the 2 graphs</p>                                                                                                                                   |  |  |

**Example 9**

The following table shows the marks scored by 10 randomly selected students in a French test and in an English test.

|                    |    |    |    |    |    |    |    |    |    |    |
|--------------------|----|----|----|----|----|----|----|----|----|----|
| French score, $x$  | 20 | 43 | 33 | 56 | 50 | 67 | 73 | 68 | 77 | 43 |
| English score, $y$ | 19 | 42 | 44 | 52 | 51 | 53 | 66 | 56 | 60 | 37 |

Assume there is a linear correlation between the two scores.

- (i) A student, Tom, was absent for the French test. Predict a suitable mark for him if he scored 65 in his English test.
- (ii) Another student, Jerry, was absent for the English test. Predict a suitable mark for him if he scored 70 in his French test.

**Solution:**

Since neither variable is dependent on another, the choice of regression line depends on which value is being estimated.

- (i) Here we need to predict the  $x$ -value given a  $y$ -value, so we use the least squares regression line of  $x$  on  $y$  :

$$x = 1.2832y - 8.5940 \text{ (5 s.f.)}$$

When  $y = 65$ ,  $x = 75$  (nearest whole number).

Predicted mark for his French test is 75.

- (ii) Least squares regression line of  $y$  on  $x$  :

$$y = 0.65979x + 13.031 \text{ (5 s.f.)}$$

When  $x = 70$ ,  $y = 59$  (nearest whole number).

Predicted mark for his English test is 59.

## 5 LINEARISATION OF DATA

There are many cases where the relation between the variables will be non-linear. However, through a suitable transformation on the data, it may still be possible to find a linear relation between the variables for the transformed data.

Suppose variables  $x$  and  $y$  have a quadratic relationship  $y = ax^2 + b$ .

If we let  $X = x^2$ , then the relationship becomes  $y = aX + b$ .

This is a linear relation between  $y$  and  $X = x^2$ .

### Example 10

In each of the following, variables  $x$  and  $y$  do not have a linear relationship. Identify the variables to be plotted on the axes in order to obtain a scatter diagram depicting a linear relationship.

| <u>Equation</u>                            | <u>Variables</u>      |
|--------------------------------------------|-----------------------|
| (a) $y = ax^2 + b$                         | $y$ and $x^2$         |
| (b) $y = \frac{a}{x} + b$                  | $y$ and $\frac{1}{x}$ |
| (c) $y^2 = a\sqrt{x} + b$                  | $y^2$ and $\sqrt{x}$  |
| (d) $y = ab^x$ , where $a > 0$ and $b > 0$ | $\ln y$ and $x$       |

For (d), take the natural log of both sides of  $y = ab^x$  to get:

$$\ln y = \ln(ab^x)$$

$$\ln y = \ln a + \ln b^x$$

$$\ln y = x \ln b + \ln a$$

**Example 11**

The daily rate charged by a car-hire firm varies with the length of the hire period. The firm's brochure gave the following data.

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

- (i) Find the value of the product-moment correlation coefficient, and comment on this value.  
(ii) Explain which of the following cases is the most appropriate for modelling these values.

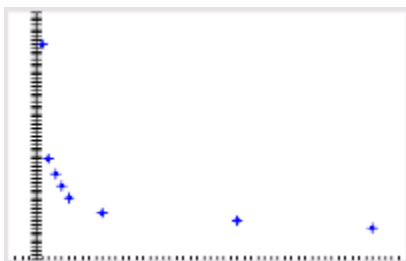
A:  $y = a + bx$ , where  $a$  is positive and  $b$  is negative,

B:  $y = a + bx^2$ , where  $a$  is positive and  $b$  is negative,

C:  $y = a + \frac{b}{x}$ , where  $a$  is positive and  $b$  is positive,

D:  $y = a + b \ln x$ , where  $a$  is positive and  $b$  is negative.

For the appropriate model, calculate the least squares regression line.

**Solution:**

- (i) The product moment correlation coefficient is  $-0.563$  (3 s.f.). There is a negative correlation between the hire period and the daily rate. Since the coefficient is not close to  $-1$ , there is no strong linear correlation between the variables, as shown in the scatter plot.

- (ii) **Remark:** Always look at the scatter diagram before proceeding to find  $r$ , in order to have an idea of the appropriateness of each model.

We see from the scatter diagram that as the length of the hire period increases, the corresponding daily rate charged decreases at a decreasing rate. This trend is consistent with Model C or D.

We thus proceed to compare the product moment correlation coefficient for Model C and Model D.

| <u>Model</u> | <u>Variables</u>      | <u>Product moment correlation coefficient</u> |
|--------------|-----------------------|-----------------------------------------------|
| C            | $\frac{1}{x}$ and $y$ | 0.992 (3sf)                                   |
| D            | $\ln x$ and $y$       | -0.817 (3sf)                                  |

Since the magnitude of 0.992 is closer to 1, then the most appropriate model is Model C.

The least squares regression line of  $y$  on  $\frac{1}{x}$  is

$$y = 99.8 + 47.1\left(\frac{1}{x}\right)$$

with coefficients correct to 3 significant figures.

**Example 12 2008 RJC/II/13 (modified)**

A stationery retailer supplies pens to offices. In order to encourage customers to buy in bulk, the stationery retailer came up with the following price scheme.

|                             |      |      |      |      |      |      |
|-----------------------------|------|------|------|------|------|------|
| Number of pens ordered, $x$ | 5    | 10   | 20   | 40   | 80   | 160  |
| Unit price in dollar, $y$   | 1.50 | 1.20 | 1.10 | 0.99 | 0.90 | 0.80 |

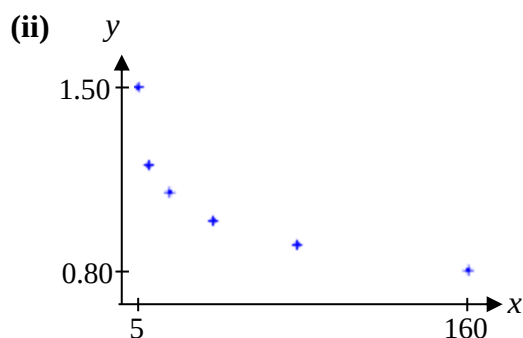
- (i) Calculate the value of the product moment correlation coefficient. [1]
- (ii) Give a sketch of the scatter diagram for the data and hence comment on the value of the product moment correlation coefficient found in (i). [3]
- (iii) State, with a reason, which model of the form  $y = a + bw$  is appropriate, where  $a$  and  $b$  are constants,  $b > 0$  and  $w$  is as follows.
- A:  $w = x^2$ ,
- B:  $w = \frac{1}{x}$ ,
- C:  $w = \ln x$ . [2]

Using the appropriate model selected above, find

- (iv) the value of the product moment correlation coefficient, [1]
- (v) the equation of the regression line of  $y$  on  $w$ , and hence obtain an estimate of the number of pens ordered when a unit price of \$1.40 is quoted. Comment on the reliability of your answer. [3]

**Solution**

- (i) Value of the product moment correlation coefficient is  $-0.807$  (3 s.f.).



The value found in (i) suggests there is some form of negative linear correlation between  $x$  and  $y$  but the scatter diagram shows that the relationship between  $x$  and  $y$  is non-linear. Thus, the value of the product moment correlation coefficient is not a good indication of the linearity between  $x$  and  $y$ .

(iii) From the scatter diagram, we see that  $y$  decreases as  $x$  increases, which is the case for model **B** ( $y = a + \frac{b}{x}$ ,  $b > 0$ ).

For models **A** ( $y = a + bx^2$ ) and **C** ( $y = a + b \ln x$ ),  $y$  increases as  $x$  increases, since  $b > 0$ . Hence model **B** is appropriate.

(iv) Value of the product moment correlation coefficient between  $y$  and  $w$  is 0.976 (3 s.f.).

(v) Equation of the regression line of  $y$  on  $w$  is  $y = 0.86652 + 3.2785w$  (5 s.f.),  
i.e.  $y = 0.867 + 3.28w$  (3 s.f.).

When  $y = 1.4$ ,  $1.4 = 0.86652 + 3.2785w$

$\Rightarrow w = 0.16272 \Rightarrow x = 6.15 = 6$  (to nearest whole number)

An estimate of the number of pens ordered when a unit price of \$1.40 is quoted is 6.

This estimate is highly reliable since  $y = 1.4$  is between the range of values of  $y$  from 0.80 to 1.50, and the product moment correlation coefficient between  $y$  and  $w$  is 0.976 (3 s.f.), which suggests a strong positive linear correlation between  $y$  and  $w$ .

### Remarks:

A near-zero value for the product moment correlation coefficient only implies the absence of a linear relation between the variables. The scatter diagram must be considered when drawing conclusions about the relation, or absence of relations between variables. A non-linear relation may still exist. Even if the product moment correlation coefficient suggests a fairly strong linear relation between the variables, a non-linear model may still provide a better fit and consequently better predictive power.

### CONCLUSION

The ideas presented in this chapter are but an introduction to the study of correlation and regression analysis on a set of bivariate data. In practice, statistical analysis is often done with the aid of computer programs. Microsoft Excel provides a simple and convenient avenue for data entry and analysis, while those seeking more rigorous analytical tools may want to consider more specialized software such as SPSS or Minitab. The study of correlation between variables and the addition of predictive power via regression analysis underpins much of scientific discovery and formation of business strategies – the same fundamental concepts are employed in the early detection of fetal abnormalities, in the study of the nature of galaxies as well as in the prediction of stock-market trends. Those wishing to take a little time off to explore further considerations in this study will find a wealth of resources online by a simple search of the chapter's title.

Many examples drawn from the study of biostatistics are available here -

[http://www.uth.tmc.edu/uth\\_orgs/educ\\_dev/osser/L3\\_1.HTM](http://www.uth.tmc.edu/uth_orgs/educ_dev/osser/L3_1.HTM)



**Appendix:****Equivalence of the 2 Formulae for the Product Moment Correlation Coefficient**

To show that

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

are equivalent, we need to show the following results:

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

**Proof:**

$$\begin{aligned} \sum (x - \bar{x})^2 &= \sum (x^2 - 2x\bar{x} + (\bar{x})^2) \\ &= \sum x^2 - 2\bar{x} \sum x + n(\bar{x})^2 \\ &= \sum x^2 - 2 \left( \frac{\sum x}{n} \right) \sum x + n \left( \frac{\sum x}{n} \right)^2 \\ &= \sum x^2 - \frac{2}{n} (\sum x)^2 + \frac{1}{n} (\sum x)^2 \\ &= \sum x^2 - \frac{1}{n} (\sum x)^2 \end{aligned}$$

The proof for  $\sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$  is similar.

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

## Summary