Measures of Relationship

The world is made of co-variation! Everywhere we look, we see patterns. From the spread of stars and the orbits of planets to the pathways of ants and details of a microchip, we can see that things go together in predictable ways. The co-variation between a set of variables provides the underlying building blocks for all the major types of statistical modelling**.** ***Measures of Relationship or Association*** refer to a wide variety of coefficients that measure a relationship's strength, defined in various ways.

In our life, the values of various variables are interrelated.

Example: when *our* ***height*** *increases,* ***weight*** *also increases, as* ***the price of good*** *increases* ***demand*** *decreases, when* ***the age of the husbands*** *increases, the* ***age of the wife*** *also increases.'*

1.2 Correlation

The ideas of relationships between variables are measured using statistics known as the **correlation coefficient**. The two variables are said to be related if one variable changes in any manner whenever the other variables vary in a particular manner. The concept of correlation has a **philosophical foundation**. Correlation at the descriptive level indicates that whatever phenomenon varies in a particular manner, another phenomenon varies in a particular manner, and then these two phenomena are correlated.

Correlation measures relationships. The correlation technique is decided by the type of scales used in the variables measured. It also serves as a basis for inferring causation, to explain which variable could be the cause and which could be the effect. A correlation is a statistical method used to measure and describe the relationship between two variables. A relationship exists when changes in one variable tend to be accompanied by consistent and predictable changes in the other variable.

A correlation typically evaluates three aspects of the relationship*: the direction, the form, the degree*. The **direction** of the relationship is measured by the sign of the correlation (+ or -). A positive correlation means that the two variables tend to change in the same direction; as one increases, the other also tends to increase. A negative correlation means that the two variables tend to change in opposite directions; as one increases, the other tends to decrease. The most common **form** of relationship is a straight line or linear relationship measured by the Pearson correlation. The ***degree*** of the relationship (the strength or consistency of the relationship) is measured by the numerical value of the correlation. *A value of 1.00 indicates a perfect relationship and a value of zero indicates no relationship.* To compute a correlation, we need two scores, X and Y, for each individual in the sample.

1.2.1 Linear Correlation

This is the simplest kind of correlation to be found between the two sets of scores or variables. When the relationship between two sets of scores or variables can be represented graphically by a straight line, it is known as a ***linear correlation***. Such a type of correlation reveals the change in one variable is accompanied by a change or to what extent an increase or decrease in one is accompanied by the increase in or decrease in the other. The correlation between two sets of measures of variables can be positive or negative. It is said to be positive when one variable's increase (decrease) corresponds to an increase or decrease in the other. It is said to be **negative** when an increase corresponds with a decrease. There is also the possibility of a third type of correlation which is zero correlation between the two sets of measures of variables if there exists **no relationship** between them.

1.2.2 Partial Correlation

If two variables are not related directly but are related through a third variable, then two variables are said to have a **partial correlation.**

Example: *The two variables **"amount of rainfall"** and "**prices of food grains"** are related through a third variable **"food grain production** ". The relationship existing between amount of rainfall and "prices of food grains" is of partial correlation.*

1.2.3 Positive, negative and zero correlation

Consider the example of the height of men as one phenomenon and the arm-length as the other phenomenon. As the **height** increases correspondingly, the **arm –length** also increases.

Positive correlation

***Height** increases, the **arm –length** also increases.*

Height of men and the arm –length of men

The taller the person, the longer his arm will be. These two phenomenon variables change together concomitantly. They are correlated; the changes in both variables are in the same direction.

When two variables are related, they vary; the variations may be in the same or different directions. If the variables change in the same direction, the correlation is positive; if they change in a different direction, it is said to be negative.

Example: **Zero correlation**

Variable intelligence and the skin-color of students

Think of the variable intelligence and the skin colour of students. They do not change together: they are not related, and they are not correlated. They have zero correlation because the two variables change uniformly in the same direction in affixed proportion. Here we can perfectly say what would be the change in one variable for a given change in the other variable without actually measuring the change.

When two variables change concomitantly in **the same direction,** the variables are positively correlated. When these changes follow a **fixed proportion**, the relationship (correlation) between them is perfectly positive. Suppose they change in **opposite** (different) directions in the same proportion. Then the correlation is perfectly negative.

Example: *Perfect negative correlation: speed and time, volume and pressure.*

As one increases, the other decreases in a fixed proportion. They are perfectly negatively correlated two variables. Therefore, the relationship between two variables then may be of three types.

- *If the variables change together in magnitude in the same direction, the relationship is* ***positive.***
- *If the variables change in magnitude in the opposite direction (as one increases, the other decreases), the relationship is* ***negative.***
- *If the changes in the magnitude of one variable tell nothing about the likely condition of the other variable, there is no (**zero**) relationship between them.*

16.2.4 Perfect Correlations

If two variables change together (as scores on one variable deviate from their mean, scores on the other also predictably deviate from their mean), we say the two sets of scores "co-vary"; they vary together. The extent to which the two sets of scores co-vary, that is, the extent of their variability can be quantified. If proportionate changes in the other variable accompany the changes in one variable, that correlation is perfect. A perfect correlation is quantified as one. So, the *perfect positive correlation is plus one, and the perfect negative correlation is minus one.*

If two variables change together in the same direction (as one increases, the other also increases, or as one increases, the other also decreases) and if the change is not proportionate, the correlation could be better. So, it is not plus one; it can only be less than one; it may be 0.8, 0.7 or anything below but above zero.

Similarly, negative relationships may be anything above -1 and below 0. If the variables change in the same direction, the measure of the extent of the relationship is positive; if they change in the opposite direction, then the correlation is negative.

Think of a relationship between time and distance covered. If the speed is uniformly 20k.m per hour, the distance covered by 1 hour, 2 hours, 3 hours etc., will increase by 20 hours for every one-hour increase. If the time increases by 1 unit (1 hour), the distance covered will increase by 20km. Both time and distance increase and they change in the same direction. Therefore, the relationship (correlation) is positive. The two variables (time and distance) increase in a fixed proportion (1:20). **The relationship here is positive and also perfect.**

This change in one variable affects the change in the other variable. Two variables are said to be related when a change in one variable affects the change in the other. In statistics, the degree to which two or more attributes or measurements on the same group of elements show a tendency to vary together.

Correlation is a statistical measurement of the relationship between two variables. **Possible correlations range from +1 to –1.** A zero correlation indicates that there is no relationship between the variables. A correlation of –1 indicates a **perfect negative** correlation, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that both variables move in the same direction. The degree or amount of relationship between the two variables is measured by the known correlation coefficient, whose values range from -1 to+1. In statistics, correlation refers to the relationship between two variables. When the value of one variable varies closely with the variation in another, the two variables are said to be correlated.

Example: *If the sale of umbrellas in a town during the rainy season varies with the amount of rainfall, then the sale of umbrellas and the amount of rainfall are correlated.*

Correlation coefficient is a statistical measure of how close this relationship is. A correlation coefficient of 1 indicates a perfect or total dependence between two variables. For example, if we were to calculate the correlation coefficient between temperature in Fahrenheit and Celsius, it will be equal to 1. Correlation coefficients of -1 mean the variables are inversely correlated.

Example: *The volume of a given weight of different substances is inversely proportional to their densities; therefore, the correlation coefficient between these two will be -1.*

A correlation coefficient of 0 implies no correlation. But in reality, this only implies that there is no linear relationship. For example, some graphs of two related variables have shapes like 'U' or 'inverted U'. In these cases, the relationship exists but is not linear. Therefore, the correlation coefficient is likely to be close to 0.

1.3 Pearson’s Product Moment Correlation

Pearson’s Product Moment Correlation is used to quantify the relationship between two variables and find the amount of correlation.

We know that when changes in the other variables accompany the changes in one variable, both are correlated. We have several changing measures of the same variable from different persons in a group. (For example, varying marks of achievement of the students in mathematics). We **require a reference point.** The mean of this measure provides this reference point. So, the changes in the measures of one variable are considered with reference to the mean of the measures of the other variable. In other words, the changes are the deviations (differences) from the mean.

Similarly, the changes in the measures from the second variable are also considered with reference to the mean of the measures of the second variable. For changes, we consider the deviations. We know that z scores are the deviation scores expressed in terms of standard deviations. Comparisons of two measures possibly become truly meaningful when expressed in z scores. When we want to know whether changes (i.e.) deviations in one variable are accompanied by changes (that is, deviations) in another variable and quantify the extent of changes, we use the z scores of the two variables. So, *to find the relationship between two sets of scores (variable), we use the z scores of these two sets.*

To quantify the relationship between two variables, we require the measures of the same persons or things on the two variables. *For example, emotional and social intelligence scores of the same person and this way for a group of persons.* For each subject, we will have two measures (two scores). The changes (deviations) of these two scores from the respective means can be expressed as z scores to make other calculations meaningful. By converting them into z scores, we express them in the same units (and also know that z scores can be added, subtracted, multiplied or put into any arithmetical operation meaningfully, even though the original measures may be in varying units.

1.4 Steps to Find Out the Extent of Relationship between Two Variables

- Express the measures of the variables in the respective z scores
- Cross multiply the z scores (i.e. each student, there will be two measures (English marks and Tamil marks and therefore, each student will have two z scores, one for English and the other for Tamil. We multiply these two z scores. This is known as cross multiplication.)
- Find the means of the products of the z scores, which gives us the extent of the relationship. This mean of the products of the z scores of the two measures of individuals in a group is known as the product-moment correlation coefficient denoted by the symbol ”r.”

$$r = \sum \frac{Z_1 Z_2}{N}$$

1.5 Perfect (Maximum) Correlation Co-Efficient is one

We know the standard deviation of z scores is always one. Their mean is zero. Z-scores are deviation scores. Standard deviation is the root-mean-squared deviation.

$$S = \sqrt{\frac{\sum x^2}{N}}$$

Z_1Z_2 is maximum when $Z_1 = Z_2$,

That is only when a student gets his z scores in the two subjects are equal (but the z scores of the students may and often shall vary)

then $\sum Z_1Z_2$ will be the maximum, and

$$r = \sum \frac{Z_1 Z_2}{N}$$

will also be maximum

When $Z_1 = Z_2$ then Z_1Z_2 will be Z_1^2 or Z_2^2

Then

$$r = \sum \frac{Z_1 Z_2}{N}$$

$$= \sum \frac{Z^2}{N}$$

And we know z is a deviation score.

So $\sum \frac{Z^2}{N}$ will be the mean squared deviation of z scores.

The root mean square deviation of z scores (s of Z score) is one.

So, the mean squared deviation will be 1^2.

Mean squared deviation is s^2 or variance.

If s is 1, then s will be $1^2=1$.

So, the value of $\sum \frac{Z^2}{N}$ will be equal to one and

we know this is the maximum value that $\sum \frac{Z_1 Z_2}{N}$ will have.

So the maximum value that r can have is only 1.

If the z scores of a student in the two sets of subjects are in different directions, that is, if one z score is plus and the other minus, and if they are equal and if this is the case of all the students, then we will have the maximum

$\sum Z_1 Z_2$ value but this will be negative.

$[(+Z_1)\,(-Z_2) = -Z_1 Z_2]$

So we get the maximum negative correlation that is -1.

1.6 Product Moment r from Raw Scores

To find the correlation coefficient, we have to calculate the z scores for each student in the two subjects, cross multiply them and find the mean of z score products and to find the z score, we need mean and standard deviation. (We know z= $\frac{X-\bar{X}}{S}$). so there is a method adopted which you can calculate all these statistics using only five values. So from the data we collect, first calculate these five values and using the required values in different formulae we can calculate mean s and r.

The five values are $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum x\,y$.

Raw Score Formula

$$r = \frac{N \sum x\,y - \sum x \sum y}{}$$

$$\sqrt{[N\sum x2\ -\ (\sum x\ )2\ ][N\sum y2\ -(\sum y)2\ ]}$$

1.6.1 Product moment 'r' for the grouped data

The above formula is used for smaller number of cases. For larger number of cases, we group the data that is form a correlation matrix or scatergram or bivariate distribution. The correlation matrix we develop will represent two variables. So it is a bivariate distribution.

Take convenient class intervals and represent one variable in columns and the other in rows. The size of the class intervals and the number of class intervals need not be the same for two variables

Step-1: Decide the size of the class interval for each variable. The sizes need not be the same

Step-2: Determine the matrix

Step- 3: 'r' for bivariate distribution is

$$r\ =\ \frac{N\sum f\,x\,y-\ (\sum fx)(\sum y)}{\sqrt{[N\sum fx2\ -\ (\sum fx\ )2\ ][N\sum fy2\ -(\sum fy)2\ ]}}$$

1.6.2 Spearman's rho (r_s)

The **Spearman correlation** is used in two general situations:

- It measures the relationship between two ordinal variables, X and Y, consisting of ranks.
- It measures the consistency of the direction of the relationship between two variables. In this case, the two variables must be converted to ranks before the Spearman correlation is computed.

The calculation of the Spearman correlation requires the following steps:

- Two variables are observed for each individual.

- □ The observations for each variable are rank ordered. The X values and the Y values are ranked separately.
- □ After the variables have been ranked, the Spearman correlation is computed by either:
  - ❖ Using the Pearson formula with the ranked data.
  - ❖ Using the special Spearman formula

(assuming there are few, if any, tied ranks).

1.7 Biserial Correlation and Point Biserial Correlation

Some variables can be measured using continuous scales. For example, height, weight and achievement can be measured using continuous scales. They are continuous variables. Some variables cannot be measured using a continuous scale. For example, sex and locality cannot be measured using a continuous scale. There is no continuity between males and females. A person He or she cannot move on a continuum from being a male to being a male. Such a variable is known as **dichotomous variable**. The other

Examples of dichotomous variables are being a farmer and not being a farmer, living and being dead, married and unmarried, and owning a car and not owning a car.

Is there a relationship between body weight and the sex of high school students?

Weight is a continuous variable, and sex is a dichotomous variable. The correlation between these two variables cannot be calculated using either Pearson's variable or the spearman technique. ***Pearson requires both variables to be continuous,*** *and* ***spearman's 'rho' insists that both measures should be ordinal***. Here we have one variable (that is, height continuous and the other dichotomous (that is, sex).In such cases, we can use another technique known as **point biserial correlation.**

Suppose we want to find out the relationship between the height of the students and their passing or failing the examination. Here height is a continuous variable. Passing or failing is a dichotomy. There lies a continuum of scores. This distribution of continuous scores is divided into two(dichotomised). Those whose scores are above a point in the achievement continuum are classified as pass, and those below that point are classified as fail. Pass-fail is not dichotomous but a **dichotomised variable**, whereas boy-girl is a dichotomous variable proper. So, in our second example, we want to find out the relationship between a continuous and a dichotomized variable. In the first example (weight and sex) the correlation is between a continuous and a dichotomous variable.

When we want to find to find out correlation between a continuous and a dichotomized variable the appropriate correlation statistic is **biserial correlation.** The biserial and point biserial correlation coefficients are **distinguished by only a conceptual difference**. These correlation coefficients are used when one of the two variables is dichotomous (i.e. it is categorical with only two categories). A dichotomous variable is one for which there are exactly two categories.

Example: *A dichotomous variable* ***being pregnant*** *because a woman can be either pregnant or not (she cannot be a bit pregnant) Men/women or succeed/fail.*

Often it is necessary to investigate relationships between two variables when one of the variables is dichotomous. The difference between the use of biserial and point biserial correlations depends on whether the dichotomous variable is **discrete or continuous**. This difference is slight. A discrete or true dichotomy is one for which there is no underlying continuum between the two categories.

Example: Someone is dead or alive.

Whether someone is dead or alive? Here there is no continuum between the two categories. A person can either be dead or alive; he cannot be a bit dead only, although a person can be half dead. (He will still be breathing). However, there is a possibility of a continuum

Example: Passing or failing of a statistics test.

Passing or failing a statistics test. Some people will fail, while others will fail by a large margin. Likewise, some people will scrape a pass whilst others excel. So, although participants fall into two categories, ***there is an underlying continuum*** along which people lie. It is clear that, in this case, there is some continuum underlying the dichotomy because some people passed or failed more dramatically than others. ***The point-biserial correlation coefficient is used when one variable is a discrete dichotomy, whereas the biserial correlation coefficient is used when one variable is a continuous dichotomy.***

The Pearson correlation formula can also measure the relationship between two variables when one or both are dichotomous. If achievement can be dichotomized as pass, fail, why not height into tall and short? *Fix an acceptable arbitrary point* and call those above ‘tall’ and those below as ‘short’. Now you can have a dichotomized height also. Tall short is dichotomized variable because behind this lies a continuum, that is, height. We can find the correlation between two dichotomized variables also, as in the case of tall short and pass-fail. For this, we use **tetra chloric correlation**.

Suppose we have two dichotomous variables, say sex and colour blindness. It can either be male or female or either colour blind or not. In such cases, the appropriate correlation statistic is **phi coefficient.**

1.8 Correlation and causation

Correlation does not indicate causation. Correlation says that there is an association between the two variables. It does not say that one variable is the cause and the other is the effect. The relationship indicated by correlation is only associative; it is not causal. *Causal relation implies necessary and sufficient conditions.* Correlation does not imply these conditions. A causal relation includes correlation. Naturally, the cause and effect are related. But the mere association, mere concomitant changes in the two variables, need not imply cause and effect relationship. The relationship may be due to another common factor influencing both variables.

Example: Correlation between economics and history scores

Between economics and history scores, suppose the correlation we found was ▪81. Is there a cause-and-effect relationship between these two sets of scores? Is Economics the cause and history the effect? Suppose the students study economics better and increase their scores in Economics. Will the increase in economics affect the increase in the history scores automatically? No, it will not. Economics is not a cause; change in it will not produce a given effect in History. The relationship that correlation that we observed is relative to the situation under which it is observed or obtained.

1.8.1 Significance of ‘r.’

The relationship between two variables may be a chance factor or real. Suppose the amount of relationship (the value of correlation calculated) is so much that it cannot occur due to the chance factor. In that case, the relationship is said to be significant. The significant correlation (the amount required to decide that the relationship is real and not due to a chance factor) *depends on the number of cases.* we interpret “r” we have to consider the number of cases also. The interpretation of ‘r” has to base on the type of relationship that one normally expects to exist in a given situation.

For example, if we consider the relationship between height and length, an r of .5 will not be considered high, even though it may not be statistically significant. (i.e. it may be obtained even with a large number of cases, say 300 and so.). We usually expect a near-perfect correlation between height and arm length. Even the r=.5 correlation is very low and even meaningless.

For general purposes, we interpret r as follows

Below . 2	Very low
.2 to . 4	Low
.4 to .6	Average
.6	high
.8 to 1	Very high

The general principle to be followed in the interpretation of r is that r is “purely relative to the circumstances in which it was obtained and should be interpreted in the light of those circumstances, very rarely, certainly, in any absolute sense’ **(Guilford).**

Prediction and Regression

Prediction is based on correlation. When the correlation is perfect, the predictions will be exact. The correlated variables vary together. The extent of correlation is expressed as a correlation coefficient. The square of the correlation coefficient indicates the percentage of variance of one variable explained by the variance of the other correlated variable. For example if the correlation between emotional intelligence of teachers and teaching effectiveness is .6, then $.6^2$ =.36, that is 36 percent of variance in teaching effectiveness is explained by the variance (change) in emotional intelligence of teachers. We can also say that the variance in emotional intelligence explains

36% of variance in teaching effectiveness. Prediction works only with the product moment correlation only.

When the correlation is not perfect, the points will not lie on the straight line, but will lie about a straight line. Our predictions will not be completely perfect. When there is no correlation that is when there is zero correlation prediction will be futile. If at all we predict, when the correlation is zero, the best prediction would be the mean of the second variable. The mean is the representative score of a variable. For any given value of the first variable, when correlation is zero, the mean of the second variable will be the best predictor. When the correlation becomes higher and higher, the prediction will also become more and more perfect. When r is not equal to 1 and is above zero, predictions are not perfect.

But we can think of best possible predictor if not perfect, when "r" takes different values between 0 and 1. When r=0, the best possible prediction is the mean of the second variable. When r=1, we can predict the exact value of the second variable for any given value in the first variable. ***That is when the correlation is perfect, the prediction is exact, when it is zero the prediction is the mean of the second variable. This tendency of the predicted value to move from the exact value when r=1 to the mean of the second variable when r=0 is known as regression.*** Regression is the tendency to move towards the mean. When r takes any value between 1 and 0, the values of the second variable move from its first "perfect" value (that is when r=1) to the mean of the second variable. This going back towards the mean is known as ***regression.***

The Latin regredi means to go back. 'Regression' (latin) means 'retreat', 'going back to', 'stepping back'. In a 'regression' we try to (stepwise) retreat from our data and explain them with one or more explanatory predictor variables. We draw a 'regression line' that serves as the (linear) model of our observed data. 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.

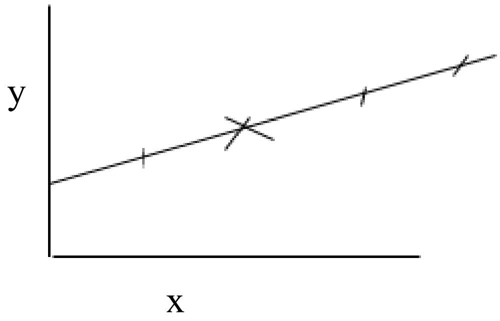
1 predictor=simple regression

>1 predictor=multiple regression

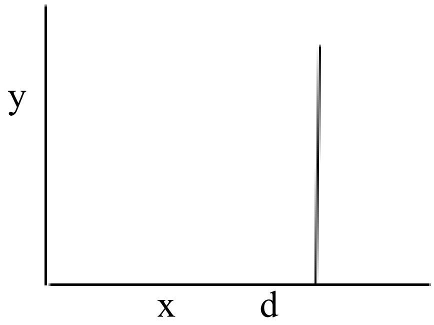
For a regression you do want to find out about those relations between variables, in particular, whether one 'causes' the other. Therefore, an unambiguous causal template has to be established between the causer and the causee before the analysis! This pattern is inferential. Regression is the statistical method underlying all inferential statistics (t-test, ANOVA, etc.). All that follows is a variation of regression.

16.9.1Regression Line

When r=1, the values of the two correlated variables when plotted on a graph will lie on the same line.

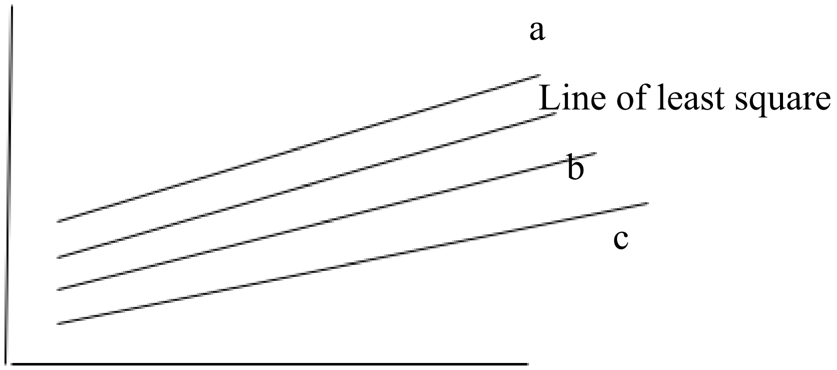


When r moves away from 1, the points indicating the values of the variables will not lie on a straight line.



We would like to predict the value of the y variable for the 'd' value of the x variable. But we do not find the corresponding value of the y variable in the graph. Only if there is a straight line joining the points, we can locate the value of one variable for the given value of one variable for another variable. Such straight lines are available only for perfect correlations, positive and negative. For simple correlations we have to find out straight lines based on, which predictions can be made. Such straight lines should help us to get the best possible **predictions**.

A line that is possible nearest to all possible points of the variables will give the best possible predictions. That is the lines should be at the least possible distance from all the points, therefore has to be drawn for prediction purposes. Naturally some points will be above this line and some below this line. The points above the line will have positive points and the distances those below the line will have negative distances. If so, the total distance as elsewhere (total deviation from the mean, total rank differences) will be zero. In order to escape from this zero we simply square the distances. Now we can find out a line that is at the least squared distance from the point (not all the least possible distance, but least squared distance possible).This line that is at the least squared distance from all the points is known as the line of least square.



Predictions are based on this line of least square. For each value of r we can have a line of least square. Least square lines are straight lines. Every straight line has an equation. In fact we do not draw graphs and predict values of one variable for the given values of the other correlated variable. We use the equation and make predictions. ***The line of least square is known as regression line and the equation of that line is known as regression equation.***

On a graph the perfect correlation will form a straight line. So if we want to predict the measure (value) of one variable (say the one represented on y-axis) corresponding to a given measure (value) of the other variable (say the one represented on the x-axis), we can easily find the required value by erecting a vertical line from the point on the x-axis that denotes the given measure and finding out the point on the straight line graph where the vertical line meet it and by finding the y-value of that point.

In graph, every straight line has an equation, if we know the equation of the straight line, then we can predict using the equation without drawing the graph. Every change in x variable is accompanied by proportionate change in the y variable. The correlation will be perfect and form a straight line on the graph. The equation of this straight line is y=2x. If we know this relationship then for any value of x, we can find the corresponding y value using this equation.

16.9.2 Regression Equation and Regression Coefficients

When the correlation is perfect, the z values of the corresponding x and y variable measures will be same. Suppose we want to predict the y-score for a given x score (that is mark in English for a given mark in Tamil) convert the x score into the respective z score. Let it be Z_x. Let the predicted y score be Z_y. When the correlation is perfect then the score for the given Z_x score will be Z_y' and this $Z_{y'}$ will be equal to $Z_{y.}$ When the correlation is zero, then the best predicted value of Y will be the mean of Y. The mean of Y, in Z-score form is zero (the mean of any value Z score set will be 0).So, when the correlation is zero, the predicted Y score for Z_x will be zero, that is Z_y will be zero.

When the correlation ranges between 1 and 0, the value of Z_y will range from Z_x to 0 and 0 is the mean of the y scores. The predicted values regress towards the mean. For various amounts of correlation, the best predictions of Z_y will be

$$Z_{Y'}=rZ_{X'}$$

If r=1, then $Z_Y' = Z_X$

If r=0, then $Z_{Y'=0}$ and 0 is the mean of the Z_Y scores.

As r comes down from 1 to zero, the amount of regression increases and this regression is towards the mean.

The general form of the regression equation then will be

$$Z_{Y'}$$

$Z_{Y'}$ (predicted) =r Z_X (given)

Similarly if we want to predict X from a given Y, the equation will be

$Z_{X'}$ (predicted) =r Z_y(given)

We have thus two regression equations, one to predict Y from given X and the other to Predict X from a Given Y when X and Y are correlated. **Predictions work one way only**: when the correlation is perfect: it will work both ways.

We Know that $Z_x = \dfrac{X-\overline{X}}{S_X}$

$Z_Y = \dfrac{Y-\overline{Y}}{S_Y}$

A predicted score in the Z form is $Z_Y{}'$ (predicted) =r Z_X (given)

Converting Z scores into raw scores, the equation becomes

$$Y - \overline{Y} = r\frac{S_X}{S_Y}(X - \overline{X})$$

and $$X - \overline{X} = r\frac{S_y}{S_x}(Y - \overline{Y})$$

16.9.3 Regression Coefficients

The regression equation in the deviation form becomes

$$Y - \overline{Y} = r\frac{S_X}{S_Y}(X - \overline{X})$$

The term $r\frac{S_X}{Sy}$ is known as regression co –efficient.

Usually it is denoted by “b” co-efficient.

16.9.4 What do we do in regression?

In a regression, the predictor variables are labelled **'independent' variables**. They predict the outcome variable labelled '**dependent' variable.** A regression is always a **linear** regression, i.e., a **straight line** represents the data as a **model**.

□ **Method of least squares**

In order to know which line to choose as the best model of a given data cloud, the method of least squares is used. We select the line for which the sum of all squared deviations (SS) of all data points is lowest. This line is labelled '**line of best fit**', or '**regression line**'.

□ **Simple regression**

The linear regression equation is:

$$Y_i = \left(b_0 + b_1 X_i\right) + e_i$$

Y_i= outcome we want to predict

b_0 = intercept of the regression line

b_1 = slope of the regression line coefficients

X_i = Score of subjects on the predictor variable

e_i = residual term, error

- **'Goodness-of-fit'**

The line of best fit (regression line) is compared with the most basic model. The former should be significantly better than the latter. The most basic model is the mean of the data.

- ***Mean of Y as basic model***

The summed squared differences between observed values and the mean, SS_T, are big; hence the mean is not a good model of the data.

- **Sum of squares total: SS_T**

Regression line as a model

- **Sum of squares residual SS_R**

The summed squared differences between observed values and the regression line, SS_R, are smaller; hence this regression line is a much better model of the data

- **Sum of squares model, SS_M**

SS_M: sum of squared differences between the mean of Y and the regression line (as our model)

- **Comparing the basic model and the regression model: R^2**

The improvement by the regression model can be expressed by dividing the sum of squares of the regression model SS_M by the sum of squares of the basic model SS_T:

$$R^2 = \frac{SS_M}{SS_T}$$

This is the same measure as the R^2 in correlation. Take the square root of R^2 and you have the Pearson correlation coefficient r!

R^2 ---The basic comparison in statistics is always to compare the amount of variance that our model can explain with the total amount of variation there is. If the model is good it can explain a significant proportion of this overall variance.

□ Comparing the basic model and the regression model: F-Test

In the F-Test, the ratio of the improvement due to the model SS_M and the difference between the model and the observed data, SS_R, is calculated.

We take the mean sum of squares, or mean squares, MS, for the model, MS_M, and the observed data, MS_R:

$$F = \frac{MS_M}{MS_R}$$

The F-ratio should be high (since the model should have improved the prediction considerably, as expressed in MS_M). MS_R, the difference between the model and the observed data (the residual), should be small.

□ The coefficient of a predictor

The coefficient of the predictor X is b_1. B_1 indicates the gradient/slope of the regression line. It says how much Y changes when X is changed one unit. In a good model, b_1 should always be different from 0, since the slope is either positive or negative.

Only a bad model, i.e., the basic model of the mean, has a slope of 0.

If $b_1=0$, this means:

- A change in one unit of the predictor X does not change the predicted variable Y.
- The gradient of the regression line is 0.

□ T-Test of the coefficient of the predictor

A good predictor variable should have a b_1 that is different from 0 (the regression coefficient of the basic model, the mean). This difference is significant, can be tested by a *t*-test.

The b of the expected values (0-Hypothesis, i.e., 0) is subtracted from the b of the observed values and divided by the standard error of b.

$$t = \frac{b_{observed} - b_{expected}}{SE_b}$$

Since $b_{expected} = 0$,

$$t = \frac{b_{observed}}{SE_b}$$

t should be * different from 0.

Summary:

This unit covers the concepts of Measures of Relationship. Measures of Relationship or Association refer to a wide variety of coefficients which measure strength of relationship, defined in various ways. The ideas of relationships between variables are measured using statistics known as correlation coefficient. The two variables are said to be related, if one variable changes in any manner whenever the other variables varies in a particular manner. The various types of correlation like Linear Correlation, Partial Correlation, Positive, Negative and Zero Correlation and Perfect Correlations and the commonly used correlation namely product moment correlation and rank difference correlation are discussed here. Finally, Regression is the statistical method underlying all inferential statistics which is also discussed.

This is an excellent reading for me. It is good to be reminded of these procedures.

The explanations are very clear to me.