CORRELATION

Correlation

- **key concepts:**

Types of correlation

Methods of studying correlation

a) Scatter diagram

b) Karl pearson’s coefficient of correlation

c) Spearman’s Rank correlation coefficient

d) Method of least squares

Correlation

- **Correlation**: The degree of relationship between the variables under consideration is measure through the correlation analysis.
- The measure of correlation called the correlation coefficient
- The degree of relationship is expressed by coefficient which range from correlation **(-1 $\leq$ r $\geq$ +1)**
- The direction of change is indicated by a sign.
- The correlation analysis enable us to have an idea about the degree & direction of the relationship between the two variables under study.

Correlation

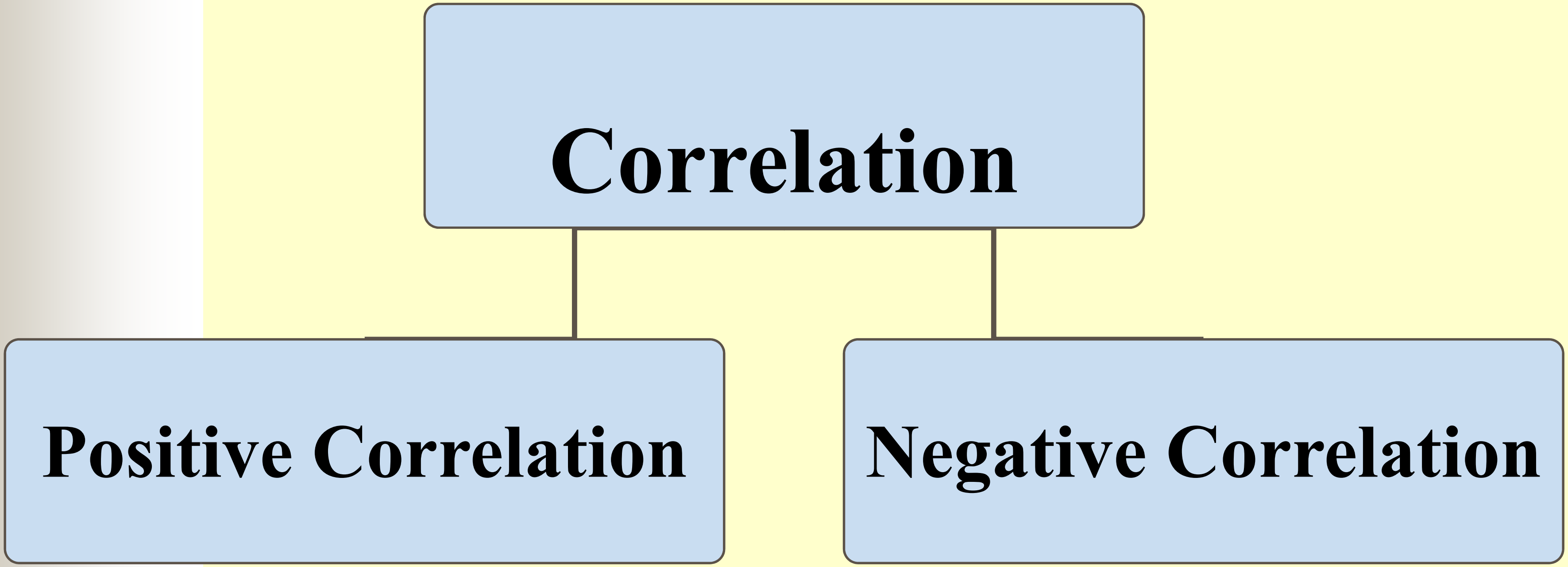
- Correlation is a statistical tool that helps to measure and analyze the degree of relationship between two variables.
- Correlation analysis deals with the association between two or more variables.

Correlation & Causation

- Causation means cause & effect relation.
- Correlation denotes the interdependency among the variables for correlating two phenomenon, it is essential that the two phenomenon should have cause-effect relationship,& if such relationship does not exist then the two phenomenon can not be correlated.
- If two variables vary in such a way that movement in one are accompanied by movement in other, these variables are called cause and effect relationship.
- Causation always implies correlation but correlation does not necessarily implies causation.

Types of Correlation

Type Ⅰ



Correlation

Positive Correlation

Negative Correlation

Types of Correlation Type Ⅰ

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Pub. Exp. & sales, Height & weight.

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Price & qty. demanded.

Direction of the Correlation

- **Positive relationship** – Variables change in the same direction.
    - As X is increasing, Y is increasing
    - As X is decreasing, Y is decreasing
  - E.g., As height increases, so does weight.

Indicated by sign; (+) or (-).

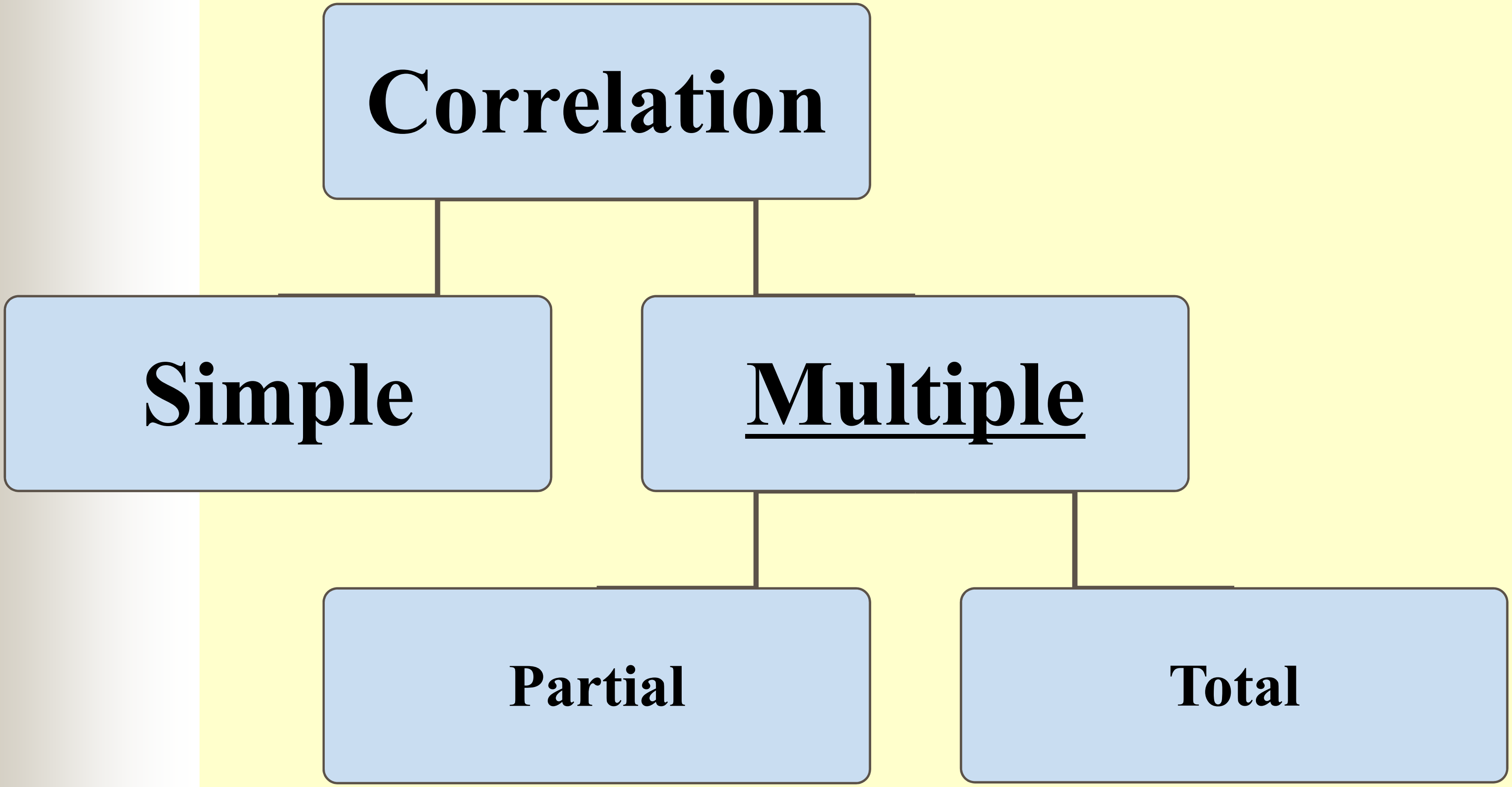
- **Negative relationship** – Variables change in opposite directions.
    - As X is increasing, Y is decreasing
    - As X is decreasing, Y is increasing
  - E.g., As TV time increases, grades decrease

More examples

- **Positive relationships**
- water consumption and temperature.
- study time and grades.

- **Negative relationships**:
- alcohol consumption and driving ability.
- Price & quantity demanded

Types of Correlation
Type II



Correlation

Simple

Multiple

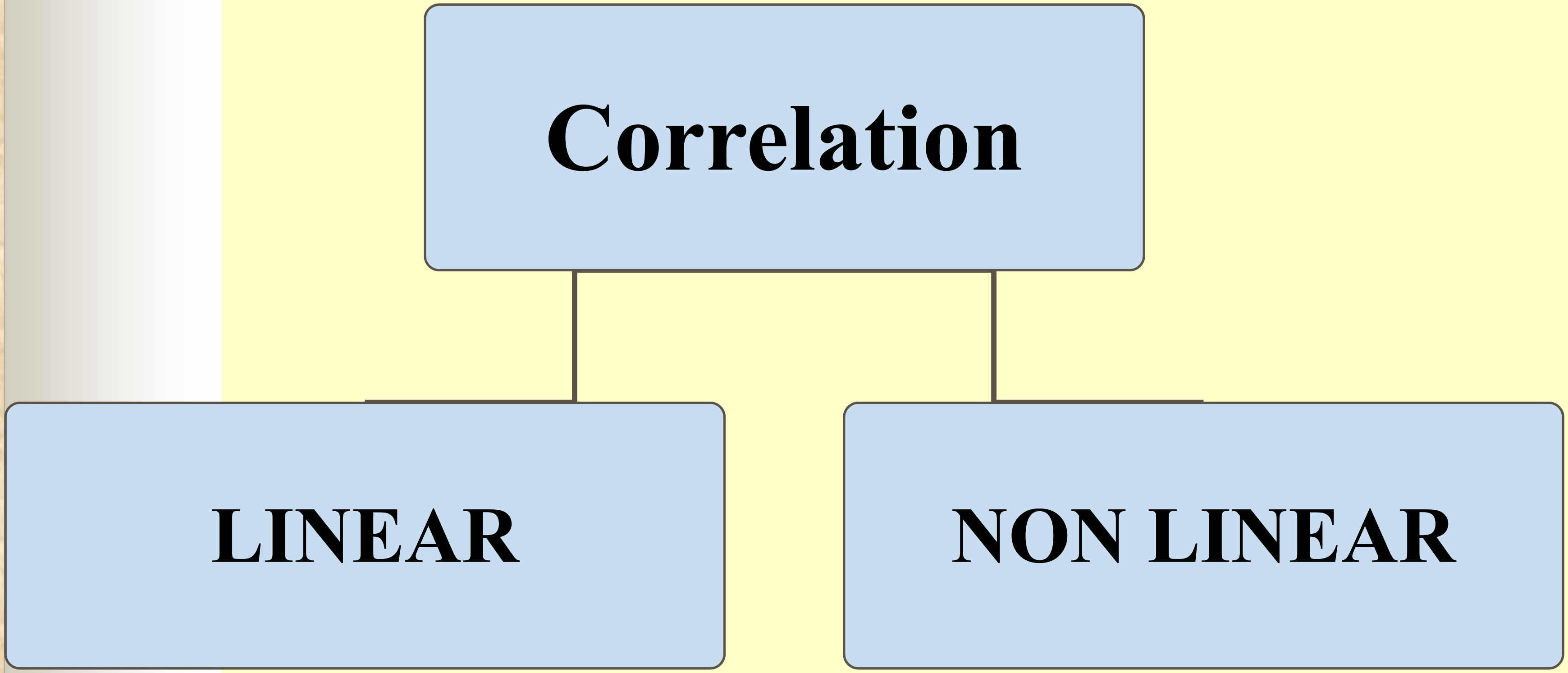
Partial

Total

Types of Correlation Type II

- **Simple correlation:** Under simple correlation problem there are only two variables are studied.
- **Multiple Correlation:** Under Multiple Correlation three or more than three variables are studied. Ex. $Q_d = f\,(P, P_C, P_S, t, y)$
- **Partial correlation:** analysis recognizes more than two variables but considers only two variables keeping the other constant.
- **Total correlation:** is based on all the relevant variables, which is normally not feasible.

Types of Correlation
Type III



Correlation

LINEAR

NON LINEAR

Types of Correlation Type III

- **Linear correlation:** Correlation is said to be linear when the amount of change in one variable tends to bear a constant ratio to the amount of change in the other. The graph of the variables having a linear relationship will form a straight line.

  Ex $X = 1, \quad 2, \quad 3, \quad 4, \quad 5, \quad 6, \quad 7, \quad 8,$

  $Y = 5, \quad 7, \quad 9, \quad 11, 13, 15, 17, 19,$

  $Y = 3 + 2x$

- **Non Linear correlation:** The correlation would be non linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

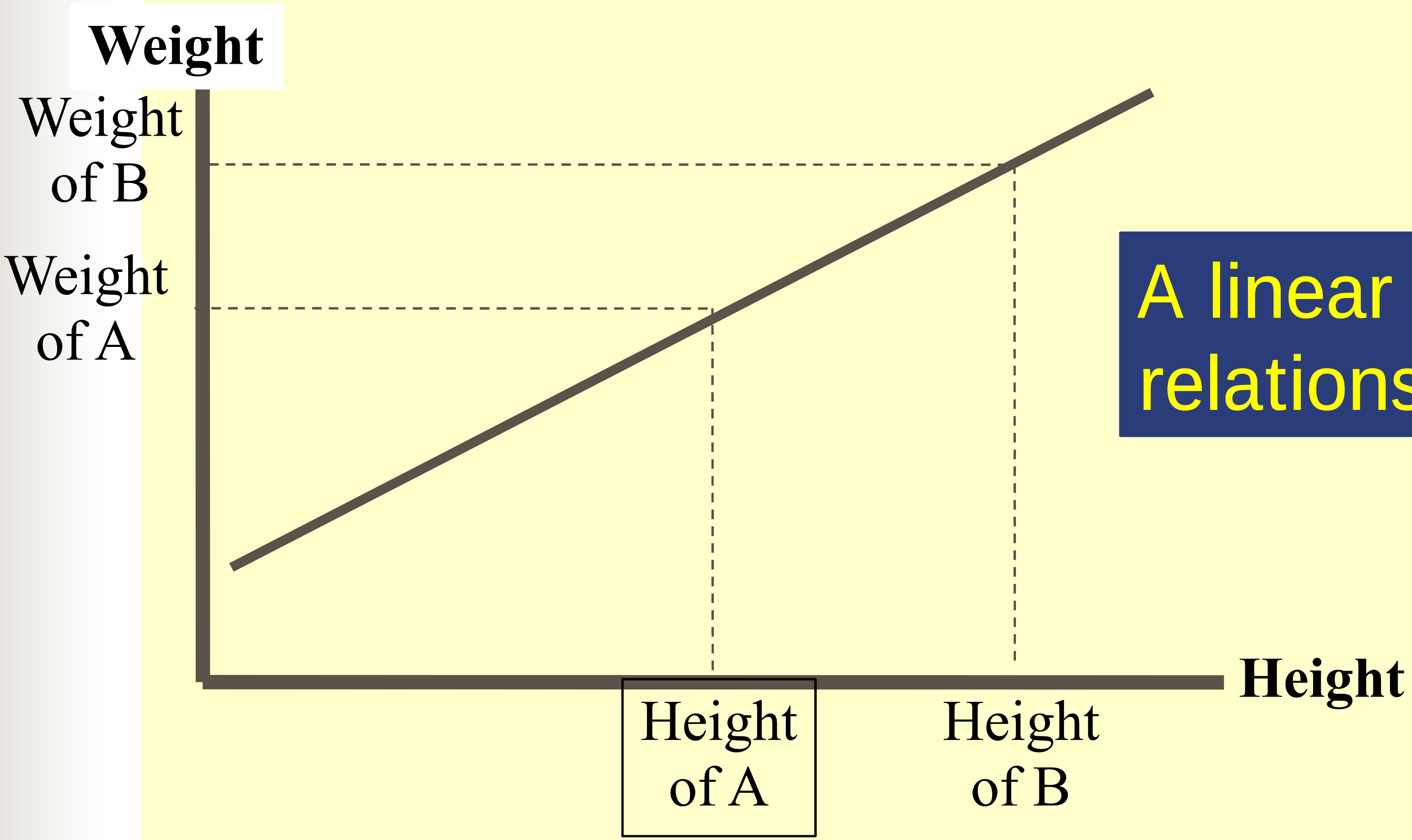
Methods of Studying Correlation

- Scatter Diagram Method
- Graphic Method
- Karl Pearson’s Coefficient of Correlation
- Method of Least Squares

Scatter Diagram Method

- Scatter Diagram is a graph of observed plotted points where each points represents the values of X & Y as a coordinate. It portrays the relationship between these two variables graphically.

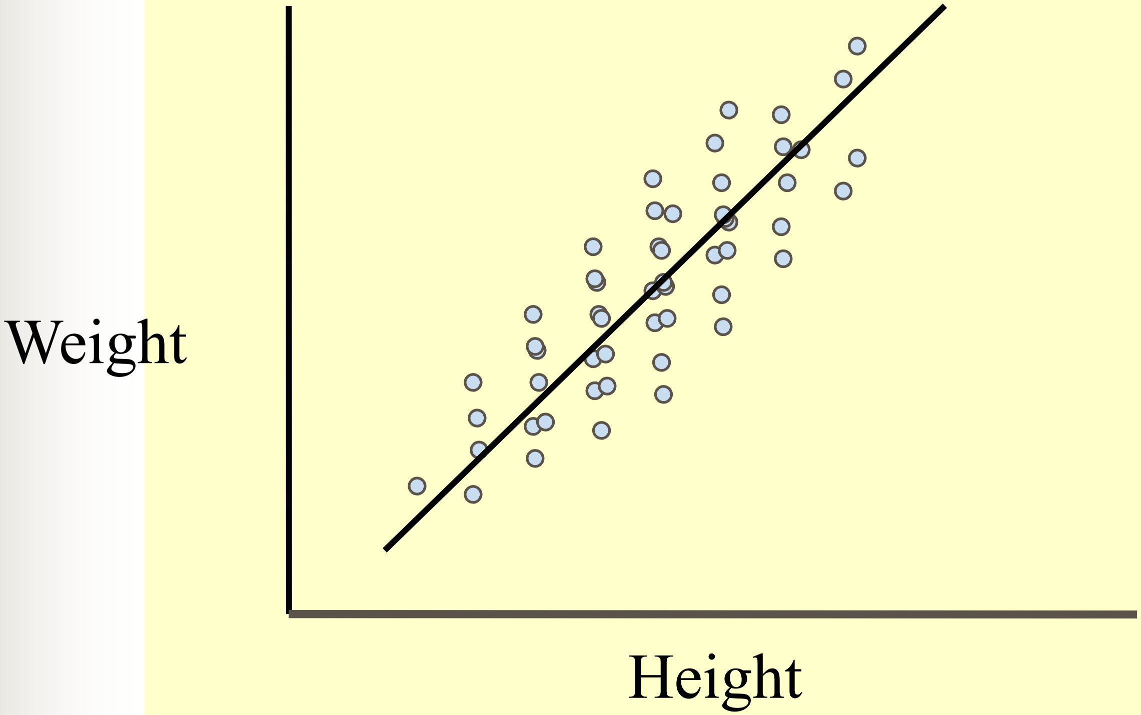
A perfect positive correlation



A linear relationship

High Degree of positive correlation

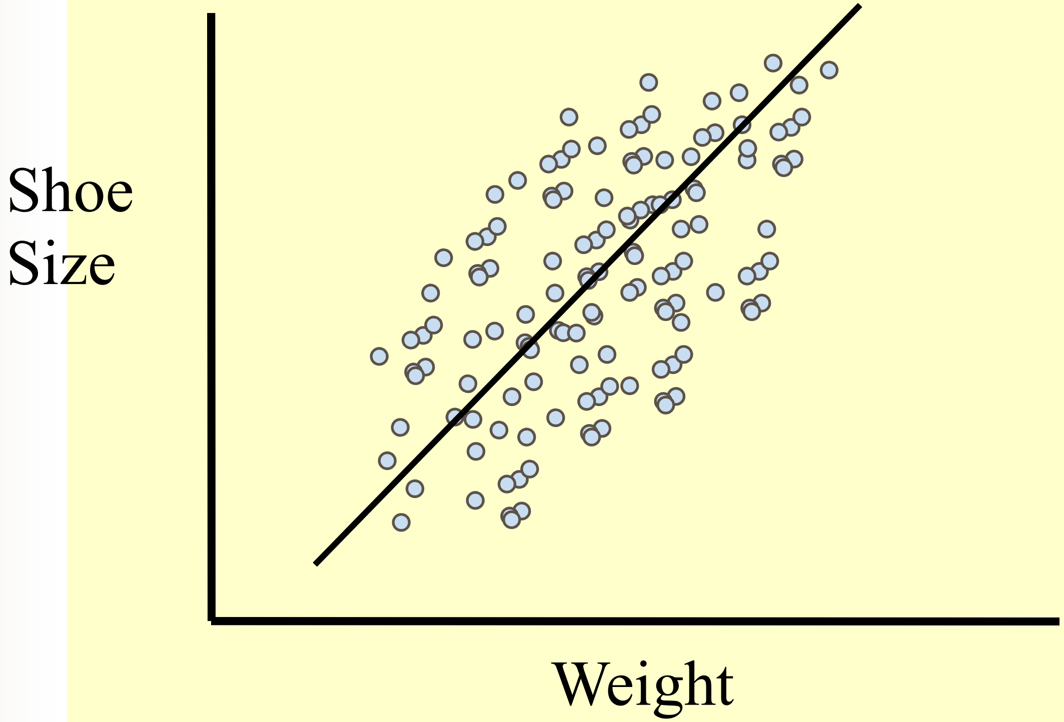
- Positive relationship



r = +.80

Degree of correlation

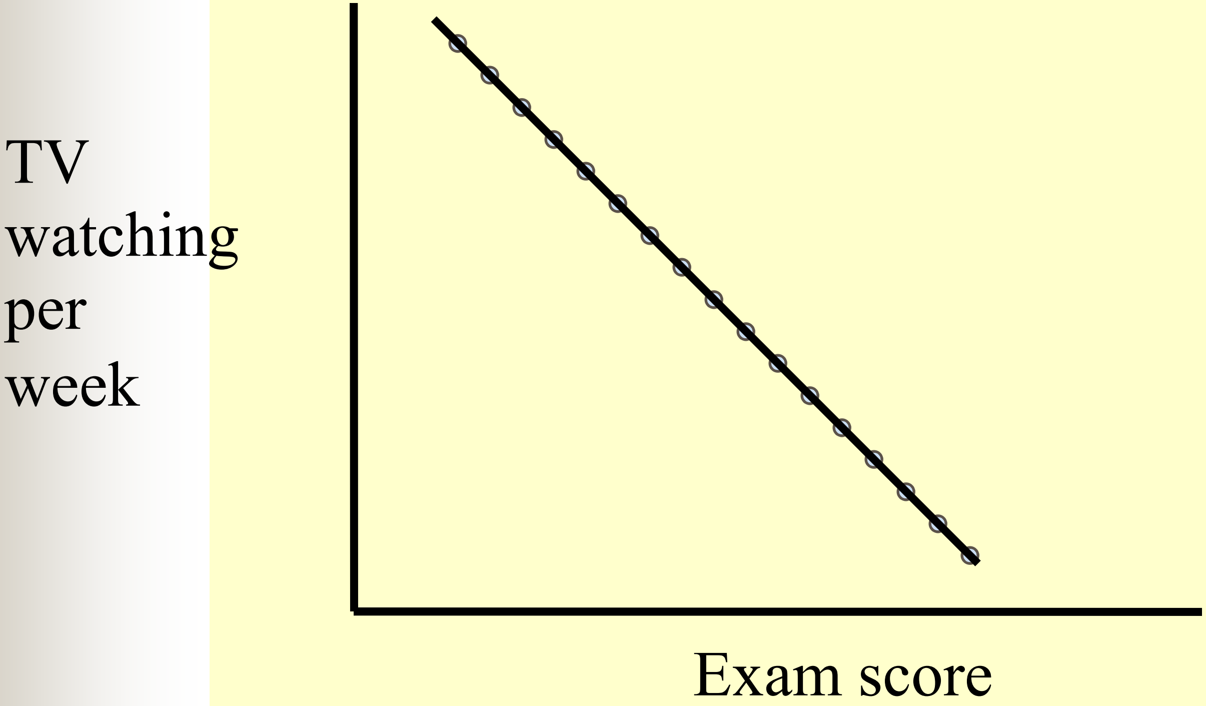
- **Moderate Positive Correlation**



r = + 0.4

Degree of correlation

- **Perfect Negative Correlation**

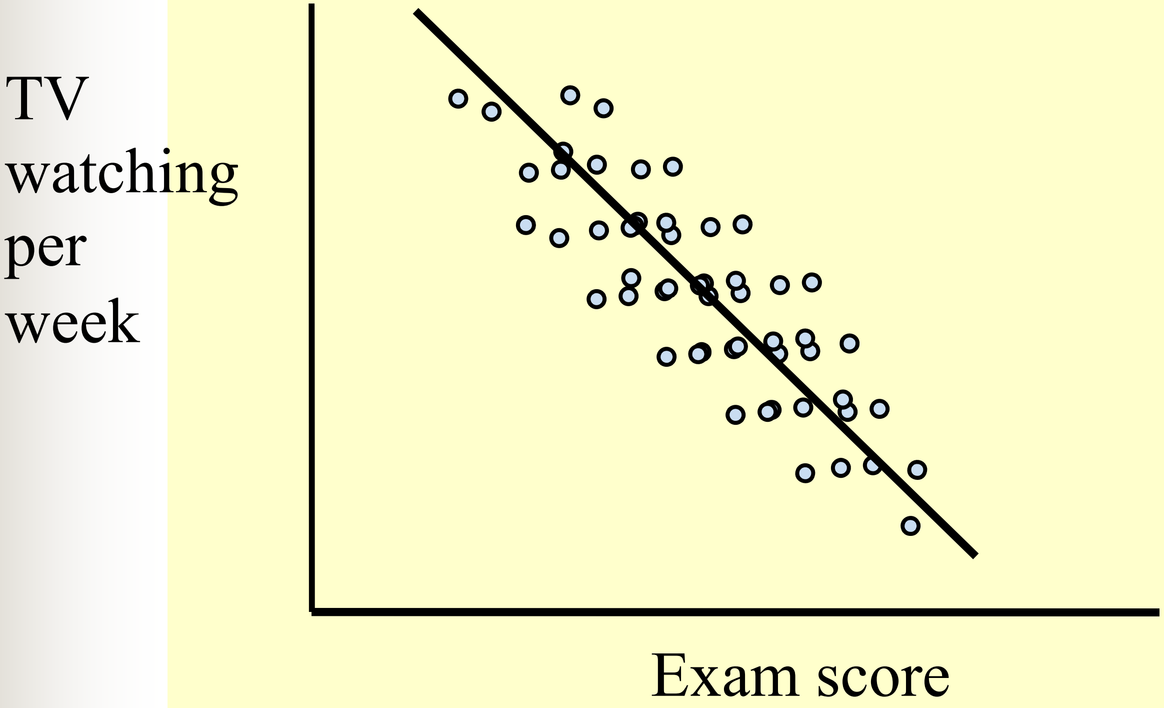


r = -1.0

Degree of correlation

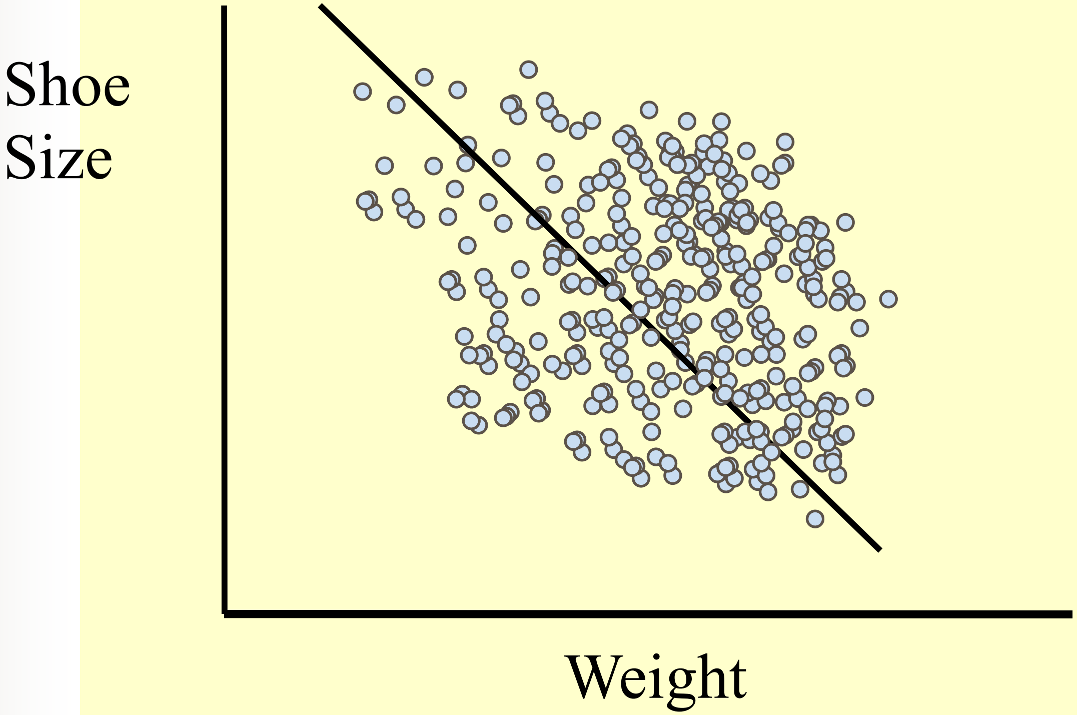
- **Moderate Negative Correlation**

r = -.80



Degree of correlation

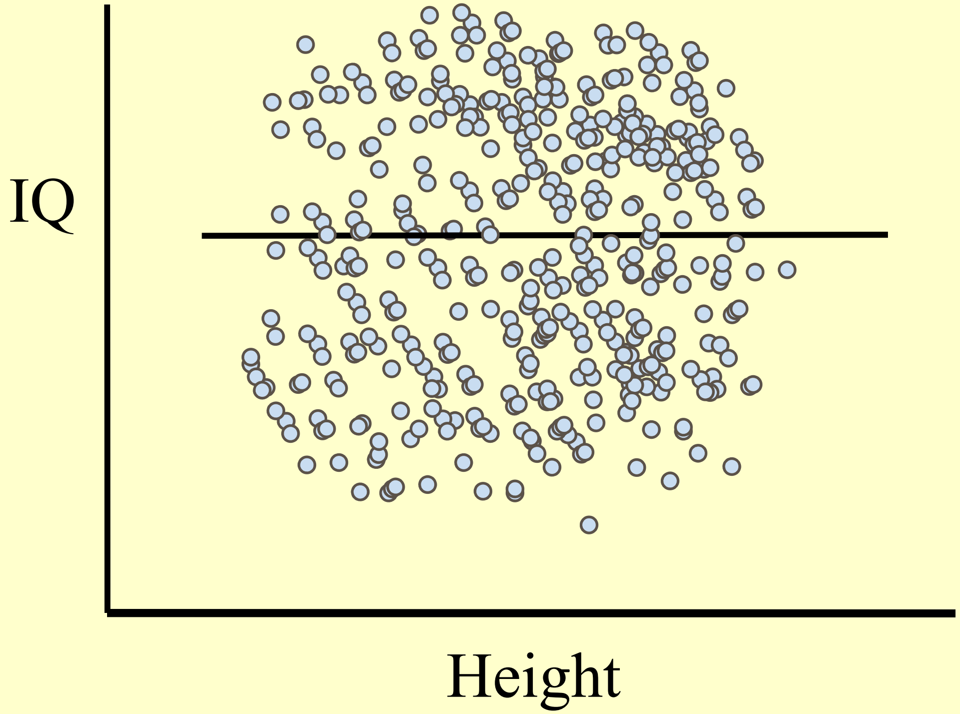
- **Weak negative Correlation**



r = - 0.2

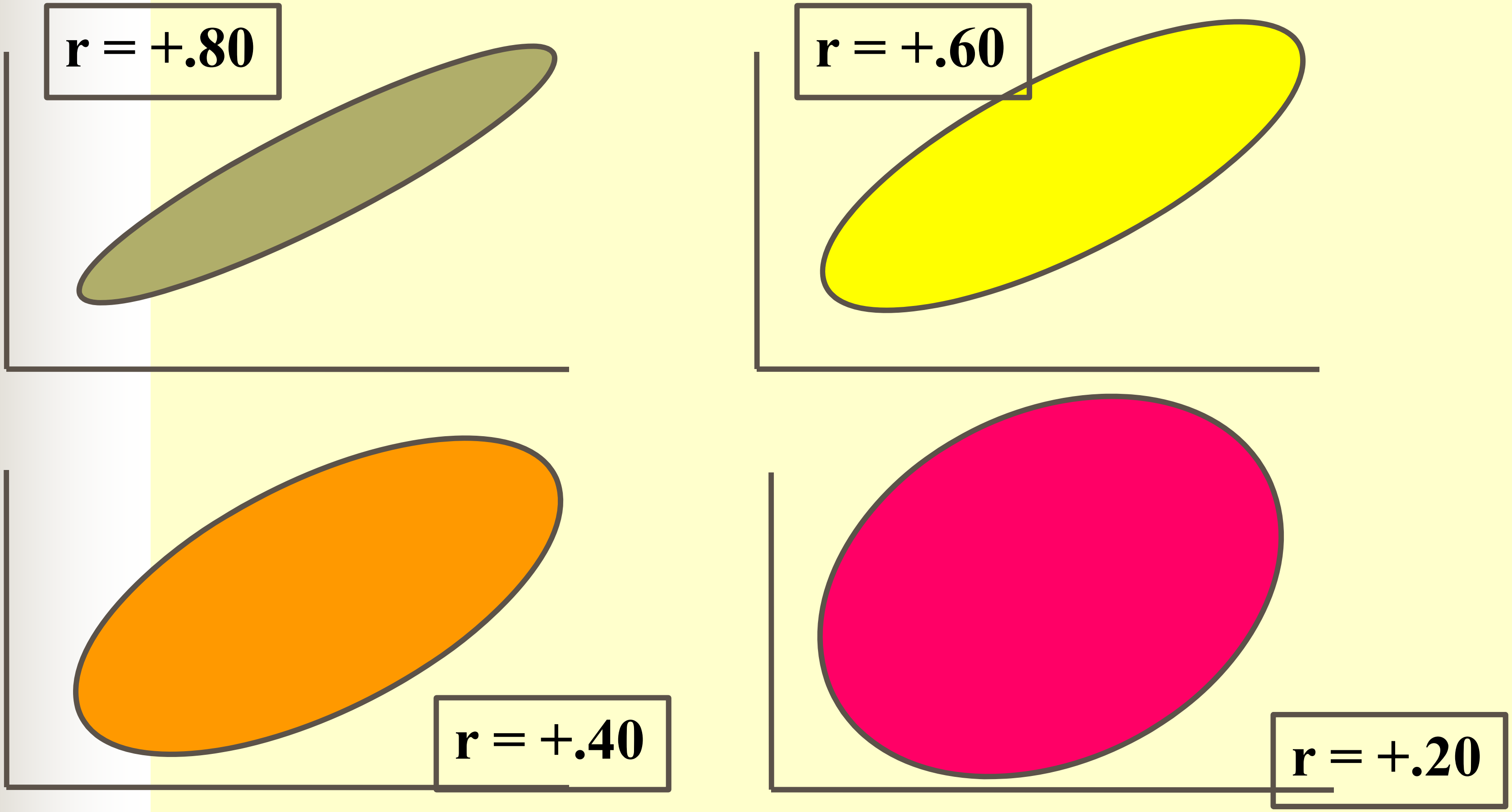
Degree of correlation

- **No Correlation (horizontal line)**



r = 0.0

Degree of correlation (r)



2) Direction of the Relationship

- **Positive relationship** – Variables change in the same direction.
    - As X is increasing, Y is increasing
    - As X is decreasing, Y is decreasing
  - E.g., As height increases, so does weight.

Indicated by sign; (+) or (-).

- **Negative relationship** – Variables change in opposite directions.
    - As X is increasing, Y is decreasing
    - As X is decreasing, Y is increasing
  - E.g., As TV time increases, grades decrease

Advantages of Scatter Diagram

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

Disadvantage of scatter diagram

Can not adopt the an exact degree of correlation

Karl Pearson's Coefficient of Correlation

- Pearson’s ‘r’ is the most common correlation coefficient.
- Karl Pearson’s Coefficient of Correlation denoted by- ‘r’ The coefficient of correlation ‘r’ measure the degree of linear relationship between two variables say x & y.

Karl Pearson's Coefficient of Correlation

- Karl Pearson’s Coefficient of Correlation denoted by- r

  $-1 \leq r \geq +1$

- Degree of Correlation is expressed by a value of Coefficient
- Direction of change is Indicated by sign (- ve) or (+ ve)

Karl Pearson's Coefficient of Correlation

- When deviation taken from actual mean:

$$r(x, y)= \Sigma xy / \sqrt{\Sigma x^2 \Sigma y^2}$$

- When deviation taken from an assumed mean:

$$r = \frac{N \Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{N \Sigma dx^2-(\Sigma dx)^2} \sqrt{N \Sigma dy^2-(\Sigma dy)^2}}$$

Procedure for computing the correlation coefficient

- Calculate the mean of the two series ‘x’ &’y’
- Calculate the deviations ‘x’ &’y’ in two series from their respective mean.
- Square each deviation of ‘x’ &’y’ then obtain the sum of the squared deviation i.e.$\sum x^2$ & .$\sum y^2$
- Multiply each deviation under x with each deviation under y & obtain the product of ‘xy’.Then obtain the sum of the product of x , y i.e. $\sum xy$
- Substitute the value in the formula.

Interpretation of Correlation Coefficient (r)

- The value of correlation coefficient ‘r’ ranges from -1 to +1
- If r = +1, then the correlation between the two variables is said to be perfect and positive
- If r = -1, then the correlation between the two variables is said to be perfect and negative
- If r = 0, then there exists no correlation between the variables

Properties of Correlation coefficient

- The correlation coefficient lies between -1 & +1 symbolically $(- 1 \leq r \geq 1)$
- The correlation coefficient is independent of the change of origin & scale.
- The coefficient of correlation is the geometric mean of two regression coefficient.

$$r = \sqrt{bxy * byx}$$

The one regression coefficient is (+ve) other regression coefficient is also (+ve) correlation coefficient is (+ve)

Assumptions of Pearson's Correlation Coefficient

- There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- Cause and effect relation exists between different forces operating on the item of the two variable series.

Advantages of Pearson’s Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

Limitation of Pearson’s Coefficient

- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming methods

Coefficient of Determination

- The convenient way of interpreting the value of correlation coefficient is to use of square of coefficient of correlation which is called Coefficient of Determination.
- The Coefficient of Determination = r^2.
- Suppose: $r = 0.9$, $r^2 = 0.81$ this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.

Coefficient of Determination

- The maximum value of r^2 is 1 because it is possible to explain all of the variation in y but it is not possible to explain more than all of it.
- Coefficient of Determination = Explained variation / Total variation

Coefficient of Determination: An example

- Suppose: $r = 0.60$

  $r = 0.30$ It does not mean that the first correlation is twice as strong as the second the 'r' can be understood by computing the value of r^2 .

When $r = 0.60$ $r^2 = 0.36$ -----(1)

$r = 0.30$ $r^2 = 0.09$ -----(2)

This implies that in the first case 36% of the total variation is explained whereas in second case 9% of the total variation is explained .

Spearman’s Rank Coefficient of Correlation

- When statistical series in which the variables under study are not capable of quantitative measurement but can be arranged in serial order, in such situation pearson’s correlation coefficient can not be used in such case Spearman Rank correlation can be used.
- **$R = 1- (6 \sum D^2) / N (N^2 – 1)$**
- R = Rank correlation coefficient
- D = Difference of rank between paired item in two series.
- N = Total number of observation.

Interpretation of Rank Correlation Coefficient (R)

- The value of rank correlation coefficient, R ranges from -1 to +1
- If R = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction
- If R = -1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction
- If R = 0, then there is no correlation

Rank Correlation Coefficient (R)

a) Problems where actual rank are given.

1) Calculate the difference ‘D’ of two Ranks i.e. (R1 – R2).

2) Square the difference & calculate the sum of the difference i.e. $\sum D^2$

3) Substitute the values obtained in the formula.

Rank Correlation Coefficient

b) **Problems where Ranks are not given :**If the ranks are not given, then we need to assign ranks to the data series. The lowest value in the series can be assigned rank 1 or the highest value in the series can be assigned rank 1. We need to follow the same scheme of ranking for the other series.

Then calculate the rank correlation coefficient in similar way as we do when the ranks are given.

Rank Correlation Coefficient (R)

- **Equal Ranks or tie in Ranks:** In such cases average ranks should be assigned to each individual. **R = 1- (6 $\sum D^2$) + AF / N (N^2 – 1)**

AF = 1/12(m_1^3 – m_1) + 1/12(m_2^3 – m_2) +…. 1/12(m_2^3 – m_2)

m = The number of time an item is repeated

Merits Spearman’s Rank Correlation

- This method is simpler to understand and easier to apply compared to karl pearson’s correlation method.
- This method is useful where we can give the ranks and not the actual data. (qualitative term)
- This method is to use where the initial data in the form of ranks.

Limitation Spearman’s Correlation

- Cannot be used for finding out correlation in a grouped frequency distribution.
- This method should be applied where N exceeds 30.

Advantages of Correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables under study.
- Can be used in many places, including natural settings, libraries, etc.
- Easier to collect co relational data

Regression Analysis

- Regression Analysis is a very powerful tool in the field of statistical analysis in predicting the value of one variable, given the value of another variable, when those variables are related to each other.

Regression Analysis

- Regression Analysis is mathematical measure of average relationship between two or more variables.
- Regression analysis is a statistical tool used in prediction of value of unknown variable from known variable.

Advantages of Regression Analysis

- Regression analysis provides estimates of values of the dependent variables from the values of independent variables.
- Regression analysis also helps to obtain a measure of the error involved in using the regression line as a basis for estimations .
- Regression analysis helps in obtaining a measure of the degree of association or correlation that exists between the two variable.

Assumptions in Regression Analysis

- Existence of actual linear relationship.
- The regression analysis is used to estimate the values within the range for which it is valid.
- The relationship between the dependent and independent variables remains the same till the regression equation is calculated.
- The dependent variable takes any random value but the values of the independent variables are fixed.
- In regression, we have only one dependant variable in our estimating equation. However, we can use more than one independent variable.

Regression line

- **Regression line** is the line which gives the best estimate of one variable from the value of any other given variable.
- **The regression line** gives the average relationship between the two variables in mathematical form.
- **The Regression would have the following properties:** a) $\sum(Y - Y_c) = 0$ and

  b) $\sum(Y - Y_c)^2 = \text{Minimum}$

Regression line

- For two variables X and Y, there are always two lines of regression –
- **Regression line of X on Y :** gives the best estimate for the value of X for any specific given values of Y
- **X = a + b Y** a = X - intercept
- b = Slope of the line
- X = Dependent variable
- Y = Independent variable

Regression line

- For two variables X and Y, there are always two lines of regression –
- **Regression line of Y on X** : gives the best estimate for the value of Y for any specific given values of X
- **Y = a + bx** a = Y - intercept
- b = Slope of the line
- Y = Dependent variable
- x= Independent variable

The Explanation of Regression Line

- In case of perfect correlation (positive or negative) the two line of regression coincide.
- If the two R. line are far from each other then degree of correlation is less, & vice versa.
- The mean values of X &Y can be obtained as the point of intersection of the two regression line.
- The higher degree of correlation between the variables, the angle between the lines is smaller & vice versa.

Regression Equation / Line & Method of Least Squares

- **Regression Equation of y on x**

$Y = a + bx$

In order to obtain the values of ‘a’ & ‘b’

$\sum y = na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

- **Regression Equation of x on y**

$X = c + dy$

In order to obtain the values of ‘c’ & ‘d’

$\sum x = nc + d\sum y$

$\sum xy = c\sum y + d\sum y^2$

Regression Equation / Line when Deviation taken from Arithmetic Mean

- **Regression Equation of y on x:**

  $$\mathbf{Y = a + bx}$$

  In order to obtain the values of ‘a’ & ‘b’

  $$a = \overline{Y} - b\overline{X} \qquad b = \sum xy / \sum x^2$$

- **Regression Equation of x on y:**

  $$\mathbf{X = c + dy}$$

  $$c = \overline{X} - d\overline{Y} \qquad d = \sum xy / \sum y^2$$

Regression Equation / Line when Deviation taken from Arithmetic Mean

- Regression Equation of y on x:

$$Y - \overline{Y} = b_{yx} (X - \overline{X})$$

$$b_{yx} = \sum xy / \sum x^2$$

$$b_{yx} = r (\sigma y / \sigma x)$$

- Regression Equation of x on y:

$$X - \overline{X} = b_{xy} (Y - \overline{Y})$$

$$b_{xy} = \sum xy / \sum y^2$$

$$b_{xy} = r (\sigma x / \sigma y)$$

Properties of the Regression Coefficients

- The coefficient of correlation is geometric mean of the two regression coefficients. $\mathbf{r = \sqrt{b_{yx\,*}\,b_{xy}}}$
- If b_{yx} is positive than b_{xy} should also be positive & vice versa.
- If one regression coefficient is greater than one the other must be less than one.
- The coefficient of correlation will have the same sign as that our regression coefficient.
- Arithmetic mean of $b_{yx\,\&}\,b_{xy}$ is equal to or greater than coefficient of correlation. $b_{yx\,+}\,b_{xy\,/\,2\,\geq}\,r$
- Regression coefficient are independent of origin but not of scale.

Standard Error of Estimate.

- Standard Error of Estimate is the measure of variation around the computed regression line.
- Standard error of estimate (SE) of Y measure the variability of the observed values of Y around the regression line.
- Standard error of estimate gives us a measure about the line of regression. of the scatter of the observations about the line of regression.

Standard Error of Estimate.

- **Standard Error of Estimate of Y on X is:**

$$\text{S.E. of Yon X } (SE_{xy}) = \sqrt{}\sum(Y - Y_e)^2 / n\text{-}2$$

Y = Observed value of y

Y_e = Estimated values from the estimated equation that correspond to each y value

e = The error term ($Y - Y_e$)

n = Number of observation in sample.

- **The convenient formula:**

$$(SE_{xy}) = \sqrt{}\sum Y^2 - a\sum Y - b\sum YX / n - 2$$

X = Value of independent variable.

Y = Value of dependent variable.

a = Y intercept.

b = Slope of estimating equation.

n = Number of data points.

Correlation analysis vs. Regression analysis.

- Regression is the average relationship between two variables
- Correlation need not imply cause & effect relationship between the variables understudy.- R A clearly indicate the cause and effect relation ship between the variables.
- There may be non-sense correlation between two variables.- There is no such thing like non-sense regression.

Correlation analysis vs. Regression analysis.

- Regression is the average relationship between two variables
- R A.

What is regression?

- Fitting a line to the data using an equation in order to describe and **predict** data
- **Simple Regression**
  - Uses just 2 variables (X and Y)
  - Other: Multiple Regression (one Y and many X's)
- **Linear Regression**
  - Fits data to a straight line
  - Other: Curvilinear Regression (curved line)

We're doing: Simple, Linear Regression

From Geometry:

- Any line can be described by an equation
- For any point on a line for X, there will be a corresponding Y
- the equation for this is y = mx + b
  - m is the slope, b is the Y-intercept (when X = 0)
- **Slope** = change in Y per unit change in X
- **Y-intercept** = where the line crosses the Y axis (when X = 0)

Regression equation

- Find a line that fits the data the best, = find a line that minimizes the distance from all the data points to that line
- Regression Equation: $\hat{Y}$(Y-hat) = bX + a
  - Y(hat) is the predicted value of Y given a certain X
  - b is the slope
  - a is the y-intercept

Regression Equation:

$$\boxed{Y = .823X + -4.239}$$

- We can predict a Y score from an X by plugging a value for X into the equation and calculating Y
  - What would we expect a person to get on quiz #4 if they got a 12.5 on quiz #3?

$$\boxed{Y = .823(12.5) + -4.239 = 6.049}$$

Advantages of Correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables studied
- Can be used in many places, including natural settings, libraries, etc.
- Easier to collect correlational data