NANYANG JUNIOR COLLEGE
DEPARTMENT OF MATHEMATICS

CHAPTER S8: Correlation and Regression

[Prepared by: Mr Leong Chong Ming & Mr Chen Jianda Alvin]

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

- use a scatter diagram to judge if there is a plausible linear relationship between two variables
- calculate and interpret the product moment correlation coefficient as a measure of the fit of a linear model to the scatter diagram
- find the equation of the least-squares regression line
- use the appropriate regression line to make prediction or estimate a value in practical situations (interpolation and extrapolation), and explain how well the situation is modelled by the linear regression model
- use a square, reciprocal or logarithmic transformation to achieve linearity

---

My Notes

8.1 Terminology

8.1.1 Dependent and Independent Variables

In statistics, **bivariate data** is data that has two variables. The following table shows 4 sets of bivariate data x and y. Is there any relationship between each pair of variables?

x	y
Student's Math Score	Student's Physics Score
Age of a tree	Trunk circumference
Length of leg of a student	Distance jumped by a student
Monthly advertising expenditure	Monthly sales

In a bivariate pair, an **independent variable** is a variable whose variation does not depend on that of another. In an experiment, the independent variable would be the one which you have "control" over, thus allowing you to vary its value to determine the value of the corresponding **dependent variable**.

It may be possible to identify *from context* the independent variable and the dependent variable. For example, if we are investigating how the length of a pendulum will affect its period of swing, then the length of the pendulum will be the **independent** variable and its period will be the **dependent** variable.

The following table shows another set of bivariate data.

x	10	20	30	40	50	60	70	80
y	20	21	23	24	23	25	28	29

The values of x varies at a *fixed increment*. This is another indication that x is likely to be the **independent** variable (controlled) and y the **dependent** variable. If context is provided, the fact that x increases at fixed intervals would give further evidence that it is the independent variable. Usually, the independent variable will be listed in the first row of the table.

Example 1

(a) A botanist wishes to investigate the effects of artificial light on the growth of a certain type of plant. A random sample of 7 seeds of the plant species is planted in 7 different pots and the amount of artificial light per day given to each pot is varied. The following results are recorded after a period of 14 days:

Amount of artificial light per day (l hours)	4	8	10	13	14	17	18
Height of plant (h cm)	6	8	10	11	14	14	15

State the independent variable and the dependent variable (if any), justifying your choice.

Solution:

The variable l is the independent variable and the variable h is the dependent variable since the investigation is on how the height of the plant, h cm, depended on the amount of artificial light per day, l cm, it received.

(b) Eight newly-born babies were randomly selected. Their head circumference, x cm, and body length, y cm were measured by the pediatrician and tabulated.

x	31	32	33.5	34	35.5	36	36.5	37.5
y	45	49	49	47	50	53	51	51

State the independent variable and the dependent variable (if any), justifying your choice.

Solution:

There is no independent and dependent variable as it is not clear from context if head circumference is dependent on body length or vice versa.

8.1.2 Scatter Diagram

A **scatter diagram**, in which pairs of data (x_i, y_i) are plotted, can be used to represent bivariate data graphically. The independent variable, if applicable, is plotted on the horizontal axis.

Example 2

Seven students from a class were selected. The number of days of absence and their corresponding final grades were recorded as follows:

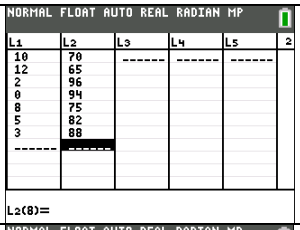
Student	*A*	*B*	*C*	*D*	*E*	*F*	*G*
No. of days of absence, *x*	10	12	2	0	8	5	3
Final grade, *y* %	70	65	96	94	75	82	88

Sketch a scatter diagram for the data.

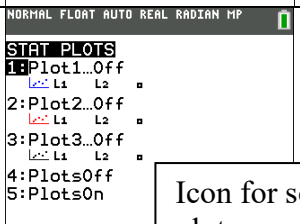
SOLUTION:

Step 1: Press [stat] and select **1:Edit**.
Key in the values of number of days of absence, *x*, in list L_1 and final grade, *y* %, in L_2.

THINKZONE:



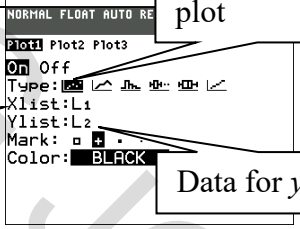
Step 2: Press [2nd] [y=] for **[STAT PLOTS]**.
Select **1:Plot1**



Icon for scatter plot

Step 3: Notice that **Plot1** is highlighted and we can turn it on by positioning the cursor over ON and press [enter].

Scroll down and select the icon for scatter plot at **Type:** and press [enter].



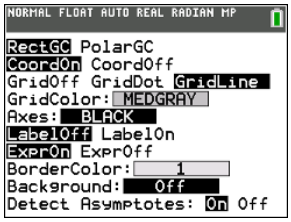
Data for *x*-axis

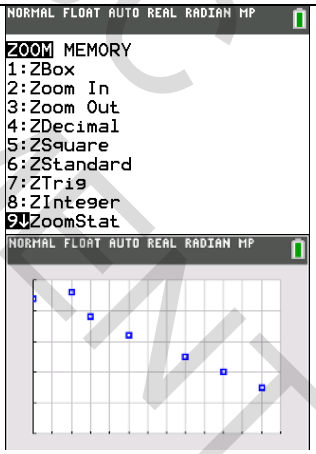
Data for *y*-axis

Step 4: Press [zoom] and select **9:ZoomStat**.

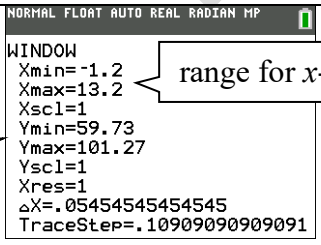
Note:

- To read the coordinates of each point in the scatter diagram, press [trace] and use the cursor control arrow keys to move from point to point.
- To add gridlines to the scatterplot, press [2nd] [zoom] and select GridLine.



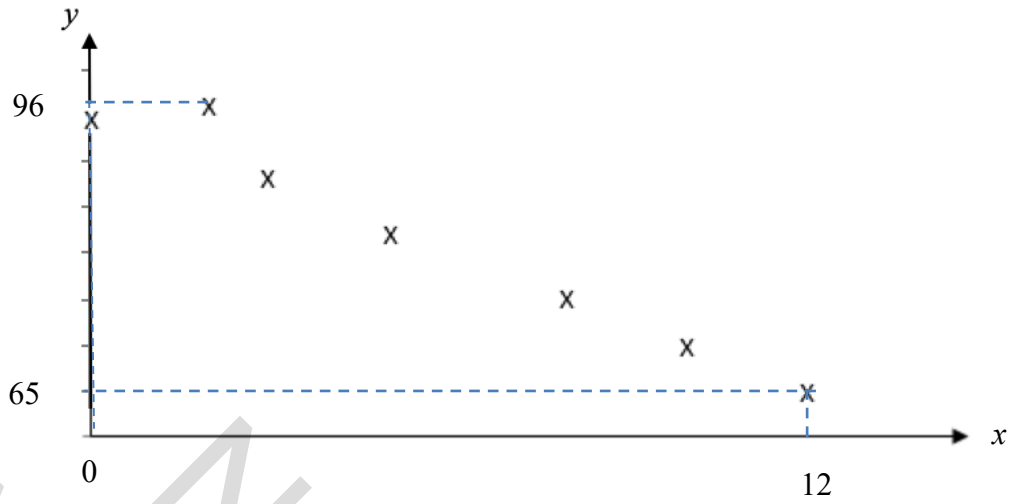


Step 5: Press [window] to see the range of *x* and *y* used by the calculator.



range for *x*-axis

range for *y*-axis



Remarks:
For scatter diagram, you need to

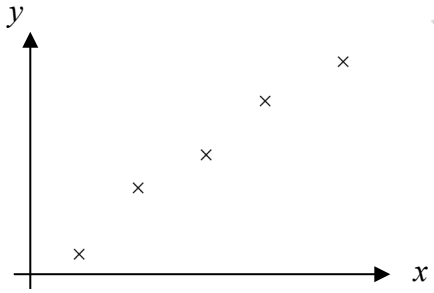
- label the axes (it suffices to label the smallest (need not start from 0) and largest values for each axis),
- draw to scale (Use GC to help gauge the range of values – see step 5 above),
- use a cross '×' to mark the data points.

A scatter diagram shows the general pattern and relationship between two variables. It is often used to determine if there is a linear relationship between two variables.

To interpret scatter diagrams, we can comment on the following features:

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

Positive linear relationship

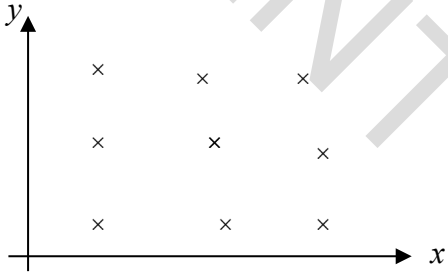


Negative linear relationship

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

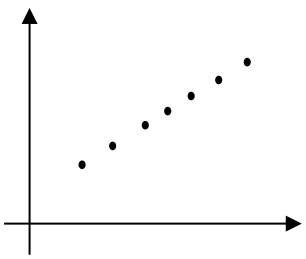
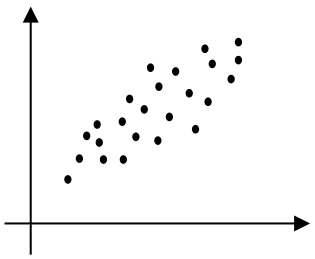
Non-linear relationship (or curvilinear relationship)

No clear relationship



There is **no** clear relationship between the variables as all the points appear to be randomly scattered, with no discernible pattern.

2. Strength

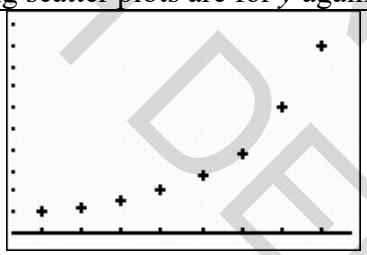
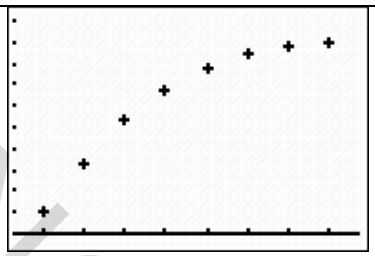
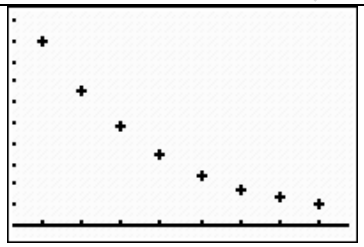
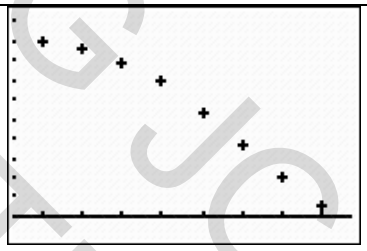
	
If the data points are close to a straight line, there is a **strong** linear correlation between x and y.	If the data are not clustered near a straight line, then the linear correlation between x and y may not be strong.

3. Direction: The two variables x and y can be positively correlated, negatively correlated or not correlated at all. This relationship is usually described in the context of the situation.

- If y generally increases as x increases, then x and y are **positively** correlated.
- If y generally decreases as x increases, then x and y are **negatively** correlated.

Suggested interpretation of a scatter plot with curvilinear trend

The following scatter plots are for y against x.

 As x increases, y *increases* at an *increasing* rate.	 As x increases, y *increases* at a *decreasing* rate.
 As x increases, y *decreases* at a *decreasing* rate.	 As x increases, y *decreases* at an *increasing* rate.

Outliers

An **outlier** is an observation that lies outside the overall pattern of the data. Outliers may arise due to error in experiment or data capture, or simply reflect anomalies that had occurred naturally. A scatter diagram can help to identify outliers.

DIAGRAMS:

Without any information on any underlying function or principle that governs the bivariate data, the scatter plot provides us an overview of the data points. A well chosen range on the (independent) variable can shed light on possible pattern or trends in the variables.

Self-Review 1:

The table below shows the ages in years (x) and the trunk circumferences in cm (y) of a random sample of 7 trees of a particular species.

Age (x)	11	13	21	28	34	42	45	51
Circumference (y)	24.4	32.1	52.8	78.3	79.2	25.1	102.7	121.2

Give a sketch of the scatter diagram for the data, as shown on your calculator.
One of the points appears to be wrongly recorded. Indicate the corresponding point on the diagram by labelling it P.

SOLUTION:	**THINKZONE:**
	• Are the axes labelled? It suffices to label the smallest (need not start from 0) and largest values for each axis. • Is the diagram drawn to scale? Use GC to help gauge the range of values. • Are the data points marked with a cross '×'? Note: Point P is called an outlier.

8.2 Linear Correlation and Product Moment Correlation Coefficient, *r*

In statistics, **correlation** is the measure of relationship between 2 variables, and this is represented by a number called the correlation coefficient. For our syllabus, we will focus on **linear correlation**.

The **product moment correlation coefficient**, ***r***, which indicates the **linear degree** of scatter, is a measure of the **linear relationship** between 2 random variables X and Y.

The **product moment correlation coefficient**, ***r***, is defined by:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\left(\sum (x-\bar{x})^2\right)\left(\sum (y-\bar{y})^2\right)}} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left\{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right)\right\}}} \quad \text{[in MF 27]}$$

Note:

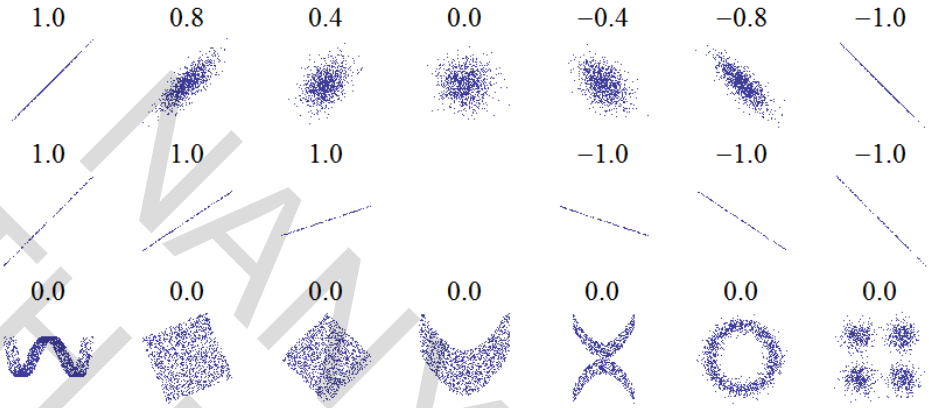
(i) The value of r ($-1 \le r \le 1$) indicates the **strength** and **direction** of the linear relation.

$r = 1$	A **perfect positive linear** relationship between x and y.	All the data points lie on a straight line with positive gradient.
$r > 0$	A **positive linear** relationship between x and y.	As x increases, y increases.
$r = 0$	**No linear relationship** between x and y.	Note that no linear relationship does not imply no relationship as there could be a nonlinear relationship.

$r < 0$	A **negative linear** relationship between x and y.	As x increases, y decreases.
$r = -1$	A **perfect negative linear** relationship between x and y.	All the data points lie on a straight line with negative gradient.

(ii) The nearer the value of r is to 1 or –1, the closer the points on the scatter diagram are to a certain straight line.

(iii) The following diagram shows different scatter diagrams with varying values of r.

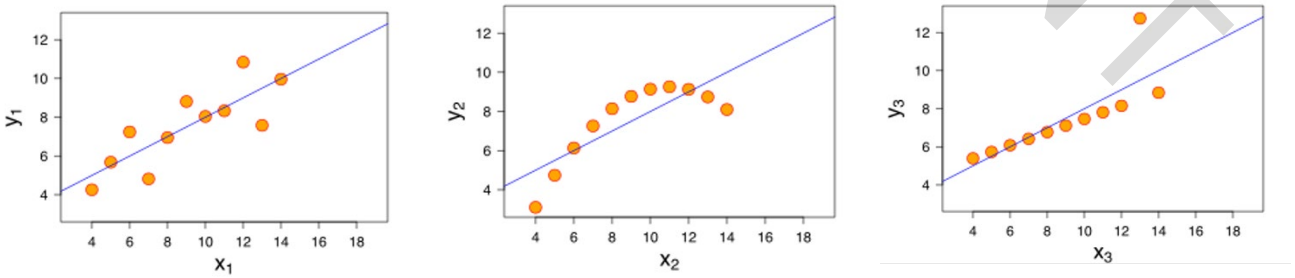


(iv) The following table gives a suggested range of values of r and its corresponding interpretation for two variables x and y.

Value of r	**Interpretation of linear correlation**
1	Perfect positive linear correlation
$0.8 \le r < 1$	Strong positive linear correlation
$0.6 \le r < 0.8$	Moderate positive linear correlation
$0 < r < 0.6$	Weak positive linear correlation
0	No linear correlation
$-0.6 < r < 0$	Weak negative linear correlation
$-0.8 < r \le -0.6$	Moderate negative linear correlation
$-1 < r \le -0.8$	Strong negative linear correlation
–1	Perfect negative linear correlation

(v) A strong linear correlation between two variables does not necessarily imply that one variable directly causes the other.

(vi) The value of r should always be interpreted together with a scatter diagram where possible. The value of r can also be affected by outliers and can give a misleading conclusion on the linear correlation of two variables. For example, the following 3 diagrams have the same value of $r = 0.816$ but appeared differently when graphed.



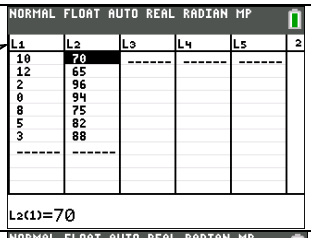
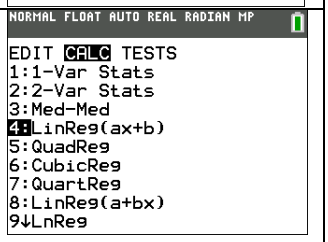
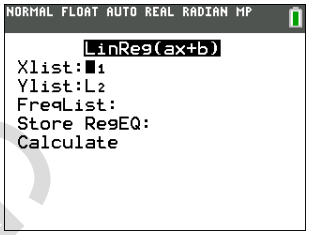
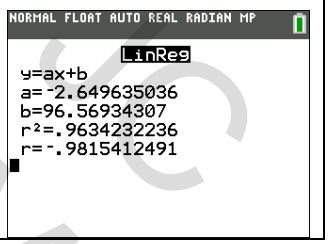
Using Graphing Calculator to obtain the Product Moment Correlation Coefficient

Example 3

Seven students from a class were selected. The number of days of absence and their corresponding final grades were recorded as follows:

Student	1	2	3	4	5	6	7
No. of days of absence, x	10	12	2	0	8	5	3
Final grade, y %	70	65	96	94	75	82	88

Calculate the product moment correlation coefficient between x and y.

GC KEYSTROKES:	**SCREEN:**
Enter the data into the calculator : **Step1**: Press [stat] and select **1:Edit**. Key in the values of x in list L_1 and y in L_2. Put the independent variable in L_1	
Step 2: Press [stat], select the **CALC** menu and select **4: LinReg(ax+b)**. **Press** [enter] • Key in L_1 into the Xlist: • Key in L_2 into the Ylist: Scroll down to **Calculate** and press [enter] . **Note:** Xlist and Ylist are the list of values that the GC will use for the horizontal and vertical axes respectively. The choice of L_1 in Xlist and L_2 in Ylist or L_2 in Xlist and L_1 in Ylist is important for the regression (best-fit) line which we will discuss in the next section. However, either choice will give the same r value. (Why is this so?) (To see the value of r, you need to turn Stat Diagnostic ON. By default, Stat Diagnostic would be switched on when you reset your GC. Press [mode] and scroll down to STAT DIAGNOSTICS and select ON to turn it on otherwise.)	  

SOLUTION:
Using GC, $r = -0.982$ (3 significant figures)

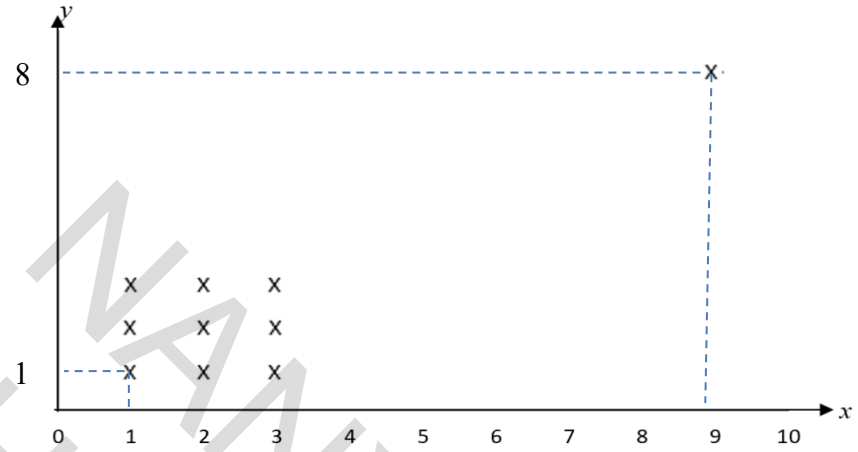
This suggests a strong negative linear relationship between the number of days of absence and the students' final grade.

Example 4

Given the following set of bivariate data, draw a scatter diagram and calculate the product moment correlation coefficient between x and y.

x	1	1	1	2	2	2	3	3	3	9
y	1	2	3	1	2	3	1	2	3	8

SOLUTION:



Product moment correlation coefficient, $r = 0.862$

Note:

The point (9, 8) is an **outlier**. Without this point, we will have $r = 0$ (check for yourself using the GC). This example shows that the product moment correlation coefficient by itself is insufficient to make a conclusion on the linear correlation (if any) between two variables. A scatter diagram should be used to visually determine if there is any linear correlation **before** the product moment correlation coefficient is calculated to quantify the correlation.

When interpreting the relationship between two variables, we may be asked to discuss the relationship based on -

- a scatter diagram only, or
- the value of r only, or
- both a scatter diagram and the value of r.

8.2.1 Scaling and/or Translation of Variables

It is not uncommon to rewrite one (or both) of your variables in a different form. For example, the data collected may have been captured with a specific unit of measurement, and needs to be rewritten in a different unit of measurement.

If the transformation required involves scaling and/or translation, the new set of transformed data will have the **same** product moment correlation coefficient. The following table shows some examples of a change in units that employs such transfomations:

Original units, x_1	**New units, x_2**	**Conversion**
Kilogram (kg)	Gram (g)	$x_2 = 1000x_1$
Degree Celsius (°C)	Degree Fahrenheit (°F)	$x_2 = \frac{9}{5}x_1 + 32$
Metre (m)	Foot (ft)	$x_2 = 3.28084x_1$

Example 5 [N2000/II/6]

The amount, x grams, of catalyst used in a chemical reaction and the corresponding time, y hours, taken to complete the reaction were recorded. The results are given in the following table.

x	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	62.1	51.2	44.1	39.2	35.0	37.3	33.0

Note: From context, the variable x (amount of catalyst used in a chemical reaction) increases at fixed intervals and is the independent variable. The variable y (resulting time taken to complete the reaction) is the dependent variable.

(a) Calculate the value of the linear (product moment) correlation coefficient for the data.

(b) State what your value indicates about the relation between x and y.

(c) Plot a scatter diagram for the data and explain how its shape is related to your answer to part (b).

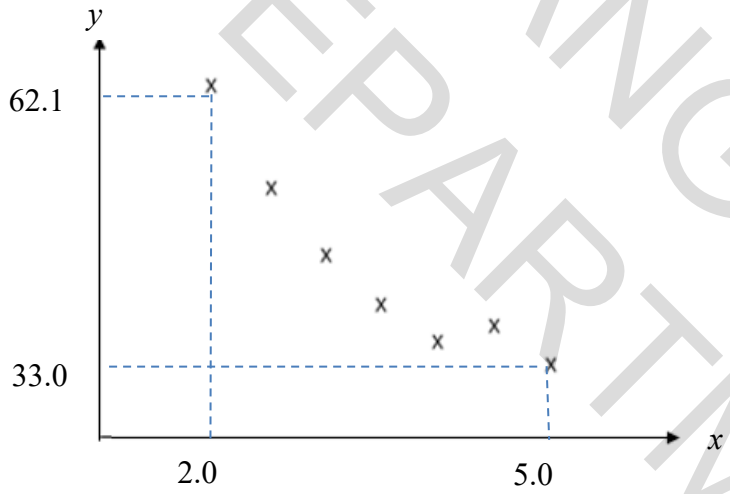
(d) Without doing further calculation, how would the value of linear (product moment) correlation coefficient change if **(i)** the amount of catalyst is recorded in kilogram,
(ii) each value of x is increased by 1.0 grams.

SOLUTION:

(a) The product moment correlation coefficient, $r = -0.925$ (3 s.f.)

(b) The value of r suggests that there is a strong negative linear relationship between the amount of catalyst used and the time taken to complete the reaction.

(c)



Since r is close to -1, it suggests a linear model is appropriate. However, the scatter diagram shows that the relationship is non-linear: as the amount of catalyst (x grams) used increases, the time taken to complete the reaction decreases at a decreasing rate. The data point at (4.5, 37.3) appears to be an outlier.

(d) There will be no change in the value of r in both cases.

(i) A change in the units for the variable corresponds to a scaling and the degree of linearity is not affected by scaling.

(ii) Increasing each value of x by the same value corresponds to a translation and the degree of linearity is not affected by translation.

8.3 Linear Regression

Correlation is used to decide if there is a relationship between two variables and to determine how strong such a relationship is. **Regression** is used to find a model for the relationship and to estimate values of one variable given values of the other variable. After obtaining the scatter diagram, we may be interested to look for a mathematical relationship $y = \mathrm{f}(x)$, i.e. by using the given points only, we have to 'work backwards' or 'regress' to find a function f that fits the given data. Hence the function is called the **regression function**.

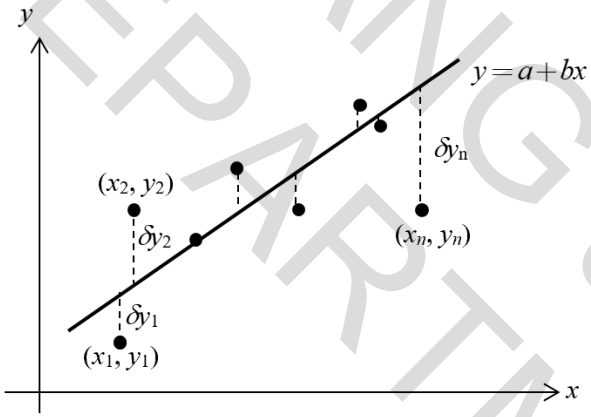
If the scatter diagram and the linear product moment correlation coefficient between two variables gives a good indication that it is meaningful to model the observed data with a straight line, we may then attempt to fit a linear model in the form of a regression line, i.e. $y = a + bx$ for some real constants a and b.

8.3.1 Method of Least Squares

The **method of least squares** is the simplest and most applied form of linear regression to find the best fitting straight line through a set of data points. A best fit line obtained using this method is called a **least-squares regression line**. Two possible best fit lines can be found – the **least-squares regression line of *y* on *x***, and the **least-squares regression line of *x* on *y***.

8.3.1.1 Least-Squares Regression Line of *y* on *x*

Let (x_1, y_1), $(x_2, y_2) \ldots (x_n, y_n)$ be a set of n observed data points and let $y = a + bx$ be the **least-squares regression line of *y* on *x***.



Define **residuals** (or **errors**), δy_i, $i = 1, 2, \ldots, n$, as the deviation between the actual and fitted (predicted) values, i.e. the vertical distances drawn from each observed data point to the regression line with

$$\delta y_i = \text{observed value} - \text{predicted value} = y_i - (a + bx_i).$$

When we fit a line through the data, some of the actual values will be larger than their predicted values (they fall above the line and $\delta y_i > 0$) and some of the actual values will be smaller than their predicted values (they fall below the line and $\delta y_i < 0$).

To measure overall error and avoid cancellation of errors, we square the errors and find a line that minimises the sum of the squared errors, $\sum_{i=1}^{n} (\delta y_i)^2 = \delta y_1^{\,2} + \delta y_2^{\,2} + \ldots + \delta y_n^{\,2}$.

The **least-squares regression line of *y* on *x*** (or simply **regression line of *y* on *x***) is obtained by finding the values of a and b in $y = a + bx$ such that $\sum_{i=1}^{n} (\delta y_i)^2$ is the least. Note that we are minimising the sum of the squared ***vertical*** deviations of the data points from the regression line.

We may use the graphing calculator to obtain the values of a and b. These values are called the **least squares estimates of *a* and *b***.

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

$$y - \bar{y} = b(x - \bar{x}) \text{ where } b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \qquad \text{[in MF 27]}$$

Remarks:

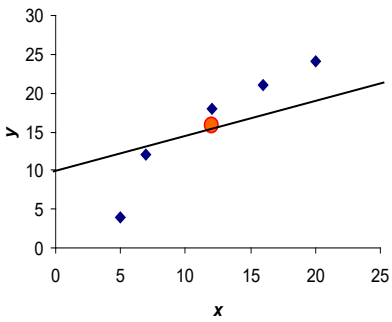
1. The regression line of y on x ***always*** passes through the point $(\bar{x}, \bar{y})$.
2. The gradient, b, of the regression line of y on x, is called the **Coefficient of Regression of *y* on *x***.
3. The y-intercept of the above equation of the regression line of y on x (i.e. $y = a + bx$) is a.
4. We need to know how to interpret the least squares estimates of a and b in context.

To illustrate the method of least squares in finding the regression line of y on x, consider the data set:

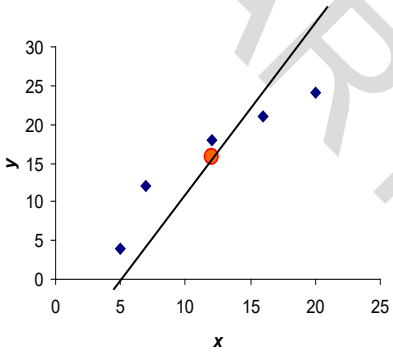
x	5	7	12	16	20
y	4	12	18	21	24

Three regression lines, each passing through $(\bar{x}, \bar{y}) = (12, 15.8)$, are drawn to fit the same set of data points and their respective sum of squared residuals, $\sum_{i=1}^{n} (\delta y_i)^2$, are computed as follows:

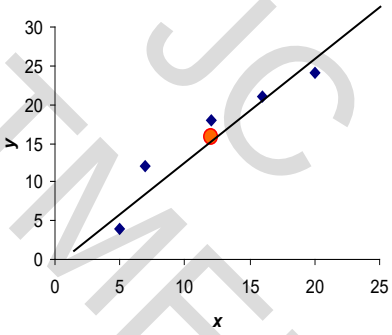
$\sum_{i=1}^{n} (\delta y_i)^2 = 107.04$



$\sum_{i=1}^{n} (\delta y_i)^2 = 188.70$



$\sum_{i=1}^{n} (\delta y_i)^2 = 23.29$



The third line has the smallest value of $\sum_{i=1}^{n} (\delta y_i)^2$. If this value is the *least among all the regression lines* drawn to fit the data points, then this line is the least-squares regression line of y on x.

Using Graphing Calculator to obtain Regression Line of *y* on *x*

Example 6

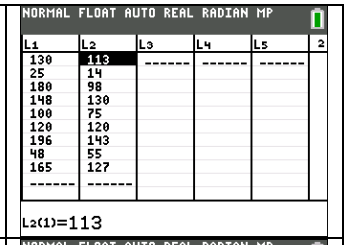
Delegates who travelled by car to a conference were asked to report the distance they travelled and the time taken. The table below shows the data collected.

Distance, y (km)	113	14	98	130	75	120	143	55	127
Time, x (min)	130	25	180	148	100	120	196	48	165

(a) Obtain the equation of the regression line for y on x.
(b) Sketch the scatter diagram with the regression line for y on x.
(c) Estimate the distance travelled (to the nearest whole number) if a delegate took 150 minutes for the journey.
(d) Interpret the coefficient of regression of y on x in the context of the question.

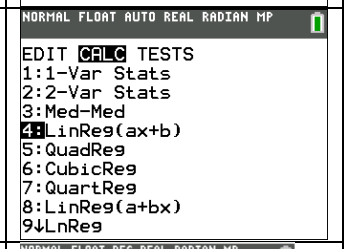
SOLUTION:

Step 1: Press [stat], select 1:Edit.
Key in the values of x in list L_1 and y in L_2.



Step 2: Press [stat], select **CALC** and scroll down to select **4:LinReg(ax+b)**.

(Remark: 8:LinReg(a+bx) will also give the same regression line, but the role of **a** and **b** are swapped.)

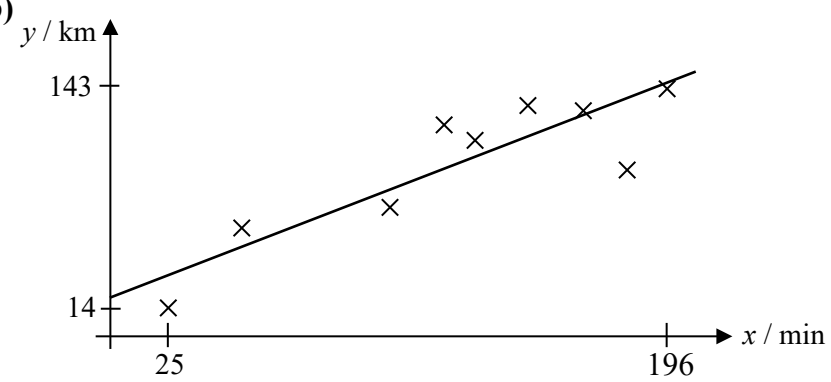


Step 3: Key in **L_1** into the Xlist
Key in **L_2** into the Ylist
Key in **Y_1** into StoreRegEQ
(Y_1 is selected by pressing [alpha] [trace])
Scroll to **[Calculate]** and press [enter] to obtain the regression line y on x

Note: Xlist and Ylist are the list of values that the GC will use for the horizontal and vertical axes respectively.
We plot the regression line of y on x on a scatter diagram with the usual x- and y-axes. Hence we select L_1 (x values) for Xlist and L_2 (y values) for Ylist.

(a) The equation of the regression line for y on x is $y = 0.647x + 17.3$

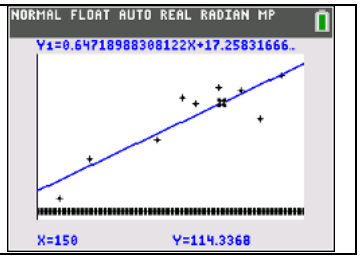
(b)



To obtain the scatter diagram with the regression line, make sure that the STAT PLOT 1 is turned on and press [ZOOM] [9] to get the plot.)

(c) By GC, when $x = 150$, $y = 114$

Press [2nd] [trace] and select '**1:value**' and press [enter]
Key in the desired x value,
i.e. $x = 150$ in this case, and press **[ENTER]**.
The GC will display the value Y = 114.35 as shown



Alternatively (by calculation), $y = 0.64719(150) + 17.258 \approx 114$

(d) The coefficient of regression of y on x (i.e the coefficient of x) is 0.647. This means that for every 1 minute increase in the time taken, the distance travelled is **expected** to increase by 0.647 km.

$y = 0.647x + 17.3$, from the regression line y on x, when x increases by 1 unit, y increases by 0.647 units. The key point here is to write in the **context of time (x) vs distance travelled (y)**.

Example 7

The following table shows the ages, x, and the systolic blood pressure, y, of 8 men.

Age (x)	56	42	72	36	63	47	55	60
Blood pressure (y)	147	125	160	118	a	128	150	149

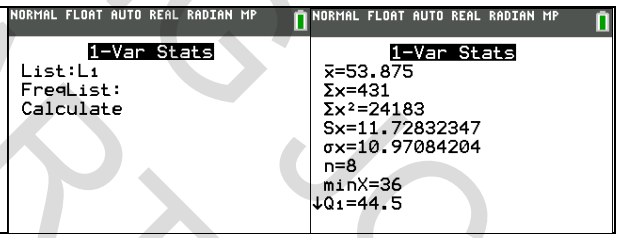
Given that the linear regression line of y on x is $y = 1.254x + 73.584$. Show that the value of a is 152, correct to the nearest integer.

SOLUTION:

Using GC, sample mean is $\bar{x} = 53.875$.

Key in the values of x in the list L_1.
Press [stat], highlight '**CALC**' and select '**1:1-Var Stats**' and press [enter]

By GC, $\bar{x} = 53.875$



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

$$\bar{y} = 1.254\bar{x} + 73.584$$
$$= 1.254(53.875) + 73.584$$
$$= 141.14325$$

Hence, $\bar{y} = \dfrac{147 + 125 + 160 + 118 + a + 128 + 150 + 149}{8} = 141.14325$

$\therefore$ $a = 152$ (correct to the nearest integer)

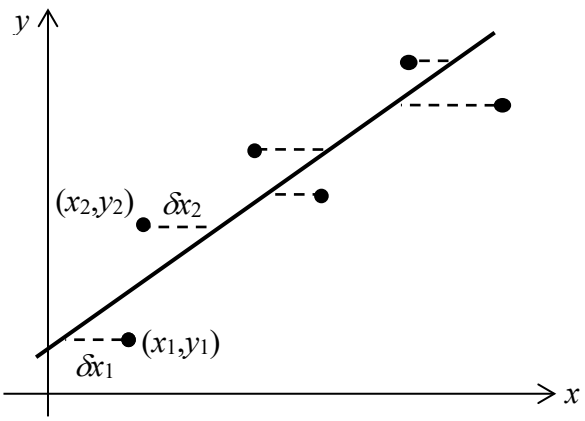
Q: Why can't we simply substitute $x = 63$ into the equation of the regression line?
A: Since $x = 63$ is an observation, it might NOT lie on the regression line.

Important Note: In general, the ***regression line may not pass through the data points***. Thus if there is a missing data, we **CANNOT use the regression line to determine the missing data**. We must make use of $(\bar{x}, \bar{y})$ to find the missing data.

8.3.1.2 Least-Squares Regression Line of x on y

Given a set of n data points (x_1, y_1), $(x_2, y_2) \ldots (x_n, y_n)$, the **least-squares regression line of x on y** (with equation $x = c + dy$) is such that $\sum_{i=1}^{n} (\delta x_i)^2$ is the *least.*

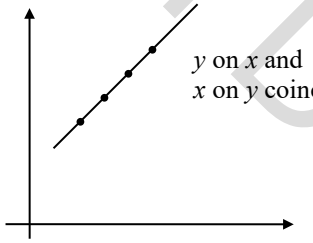
Note that we are minimising the sum of the squared *horizontal* deviations of the data points from the regression line.



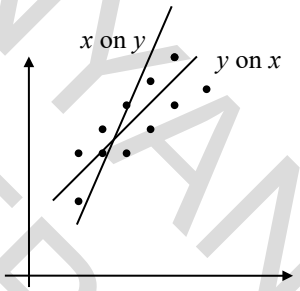
Remarks:

1. Both the least-squares regression line of y on x and the least-squares regression line of x on y pass through the point $(\overline{x}, \overline{y})$.
2. In general, the regression line of x on y is different from the regression line of y on x.

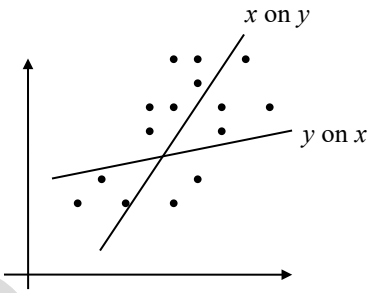
3. The nearer the value of r is to 1 or –1, the closer the two regression lines y on x and x on y are on a scatter diagram. This is illustrated in the following diagrams.



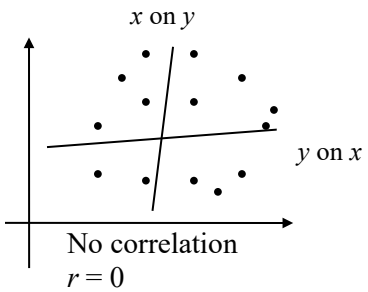
Perfect positive correlation $r = 1$



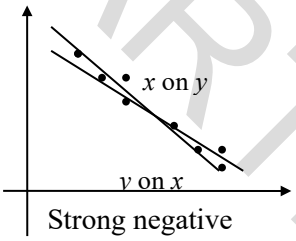
Strong positive correlation $r = 0.8$



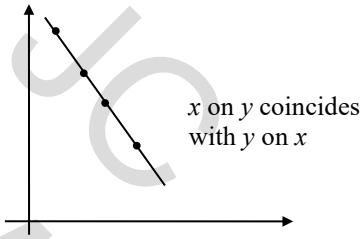
Some positive correlation $r = 0.5$



No correlation $r = 0$



Strong negative correlation $r = -0.9$



Perfect negative correlation $r = -1$

4. To plot the least-squares regression line of x on y with equation $x = c + dy$ on a scatter diagram, where the x-values are on the horizontal axis and the y-values are on the vertical axis, it is necessary to make y the subject first as $y = \frac{1}{d}x - \frac{c}{d}$ before keying this equation into the GC.

5. We always use regression line of (dependent variable) on (independent variable) regardless of which variable we are predicting/estimating. Only when there is no independent variable (unclear from context), we use the line of (variable to predict) on (given variable).

x	***y***	**Purpose**	**Equation of Regression Line**	**Remarks**

Independent	Dependent	Predict y given x	y on x	Use the same regression line *regardless* of the variable we are predicting/estimating
		Predict x given y	y on x	
No clear relationship between x and y		Predict y given x	y on x	Treat x as the independent variable
		Predict x given y	x on y i.e. $x = c + dy$	Treat y as the independent variable

Using Graphing Calculator to obtain Regression Line of x on y

Just as in the case of finding the regression line of y on x, we can key in x and y values into L_1 and L_2 respectively. However, to find the regression line of x on y, we key in L_2 (y values) into Xlist (horizontal axis) and L_1 (x values) into Ylist (vertical axis). Note that the equation of the regression line obtained is of the form $x = c + dy$.

Example 8

Delegates who travelled by car to a conference were asked to report the distance they travelled and the time taken. The table below shows the data collected.

Distance, y (km)	113	14	98	130	75	120	143	55	127
Time, x (min)	130	25	180	148	100	120	196	48	165

(a) Obtain the equation of the regression line of x on y in the form $x = \alpha y + \beta$ where α and β are real constants to be determined to 2 decimal places. Give interpretations for the least squares estimates of α and β and comment on these values.

(b) Sketch the scatter diagram (with x on the horizontal axis) with the regression line of x on y.

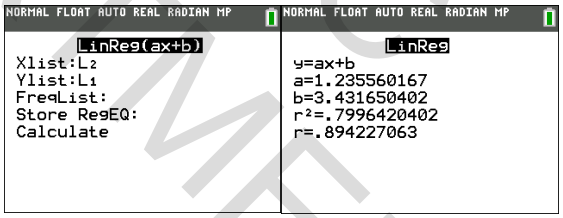
(c) Estimate the time taken for a delegate to travel 80 km for the journey to one decimal place.

SOLUTION:

(a)

In example 6, we keyed in the values of x in **L1** and y in **L2**.

To find regression x on y:
Key in **L_2** into the **Xlist:**
Key in **L_1** into the **Ylist:**
Press **[Calculate]** obtain the regression line x on y.
(Note that **L_2** (which contains the y values) is now in Xlist).

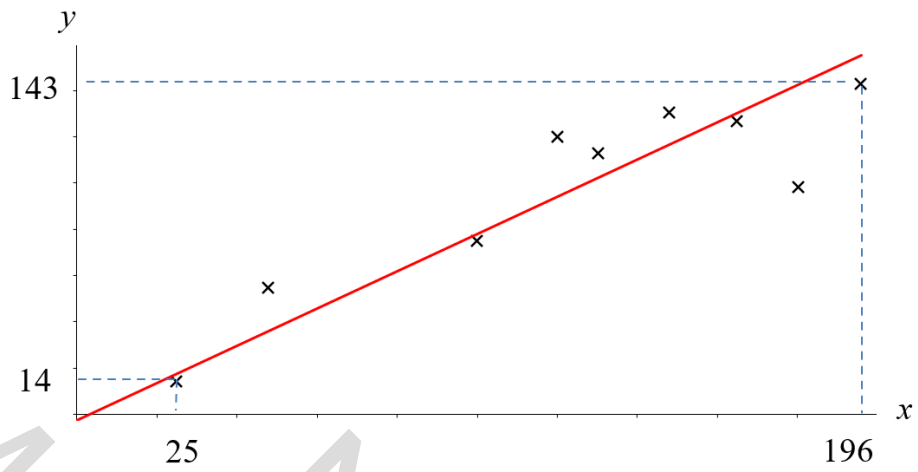


By GC, the equation of the regression line of x on y is $x = 1.24y + 3.43$ (to 2 d.p.).

The value $\alpha = 1.24$ means that with every 1 km increase in the distance travelled, the time taken is expected to increase by 1.24 mins.

The value $\beta = 3.43$ is the expected time taken (in mins) to travel zero distance ($y = 0$), which is absurd since $y = 0$ lies outside the data range and so we do not extrapolate.

(b) By GC, the equation of the regression line of x on y is $x = 1.24y + 3.43 \Rightarrow y = \dfrac{x - 3.43}{1.24}$.

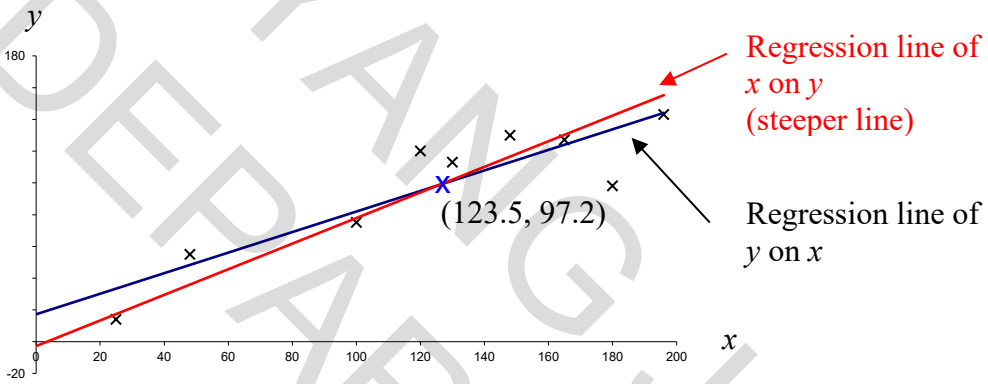


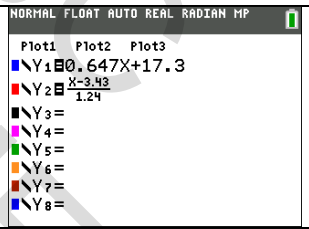
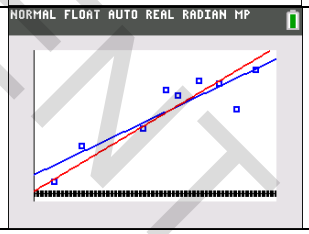
We need to write y in terms of x in sketching the regression line of x on y since the y values are on the vertical axis while the x values are on the horizontal axis.

As there is no clear relationship between x and y, to predict x given y, we will use the line x on y.

(c) When $y = 80$, estimated time taken $x = 1.2356(80) + 3.4317 = 102.2797 = 102.3$ (1 d.p)

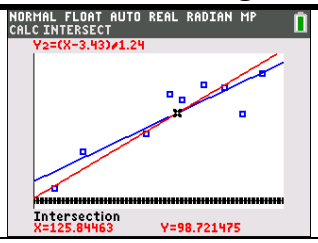
Remark: In **Examples 6** and **8**, the same data set were used and the regression line of y on x and the regression line of x on y are drawn on the same x-y plane below with x on the horizontal axis and y on the vertical axis. Observe that the two lines are **different** and they **intersect** *at* $(\bar{x}, \bar{y})$.



Press [y=] and key in $\dfrac{x-3.43}{1.24}$ at $\mathbf{Y_2} =$	
Press [zoom] and select **9:ZoomStat.** The two regression lines can be plotted on the same diagram as shown.	

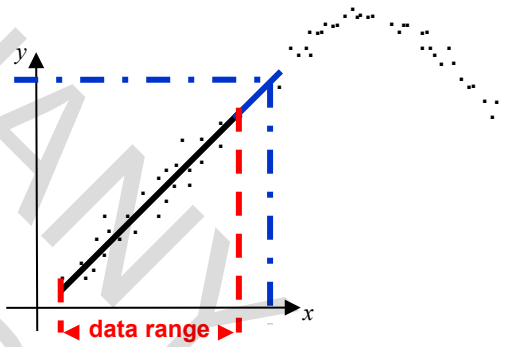
To find the point of intersection of these two lines:

Press 2nd trace for **[CALC]** and select **5: intersection**



8.3.2 Interpolation Vs Extrapolation

Interpolation is an estimation of a value within the data range. **Extrapolation** is an estimation of a value outside the data range. In general, we should NOT do extrapolation as there is no way to check if the relationship between x and y continues to hold for values that are outside the given data range.



You would be asked to estimate the value of one variable given the other and to comment on the reliability of the estimate.

For **interpolation**, there are two reasons to justify that the estimate is *reliable*: r is close to 1 (or –1), which suggests that there is a strong positive (or negative) linear relationship between the two variables, **and** the given value lies within the data range.

For **extrapolation,** we say that the estimate is *unreliable* as the given value lies outside the data range, thus the estimate may not be reliable due to extrapolation.

BIG IDEA

EXTENSIONS:

Extrapolation doesn't bode well with regression lines because the best fit lines we create are based on the data collected, so it only represents the behaviour in the region/domain of the data collected.

Example 9 [N2000/FM/II/10 (modified)]

An experiment with certain swimming animals was carried out to investigate how the speed at which they swam depended on the angle through which their feet moved. The angle θ^o through which the hind feet moved was measured, together with the swimming speed v ms^{-1}. The results are given in the table.

θ	87	92	96	97	98	101	110	114	115	115	116	123	133
v	0.35	0.30	0.50	0.40	0.35	0.45	0.60	0.55	0.55	0.65	0.50	0.70	0.75

(a) State, giving a reason, which of the least-squares regression lines, θ on v or v on θ, should be used to express a possible linear relation between v and θ.

(b) Calculate the equation of the line chosen in part **(a)** and give the values of the coefficients to a suitable accuracy.

(c) Calculate the product moment correlation coefficient.

(d) Estimate

(i) the swimming speed of the animal when the angle through which its hind feet moved is 70°,

(ii) the angle through which the animal's feet moved when its swimming speed is 0.32 ms^{-1}.

(d) Comment briefly on the reliability of the estimates in part **(d)**.

SOLUTION:	**THINKZONE:**
(a) The regression line v on θ should be used since the experiment investigated how the speed v at which the animals swam depended on the angle θ through which their feet moved, i.e θ is the independent variable.	
(b) The regression line is $v = 0.0095084\theta - 0.51025$ (5 s.f.) $v = 0.01\theta - 0.51$ (2 d.p.)	For **(b)** and **(d)**, we leave the final answer in 2 decimal places as it was the number of decimal place as given in the question, and the question does not specify the required accuracy.
(c) $r = 0.910$ (3 s.f.)	
(d) (i) When $\theta = 70$, $v = 0.0095084(70) - 0.51025$ $= 0.15534$ (5 s.f.) $= 0.16$ (2 d.p.)	
(ii) When $v = 0.32$, $\theta = \dfrac{0.32 + 0.51025}{0.0095084}$ $= 87.318$ (5 s.f.) $= 87$ (nearest whole number)	The regression line in **(b)** is still used for **(d)(ii)** because θ is the independent variable.
(e) The estimate in (i) is unreliable as $\theta = 70$ is outside the data range of θ. The estimate in (ii) is reliable as $v = 0.32$ is within the data range of v, and r is close to 1.	

Remarks on Accuracy

Do note that we will usually adopt the same number of significant figures as given in the question (data values of x or y) for our answers if the question does not specify the required accuracy.

Self-Review 2:

A study is done to investigate whether children who spend more time on reading tend to have higher English Literacy scores than children who do not read sufficiently. The table below shows the number of hours spent on reading per week of a sample of 8 children taken when they were 7 years together with the English Literacy scores determined through a test at age 9 years.

Hours spent on reading per week (x)	12	15	8	10	9	6	7
English Literacy score (y)	85	95	73	79	76	68	70

(a) Find the product moment correlation coefficient and the equation of the estimated regression line of y on x.

(b) Estimate the English Literacy score of a child who spends 11 hours reading per week.

(c) Estimate the hours spent on reading per week at age 7 years given an English literacy score of 78 at age 9 years to 2 decimal places using the regression line of

(i) x on y,

(ii) y on x obtained in **(a)**.

Comment on the values obtained in **(i)** and **(ii)**.

Solution:

8.4 Transformation of Data Points from Non-Linear to Linear

Two variables may be governed by a relation that is non-linear. For example, the equation $y = a + bx^2$ indicates that y and x are related by a quadratic relationship. However, we can view this equation as y with respect to x^2. If we let $u = x^2$, then $y = a + bu$ and y and u are linearly related. We see that in making a transformation, $u = x^2$, we have changed two variables that are non-linearly related to two variables that are linearly related.

The table below illustrates some suitable examples of transformation on the variables to obtain a linear relationship.

Equation	Independent Variable	Dependent Variable
$y = a + \frac{b}{x}$	$\frac{1}{x}$	y
$y = a + bx^2$	x^2	y
$y^2 = a + bx$	x	y^2
$y = a + b\ln x$	$\ln x$	y
$\ln y = a + bx$	x	$\ln y$
$y = ae^{bx}$ $\Rightarrow \ln y = \ln a + bx$	x	$\ln y$

Note:

1. The above list is not exhaustive. The transformation will usually be given in the question.
2. You may be given a scatter diagram and asked to compare two or more proposed models and determine which model is a better fit. Simply state which equation fits the shape of the scatter plot. If there is more than one possibility, compute the product moment correlation coefficient for each model and **break the tie by choosing the model with $|r|$ closest to 1.**

TRANSFORMATION:

By the means of transformation, we are able to achieve linearity, not with the variable directly but with the transformed variable. This allows us to study the statistical relationship between the variables.

Example 10

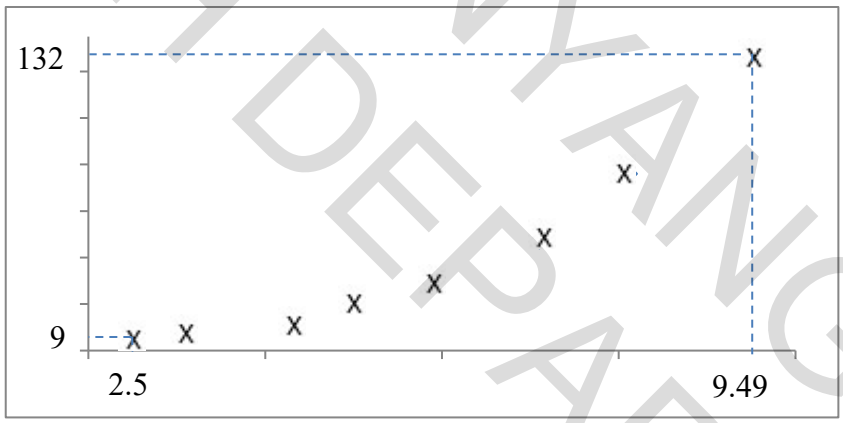
A research worker gave each of eight children a list of words of varying difficulty and asked them to define the meaning of each word. The following table shows the age in years and the number of correctly defined words for each child.

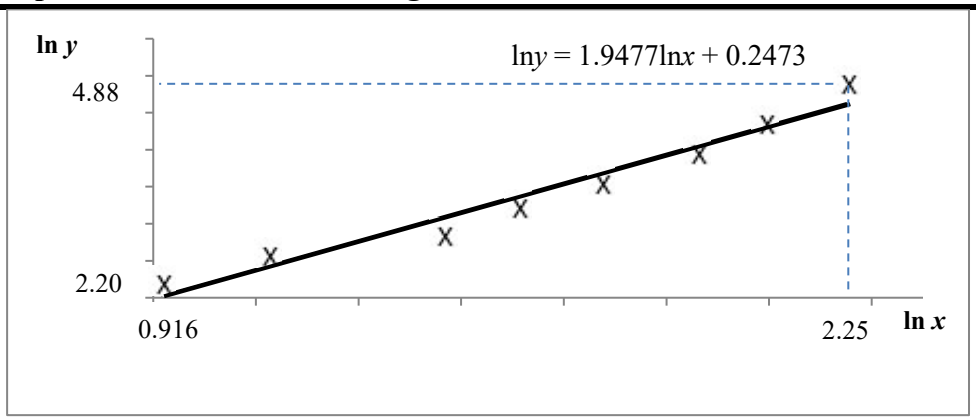
Child	1	2	3	4	5	6	7	8
Age, x (years)	2.5	3.1	4.3	5.0	5.9	7.1	8.1	9.49
Number of correct words, y	9	13	18	25	35	53	81	132

(a) Using a scatter diagram, comment on the suitability of finding the linear regression line of y on x.

(b) Find the product moment correlation coefficient between $\ln y$ and $\ln x$ and sketch the scatter diagram of $\ln y$ against $\ln x$. Comment on the suitability of the model $\ln y = a + b \ln x$.

(c) Find the least-squares regression line of $\ln y$ on $\ln x$ and sketch this line on the scatter diagram in **(b)**.

SOLUTION:	**THINKZONE**
(a)  The scatter diagram reveals that the relationship between x and y is non-linear: as x increases, y increases at an *increasing rate*. Thus it is not suitable to find the linear regression line for y on x.	Be extra alert when keying in values into L_1. The first row in the given table {1, 2, 3, …, 8} is not the variable. The correct list should be {2.5, 3.1, … , 9.49}.
(b) $\ln y = a + b \ln x \Rightarrow y = e^{a+b\ln x} \Rightarrow y = e^a x^b$ The scatter diagram shows that as x increases, y increases at an *increasing rate.* This agrees with the model $\ln y = a + b \ln x$. Also, $r = 0.982$ is very close to 1, which shows a strong positive linear relationship between the variables $\ln x$ and $\ln y$. Hence the model $\ln y = a + b \ln x$ is suitable.	
(c) By GC, the regression line of $\ln y$ on $\ln x$ is $\ln y = 1.948 \ln x + 0.247$.	See **ANNEX** for the GC keystrokes



Example 11

The amount of a chemical, x (in grams), varies with time, t (in minutes) for which a particular chemical reaction has taken place and is given by the formula $x = x_0 e^{-kt}$, where x_0 and k are constants. The values of t may be considered to be exact, while the values of x are subject to experimental error.

Time, t	0.2	0.4	0.6	0.8	1.0
Amount, x	3.22	1.63	0.89	0.41	0.36

The variable y is defined by $y = \ln x$.

(a) Using a least-squares method, calculate the equation of the appropriate regression line and hence give estimates of x_0 and k.

(b) From the equation, estimate the decrease in y when t increases by 1.

(c) Use the regression line obtained in **(a)** to give the best estimate of t when $x = 1.5$. Comment briefly on the reliability of the estimate.

SOLUTION:	**KEYSTROKES**
(a) $x = x_0 e^{-kt}$ $\ln x = \ln\left(x_0 e^{-kt}\right)$ $= \ln x_0 + \ln e^{-kt}$ $= \ln x_0 - kt$ By GC, $\ln x = -2.8811t + 1.6543$ Hence $-k = -2.8811 \Rightarrow k = 2.88$ (3 s.f.) and $\ln x_0 = 1.6543 \Rightarrow x_0 = e^{1.6543} = 5.23$ (3 s.f.) **(b)** When t increases by 1, y is expected to decrease by approximately 2.88. **(c)** When $x = 1.5$, $y = \ln(1.5)$. Using the equation of the regression line of y on t, the estimated value of t is 0.433. Since $r = -0.984$, which is very close to -1 and $x = 1.5$ is within the data range, we can use the regression line of y on t to obtain an estimate of t. Thus, the estimate is likely to be reliable.	See **ANNEX** for the GC keystrokes

Example 12

A student wishes to determine the relationship between the length of a pendulum, l, and the corresponding period, T. After conducting the experiment, he obtained the following set of data:

l (cm)	150	135	120	105	90	75	60	45	30	15
T (s)	2.45	2.31	2.22	2.07	1.91	1.74	1.56	1.35	1.10	0.779

(a) Obtain a scatter diagram of this set of data.

(b) The student proposes the following two models:

$$A:\ T = a + b\ln(l)$$

$$B:\ T^2 = a + bl.$$

(i) Calculate the product moment correlation coefficient for both models, giving your answers to 4 decimal places.

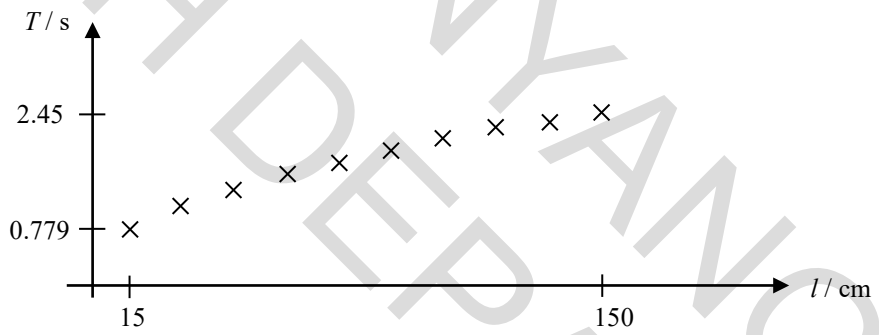
(ii) Determine which model is appropriate for this set of data.

(c) Using the model determined in **(b)** part **(ii)**, estimate the value of l when $T = 3$ to 1 decimal place. Comment on the suitability of this method.

(d) Find a value of l and its corresponding value of T such that the equation of the regression line for the chosen model will remain the same after the addition of this pair of values.

SOLUTION:

(a)



(b)(i) Using GC, for model A, $r = 0.9871$

for model B, $r = 0.9996$.

(ii) Since $0.9996 > 0.9871$, model B indicates a stronger positive linear correlation between l and T^2 than the linear correlation between $\ln(l)$ and T of model A. Hence, model B is an appropriate model.

Note that both equations can apply to this scatter diagram. Thus we compare the value of r as instructed in the question to determine the better model.

(c) Using model B, the regression line of T^2 on l is $T^2 = 0.025136 + 0.040060l$.

When $T = 3$, $3^2 = 0.025136 + 0.040060l$.

Hence $l = 224.0$.

Using the regression line may not be appropriate as $T = 3$ is not within the data range even though r is close 1.

Are we justified in using the regression line of T^2 on l to estimate l given a value of T?

(d) Note that $\left(\bar{l}, \overline{T^2}\right)$ is the *only* point that we can safely add to the set of data points and not change the regression line.

By GC, we find that $\bar{l} = 82.5$ and $\overline{T^2} = 3.3300541$

Thus, $l = 82.5$ and $T = \sqrt{3.3300541} \approx 1.82$.

Important: Note that $\overline{T^2} \neq \overline{T}^2$.

See **ANNEX** for the GC keystrokes to find the values of $\bar{l}$ and $\overline{T^2}$.

Note: For any set of data, adding the point $(\bar{x}, \bar{y})$ will not alter the equation of the regression line of y on x and the equation of the regression line of x on y.

Summary:

a.	We first determine using a scatter diagram that a **linear** relationship between the variables x and y exist.
b.	For a regression line y on x, ***x*** is taken as the **independent** variable and ***y*** as the **dependent** variable. For a regression line x on y, ***y*** is taken as the **independent** variable and ***x*** as the **dependent** variable.
c.	When it is not obvious which is the independent or dependent variable, then, • If you are given x and asked to *estimate* y, use the regression line *y on x*. • If you are given y and asked to *estimate* x, use the regression line *x on y*
d.	If it is clear from the question that x is the independent variable (e.g. controlled variable), while y is the dependent variable, then **there is only one regression line *y* on *x***. (Note: you should not find regression line x on y in this case)
e.	Uses of the regression line of y on x: • Should only be used to estimate values of y within the range of values of x used in calculating the regression line (Interpolation). • Extrapolation **should not** be used as we have no way of knowing the relationship between x and y that are not in the data range. Commenting on reliability of estimates: • Estimates obtained from the least squares regression lines are reliable if the given values are within the data range and r is close to 1 or -1. • Estimates obtained from extrapolation are not reliable.
f.	Both regression lines of y on x and x on y will pass through the point $(\bar{x}, \bar{y})$.
g.	If the scatter diagram reveals a non-linear relationship between two variables x and y, say of the form, $y = a + b\,\mathrm{f}(x)$, then it is possible to introduce a new variable $u = \mathrm{f}(x)$ so that the equation becomes $y = a + bu$ which is linear in u and y.
h.	Given a few non-linear models, we first use the scatter diagram to determine which is the appropriate model. If two or more models are applicable, then the model such that $\lvert r \rvert$ is the largest after transformation to a linear equation is in general the best model.

ANNEX

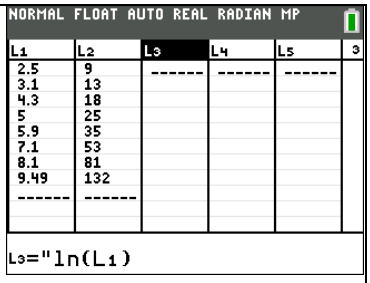
GC Keystrokes for Example 10

To use the GC to transform the data and obtain the regression line of ln$y$ on ln$x$.

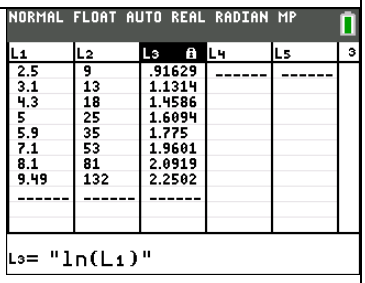
Press stat and highlight '**1:Edit…**' and press enter. Key in the values of x and y into L_1 and L_2 respectively,

In cell L_3, key in "**ln**(L_1) and press enter .

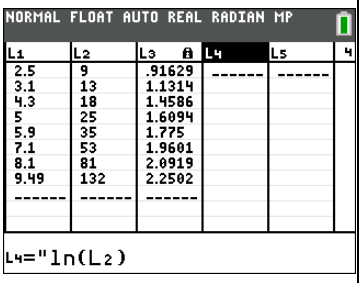
The command **"** is to ensure that the GC stores the formula. **"** can be found on the '+' button.



Notice that there is a black icon in the header of L_3 to indicate that a formula is stored. When the cursor is moved to L_3, the formula "$\ln(L_1)$" will appear below the screen to indicate that $L_3 = \ln(L_1)$. If **"** is omitted when keying in "$\ln(L_1)$" in L_3, the formula will not stored.

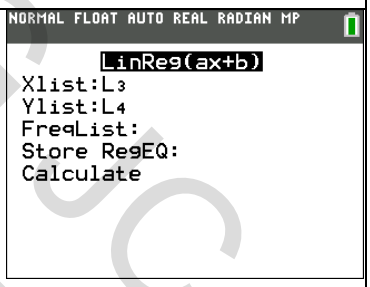


In cell L_4, key in "$\ln(L_2)$ and press enter .



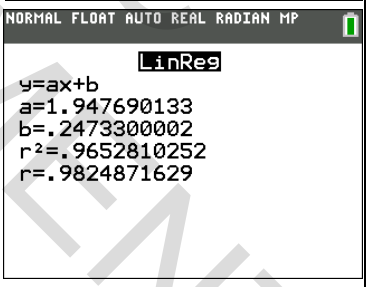
Press stat, highlight '**CALC**' and select '**4:LinReg(ax+b)**' and press enter

To obtain the product moment correlation coefficient between $\ln y$ and $\ln x$, we key in 'L_3' in '**Xlist:**' and 'L_4' in '**Ylist**'. We then scroll down to '**Calculate**' and press enter .

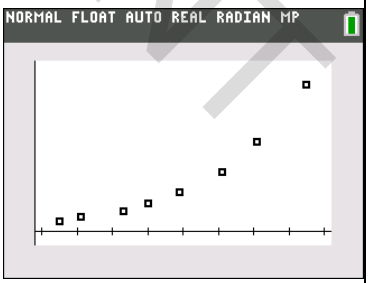


By GC, $r = 0.982$.

$\ln y = 1.948 \ln x + 0.247$



Since $r = 0.982$ is very close to 1 and also based on the scatter diagram, we can say that the model $\ln y = a + b \ln x$ is suitable.



Note:

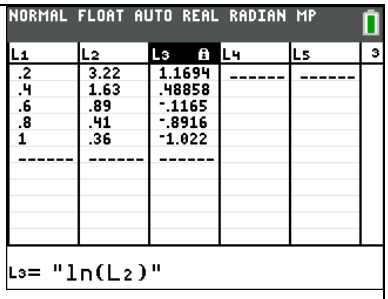
You will need to plot L_4 against L_3 in your GC to see that ln$x$ and ln$y$ are linearly related.

GC Keystrokes for Example 11

To use the GC to transform the data and obtain the regression line of ln$x$ on t.

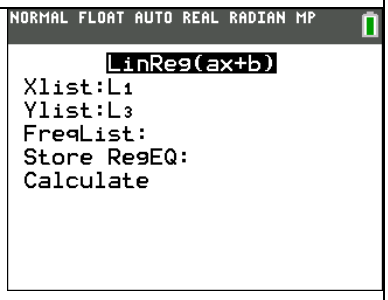
Press [stat] and highlight '**1:Edit…**' and press [enter]. Key in the values of t and x into L_1 and L_2 respectively,

In cell L_3, key in '$\ln(L_2)$' and press [enter].

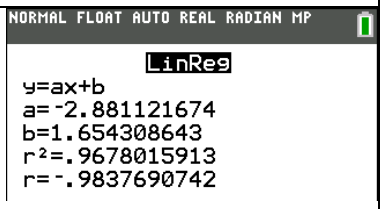


Press [stat], highlight '**CALC**' and select '**4:LinReg(ax+b)**' and press [enter].

To obtain the values of x_0 and $-k$, we key in 'L_1' in '**Xlist:**' and 'L_3' in '**Ylist**'. We then scroll down to '**Calculate**' and press [enter].



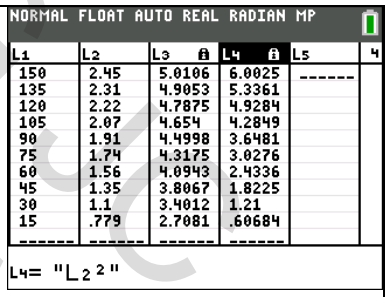
Subsequently, we get the results:
From the GC, the regression line of y on t is
$y = -2.8811t + 1.6543$ or
$\ln x = -2.8811t + 1.6543$ since $y = \ln x$.
Thus, $k \approx 2.88$ and $x_0 = e^{1.6543} \approx 5.23$.



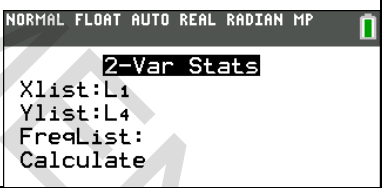
GC Keystrokes for Example 12

To use the GC to calculate $\bar{l}$ and $\overline{T^2}$

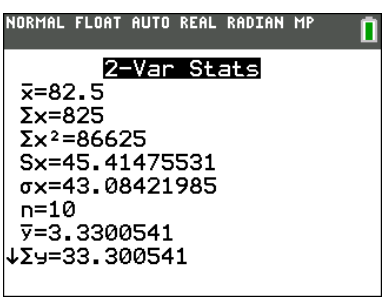
To obtain the values of $\bar{l}$ and $\overline{T^2}$, we key in 'L_1' in '**Xlist:**' and 'L_4' in '**Ylist**' respectively.

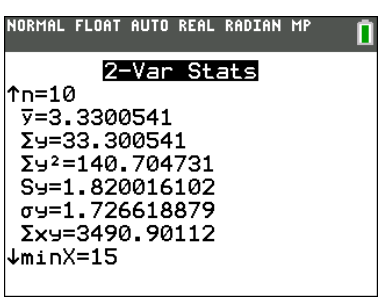


Press [stat], highlight '**CALC**' and select '**2:2-Var Stats**' and press [enter].



Scrolling down with the navigating buttons:





Appendix

We shall prove that the scaling and translation of bivariate data will not change the product moment of correlation.

It is given that $r=\dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\left(\sum(x-\bar{x})^2\right)\left(\sum(y-\bar{y})^2\right)}}=\dfrac{\sum xy-\dfrac{\sum x\sum y}{n}}{\sqrt{\left\{\left(\sum x^2-\dfrac{\left(\sum x\right)^2}{n}\right)\left(\sum y^2-\dfrac{\left(\sum y\right)^2}{n}\right)\right\}}}$.

Suppose x has been replaced by $ax+b$, and y has been replaced by $cy+d$, where $a,c>0$.

Note that $\sum(ax+b)=\sum ax+\sum b$

$=a\sum x+nb$

$$\frac{\sum(ax+b)(cy+d)-\dfrac{\sum(ax+b)\sum(cy+d)}{n}}{\sqrt{\left(\sum(ax+b)^2-\dfrac{\left(\sum(ax+b)\right)^2}{n}\right)\left(\sum(cy+d)^2-\dfrac{\left(\sum(cy+d)\right)^2}{n}\right)}}$$

$$=\frac{\sum(acxy+adx+bcy+bd)-\dfrac{\left(a\sum x+nb\right)\left(c\sum y+nd\right)}{n}}{\sqrt{\left(\sum\left(a^2x^2+2abx+b^2\right)-\dfrac{\left(a\sum x+nb\right)^2}{n}\right)\left(\sum\left(c^2y^2+2cdy+d^2\right)-\dfrac{\left(c\sum y+nd\right)^2}{n}\right)}}$$

$$=\frac{ac\sum xy+ad\sum x+bc\sum y+nbd-\dfrac{ac\sum x\sum y+nad\sum x+nbc\sum y+n^2bd}{n}}{\sqrt{\left(a^2\sum x^2+2ab\sum x+nb^2-\dfrac{a^2\left(\sum x\right)^2+2nab\sum x+n^2b^2}{n}\right)\left(c^2\sum y^2+2cd\sum y+nd^2-\dfrac{c^2\left(\sum y\right)^2+2ncd\sum y+n^2d^2}{n}\right)}}$$

$$=\frac{ac\sum xy-\dfrac{ac\sum x\sum y}{n}}{\sqrt{\left(a^2\sum x^2-\dfrac{a^2\left(\sum x\right)^2}{n}\right)\left(c^2\sum y^2-\dfrac{c^2\left(\sum y\right)^2}{n}\right)}}$$

$$=\frac{ac\left(\sum xy-\dfrac{\sum x\sum y}{n}\right)}{\sqrt{a^2c^2\left(\sum x^2-\dfrac{a^2\left(\sum x\right)^2}{n}\right)\left(\sum y^2-\dfrac{\left(\sum y\right)^2}{n}\right)}}$$

$$=\frac{\sum xy-\dfrac{\sum x\sum y}{n}}{\sqrt{\left(\sum x^2-\dfrac{a^2\left(\sum x\right)^2}{n}\right)\left(\sum y^2-\dfrac{\left(\sum y\right)^2}{n}\right)}}\because ac>0$$

$=r$