CORRELATION AND REGRESSION

CORRELATION

Introduction

CORRELATION problems which involve measuring the *strength* of a relationship.

Correlation Analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

"Two variables are said to be in correlation if the change in one of the variables results in a change in the other variable". also,

Types of Correlation

(1) Positive and Negative correlation .

(2) Linear and Non – Linear correlation.

1: Positive and Negative Correlation

- A positive relationship exists when both variables increase or decrease at the same time. (Weight and height).

A negative relationship exist when one variable increases and the other variable decreases or vice versa. (Strength and age).

***Example*:1**: The following are the heights and weights of 15 students of a class. Draw a graph to indicate whether the correlation is negative or positive.

Heights (cms)	Weights (kgs)
170	65
172	66
181	69
157	55
150	51
168	63
166	61
175	75
177	72
165	64
163	61
152	52
161	60
173	70
175	72

2.Linear and Non – Linear Correlation

- The correlation between two variables is said to be **linear** if the change of one unit in one variable result in the corresponding change in the other variable over the entire range of values.
- The relationship between two variables is said to be **non – linear** if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate.

- Since the points are dense (close to each other) we can expect a high degree of correlation between the series of heights and weights. Further, since the points reveal an upward trend, the correlation is positive. Arrange the data in increasing order of height and check that , as height increases, the weight also increases.

Correlation Coefficient (r)

- The correlation coefficient computed from the sample data measures the strength and direction of a relationship between two variables.
- The range of the correlation coefficient is.

  - 1 to + 1 and is identified by *r.*

The formula is:

to compute a correlation coefficient

$$r = [n(\sum xy) - (\sum x)(\sum y)] / \{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]\}^{0.5}$$

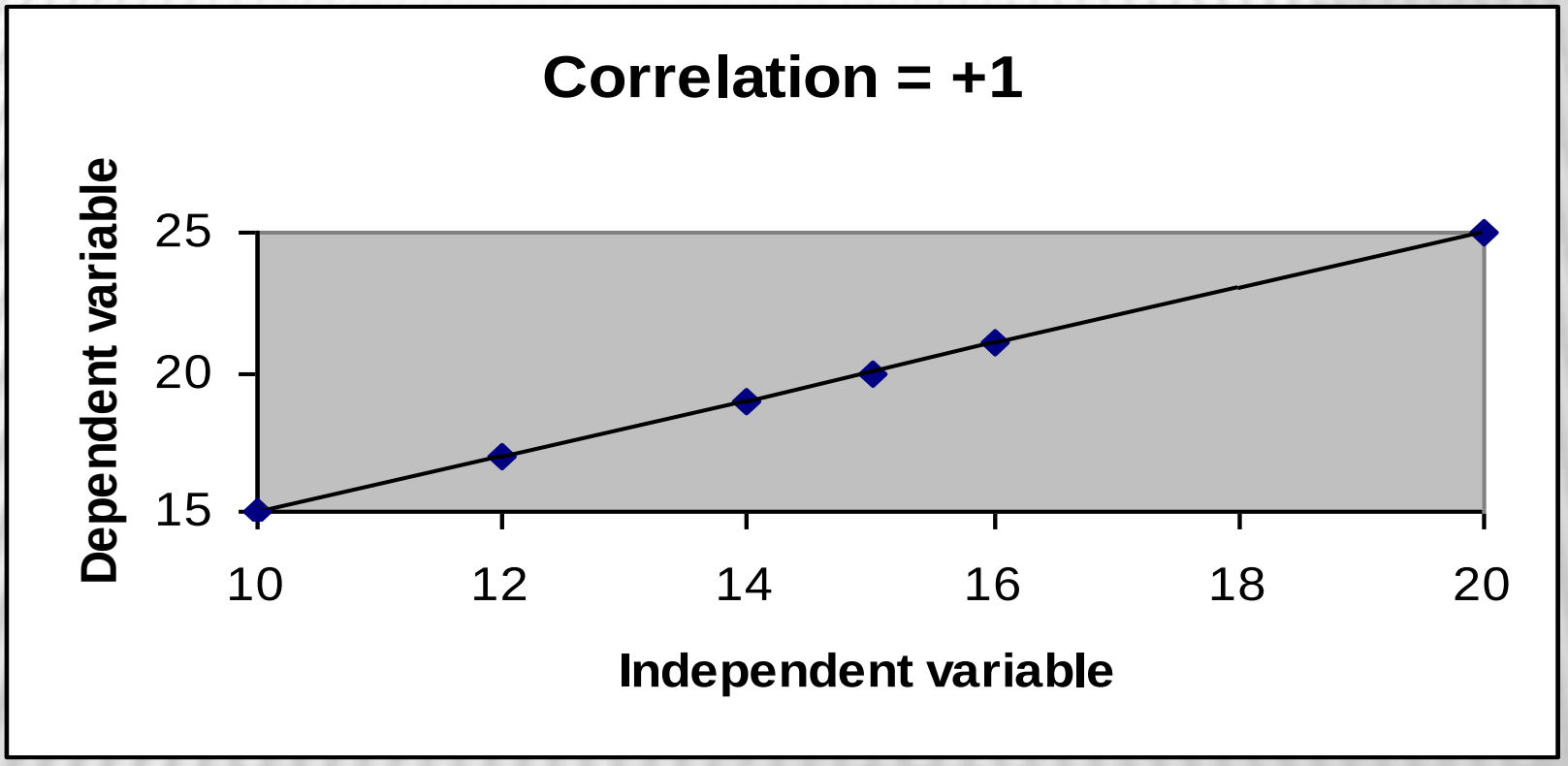
Where *n* is the number of data pairs, x is the independent variable and y the dependent variable.

The value of r can range between -1 and + 1.

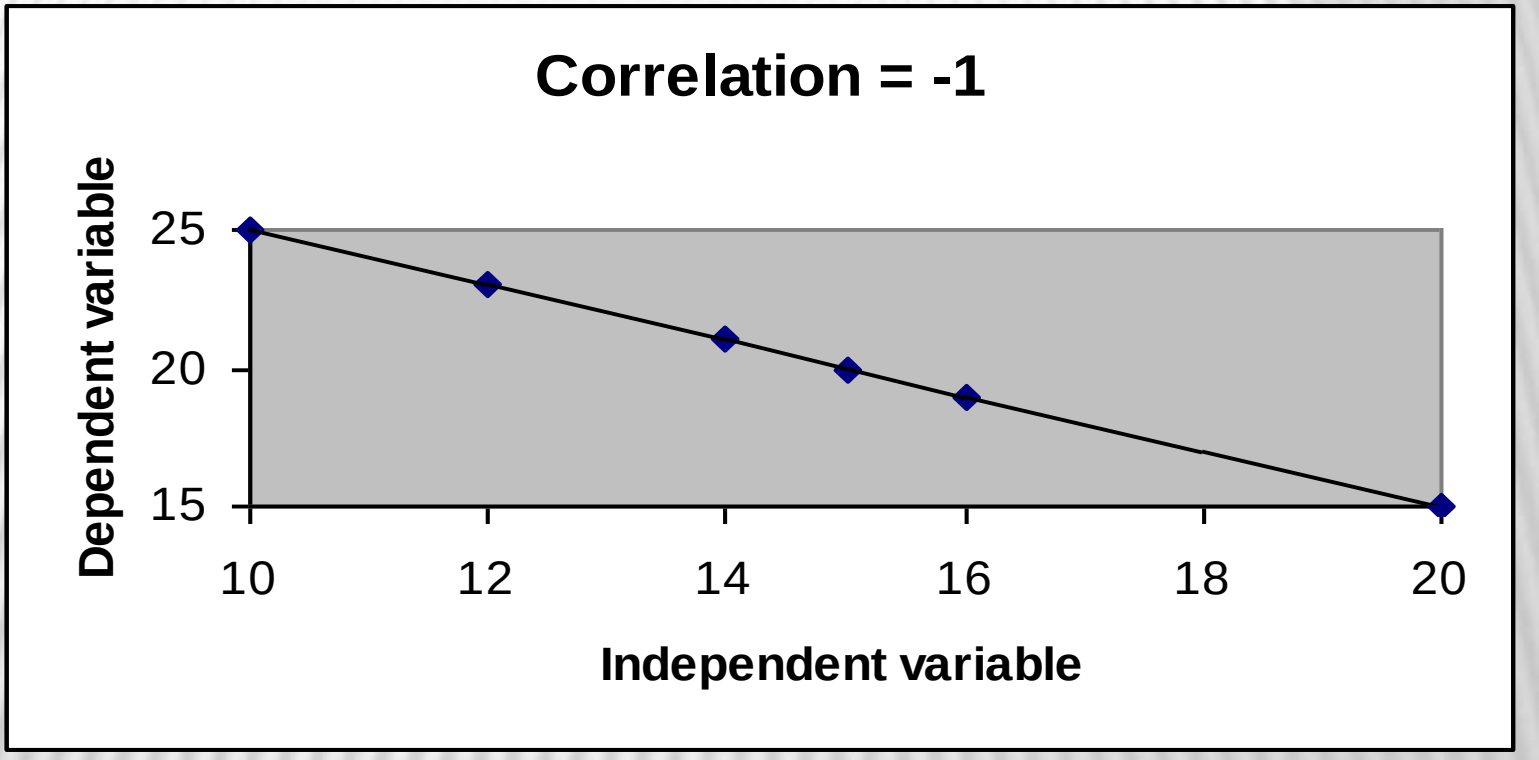
- r = 0 is no linear correlation
- r = 1 is perfect positive (slope up from bottom left to top right) linear correlation
- r = -1 is perfect negative (slope down from top left to bottom right) linear correlation

Range of correlation coefficient

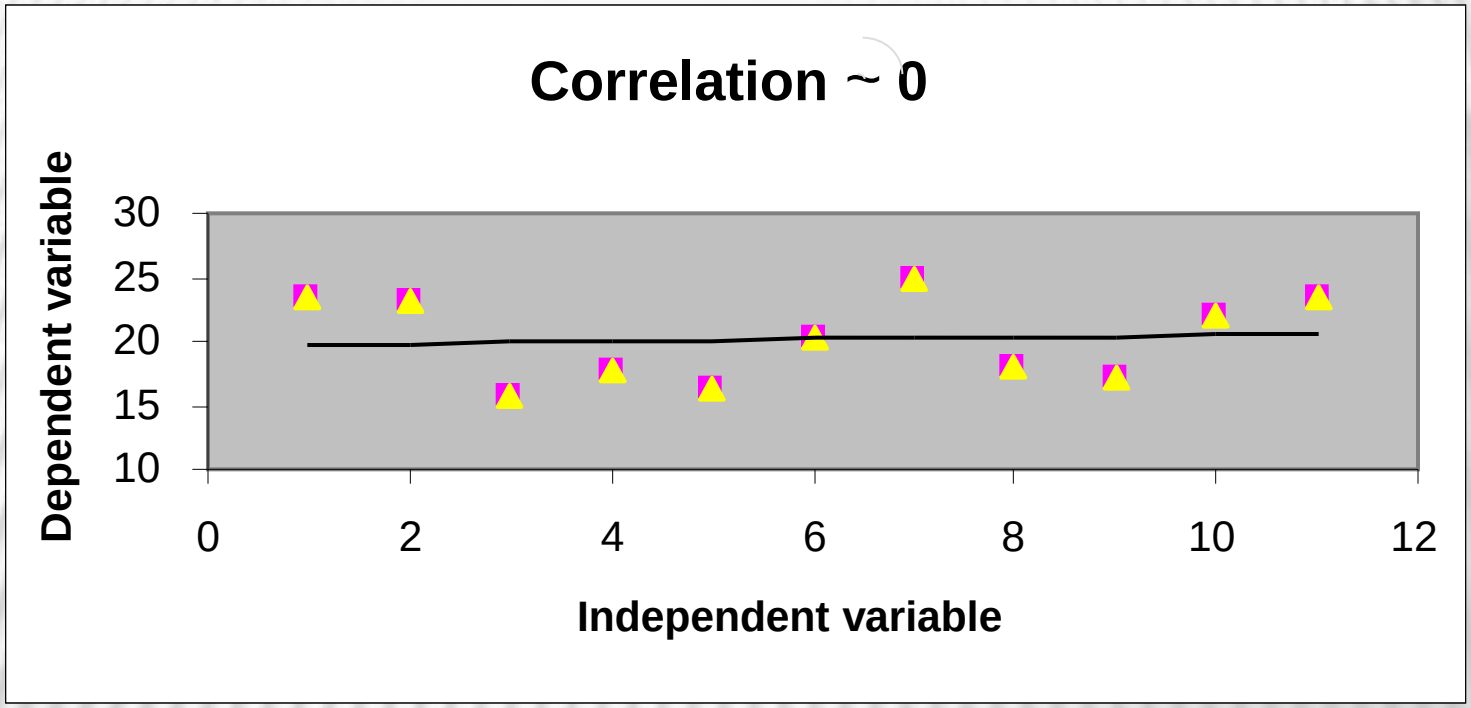
- In case of exact positive linear relationship the value of r is +1.
- In case of a strong positive linear relationship, the value of *r* will be close to + 1.



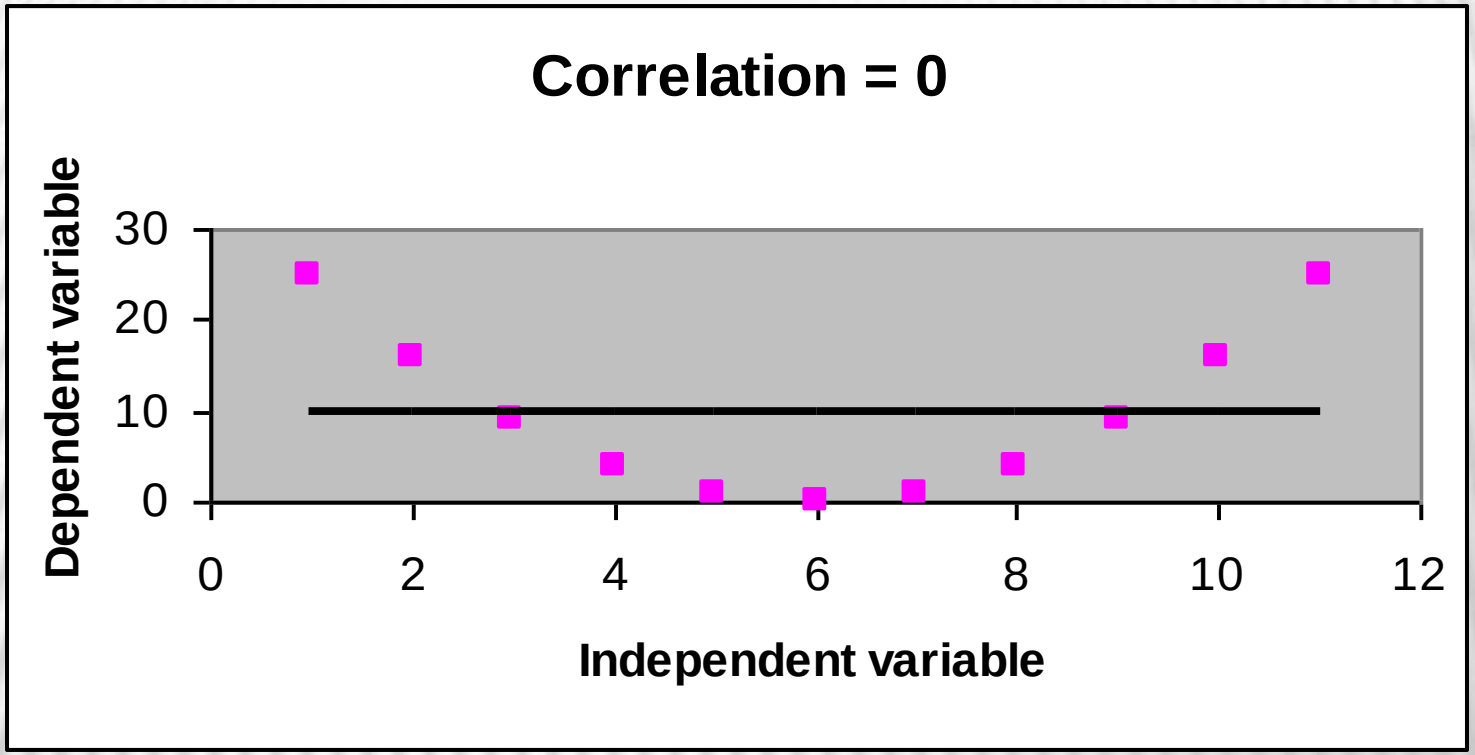
- In case of exact negative linear relationship the value of *r* is –1.
- In case of a strong negative linear relationship, the value of *r* will be close to – 1.



- In case of a weak relationship the value of *r* will be close to 0.



- In case of nonlinear relationship the value of *r* will be close to 0.



- ***Example for correlation coefficient***

Using the data on age and blood pressure, calculate the $\sum x$, $\sum y$, $\sum xy$, $\sum x^2$ and $\sum y^2$.

Student	**Age**	**Blood Pressure**	**Age*BP**	**age^2**	**BP^2**
A	43	128	5504	1849	16384
B	48	120	5760	2304	14400
C	56	135	7560	3136	18225
D	61	143	8723	3721	20449
E	67	141	9447	4489	19881
F	70	152	10640	4900	23104
Sum	**345**	**819**	**47634**	**20399**	**112443**

Substitute in the formula and solve for r:

$r= \{(6*47634)-(345*819)\}/\{[(6*20399)-345^2][(6*112443)-819^2]\}^{0.5}$.

r= 0.897.

The correlation coefficient ***r*** suggests a ***strong positive*** relationship between age and blood pressure.

Note:

Correlation measures association and not causation.

Correlation assumes linear relationship.

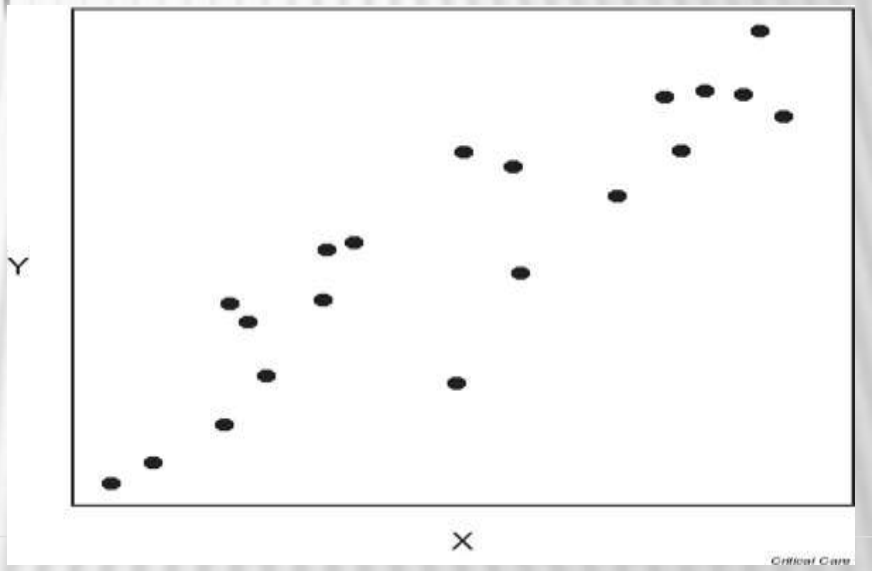
- Values between –1 and +1 and measure the strength and direction of the relationship.

INTERPRETATION

THE CORRELATION IS 0.9

THERE IS A STRONG POSITIVE RELATIONSHIP BETWEEN AGE AND BLOOD PRESSURE

	Age
Blood Pressure	0.90



Test of Correlation

Null hypothesis: correlation is zero

Test statistic is

$$t = r\left[(n-2)/(1-r^2)\right]^{0.5}$$

The statistic is distributed as Student t distribution with n-2 degrees of freedom

Excel does not calculate this statistic and you can manually calculate it

REGRESSION

***Definition*:**

REGRESSION **problems which are concerned with the *form or nature* of a relationship**

A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.

the regression statistics can be used to predict the dependent variable when the independent variable is known. Regression goes beyond correlation by adding prediction capabilities.

The regression line is the line that best fits the data:

- **The correlation tells us how well the regression line fits the data, r.**
- **The relationship between the correlation and the slope of the regression line is given by r=b*(Sx/Sy)**

X-axis	Y-axis
independent	dependent
predictor	predicted
carrier	response
input	output

REGRESSION MODELS

A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation.

Various tests are then employed to determine if the model is satisfactory.

If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables

In simple linear regression model(the first order linear model)

the model used to describe the relationship between a single dependent variable **y** and a single independent variable **x** is

y = a + bx + k **a** =y-intercept , **b** =slope of the line , **k**= probabilistic error

(**b** defined as the ratio **rise/run** or **change in y/change in x**,

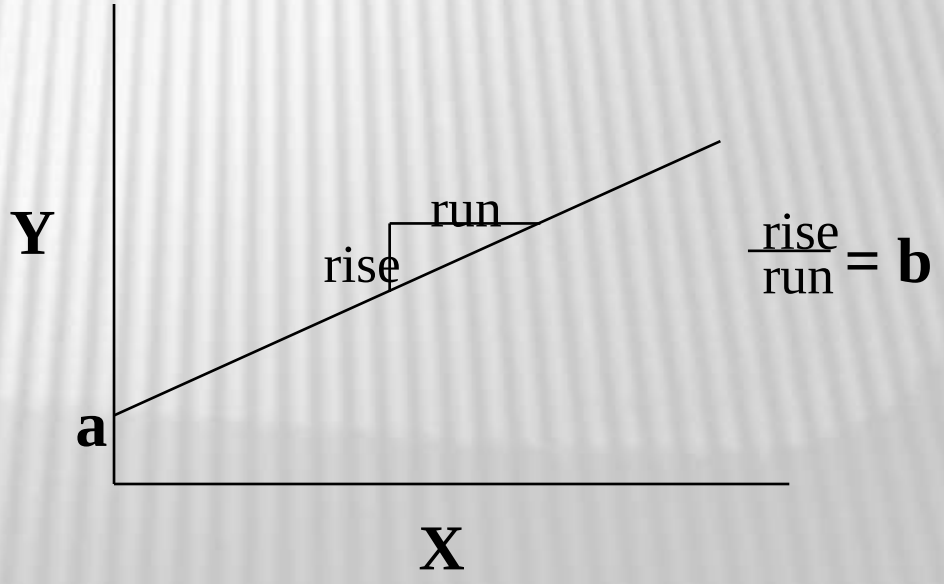
both **a** and **b** are referred to as the model parameters),

and **k** is a probabilistic **error** term that accounts for the variability in **y** that cannot be explained by the linear relationship with **x**.

If the error term were not present, the model would be deterministic; in that case, knowledge of the value of x would be sufficient to determine the value of y.

In a simple regression analysis, one **dependent variable** is examined in relation to only one **independent variable**. The analysis is designed to derive an equation for the line that best models the relationship between the dependent and independent variables. This equation has the mathematical form: **y = a + bx**

where, **y** is the value of the dependent variable, **x** is the value of the independent variable, **a** is the intercept of the regression line on the **y** axis when **x** = 0, and **b** is the slope of the regression line.



Regression Formula:

Regression Equation **$y = a + bx$**

Slope **$(b) = (n\Sigma xy - (\Sigma x)(\Sigma y)) / (n\Sigma x^2 - (\Sigma x)^2)$**

Intercept **(a)** = **$(\Sigma y - b(\Sigma x)) / n$**

where x and y are the variables.

b = is the gradient, slope or regression coefficient

a = is the intercept of the line at Y axis or regression constant

n = number of values or elements

x = First Score

y = Second Score

Σxy = Sum of the product of first and Second Scores

Σx= Sum of First Scores

Σy = Sum of Second Scores

Σx^2 = Sum of square First Scores

- **Regression Example**: To find the Simple/Linear Regression of

x values	y values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

To find regression equation, we will first find slope, intercept and use it to form regression equation..

- Step 1: Count the number of values.

  n = 5

- Step 2: Find xy, x^2 See the below table

x value	y value	x*y	$x^2=x*x$
60	3.1	60 * 3.1 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3: Find Σx, Σy, Σxy, Σx^2.

$\Sigma x = 311$

$\Sigma y = 18.6$

$\Sigma xy = 1159.7$

$\Sigma x^2 = 19359$

Step 4: Substitute in the above slope formula given.

$\text{Slope}(b) = (n\Sigma xy - (\Sigma x)(\Sigma y)) / (n\Sigma x^2 - (\Sigma x)^2)$

$= ((5)*(1159.7)-(311)*(18.6))/((5)*(19359)-(311)^2)$

$= (5798.5 - 5784.6)/(96795 - 96721)$

$= 13.9/74$

$= 0.19$

Step 5: Now, again substitute in the above intercept formula given.

$\text{Intercept}(a) = (\Sigma y - b(\Sigma x)) / n$

$= (18.6 - 0.19(311))/5$

$= (18.6 - 59.09)/5$

$= -40.49/5$

$= -8.098$

Step 6: Then substitute these values in regression equation formula

Regression Equation $(y) = a + bx$

$= -8.098 + 0.19x.$

Suppose if we want to know the approximately value for the variable x = 64. Then we can substitute the value in the above equation.

Regression Equation (y) = a + bx

= -8.098 + 0.19(64).

= -8.098 + 12.16

= 4.06

This example will guide you to find the relationship between two variables by calculating the Regression from the above steps.

NON-LINEAR REGRESSION

So far we have considered only linear relationships between variables. Before calculating the correlation coefficient or a regression line, this must be checked with a scatter plot.

If there is a curved relationship it can often be make linear simply by **TRANSFORMING** one (or both) of the variables.

Some common transformations are:

Use

log(y) and x

y and log(x)

log(y) and log(x)

y and 1/x

y and *(X)0.5*

y and x^2

In practice, try a few transformations and see which has the best linearising effect using scatter diagrams. Check the correlation coefficient as well - the closer it is to + _ 1, the greater is the linear relationship between the variables.

(Actually, it is more usual to check the equivalent $\mathbf{r^2}$ value – the closer it is to 100%, the better.)

Having linearised the data, proceed with the regression analysis as before.

- **Applications of regression analysis in pharmaceutical experimentation are numerous:**
- Used to describe the linear relationship between variables **as in Beer's law plots,** where **optical density is plotted against drug concentration;**
- when the functional form of a response is unknown, but where we wish to represent a trend or rate as characterized by the slope (e.g., as may occur when following **a pharmacological response over time);**
- when we wish to describe a process by a relatively simple equation that will relate the response, *Y*, to a fixed value of *X,* such as in **stability prediction (concentration of drug versus time).**
- Test hypothesis (Cause& effect)relationships . For example.see whether variation in X causes variation in Y(**giving people different amounts of a drug and measuring their blood pressure).**

Correlation

- In a correlation, we look at the relationship between two variables without knowing the direction of causality
- For a correlation you do not need to know anything about the possible relation between the two variables
- Many variables correlate with each other for unknown reasons
- Correlation underlies regression but is descriptive only

Regression

- In a regression, we try to predict the outcome of one variable from one or more predictor variables. Thus, the direction of causality can be established.
- 1 predictor=simple regression
- >1 predictor=multiple regression
- For a regression you do want to find out about those relations between variables, in particular, whether one 'causes' the other.
- Therefore, an unambiguous causal template has to be established between the causer and the causee before the analysis!
- This template is inferential.
- **Regression is THE statistical method underlying ALL inferential statistics (t-test, ANOVA, etc.). All that follows is a variation of regression.**