UNIT 1 PRODUCT MOMENT COEFFICIENT OF CORRELATION

Structure

1.0 INTRODUCTION

We measure psychological attributes of people by using tests and scales in order to describe individuals. There are times when you realise that increment in one of the characteristics is associated with increment in other characteristic as well. For example, individuals who are more optimistic about the future are more likely to be happy. On the other hand, those who are less optimistic about future (i.e., pessimistic about it) are less likely to be happy. You would realise that as one variable is increasing, the other is also increasing and as the one is decreasing the other is also decreasing. In the statistical language it is referred to as correlation. It is a description of "relationship" or "association" between two variables (more than two variables can also be correlated, we will see it in multiple correlation).

In this unit you will be learning about direction of Correlation that is, Positive and Negative and zero correlation. You will also learn about the strength of correlation and how to measure correlation. Specifically you will be learning Pearson's Product Moment Coefficient of Correlation and how to interpret this correlation coefficient. You will also learn about the ramifications of the Pearson's r. You will also learn the coefficient of correlation equations with numerical examples.

1.1 OBJECTIVES

After reading and doing exercises in this unit, you will be able to:

- describe and explain concept of correlation;
- plot the scatter diagram;
- explain the concept of direction, and strength of relationship;
- differentiate between various measures of correlations;
- analyse conceptual issues in correlation and causality;
- describe problems suitable for correlation analysis;
- describe and explain concept of Pearson's Product Moment Correlation;
- compute and interpret Pearson's correlation by deviation score method and raw score method; and
- test the significance and apply the correlation to the real data.

1.2 CORRELATION: MEANING AND INTERPRETATION

Correlation is a measure of association between to variables. Typically, one variable is denoted as X and the other variable is denoted as Y. The relationship between these variables is assessed by correlation coefficient. Look at the earlier example of optimism and happiness. It states the relationship between one variable, optimism (X) and other variable, happiness (Y). Similarly, following statements are example of correlations:

As the *intelligence* (IQ) increases the *marks* obtained increases.

As the *introversion* increases *number of friends* decreases.

More the *anxiety* a person experiences, weaker the *adjustment* with the stress.

As the score on *openness to experience* increases, scores on *creativity* test also increase.

More the *income*, more the *expenditure*.

On a reasoning task, as the *accuracy* increases, the *speed* decreases.

As the *cost* increases the *sales* decrease.

Those who are good at *mathematics* are likely to be good at *science*.

As the age of the child increases, the *problems solving capacity* increase.

More the *practice*, better the *performance*.

All the above statements exemplify the correlation between two variables. The variables are shown in *italics*. In this first section, we shall introduce ourselves to the concept of correlation.

1.2.1 Scatter Diagram: Graphical Presentation of Relationship

Scatter diagram (also called as *scatterplot, scattergram,* or *scatter*) is one way to study the relationship between two variables. Scatter diagram is to plot pairs of values of subjects (observations) on a graph. Let's look at the following data of five subject, A to E (Table 1.1). Their scores on intelligence and scores on reasoning task are provided. The same data is used to plot a scatter diagram shown in Figure 1.1. Now, I shall quickly explain 'how to draw the scatter diagram'.

Table 1: Data of five subjects on intelligence and scores of reasoning

Subject	Intelligence	Scores on reasoning task
A	104	12
B	127	25
C	109	18
D	135	31
E	116	19

Step 1. Plotting the Axes

Draw the x and y axis on the graph and plot one variable on x-axis and another on y-axis.

(Although, correlation analyses do not restrict you from plotting any variable on any axis, plot the causal variable on x-axis in case of implicitly assumed cause-effect relationship.)

Also note that correlation does not necessarily imply causality.

Step 2. Range of Values

Decide the range of values depending on your data.

Begin from higher or lower value than zero.

Conventionally, the scatterplot is square.

So plot x and y values about the same length.

Step 3. Identify the pairs of values

Identify the pairs of values.

A pair of value is obtained from a data.

A pair of values is created by taking a one value on first variable and corresponding value on second variable.

Step 4. Plotting the graph

Now, locate these pairs in the graph.

Find an intersection point of x and y in the graph for each pair.

Mark it by a clear dot (or any symbol you like for example, star).

Then take second pair and so on.

The scatterplot shown below is based on the data given in table 1. (Refer to Figure 1).

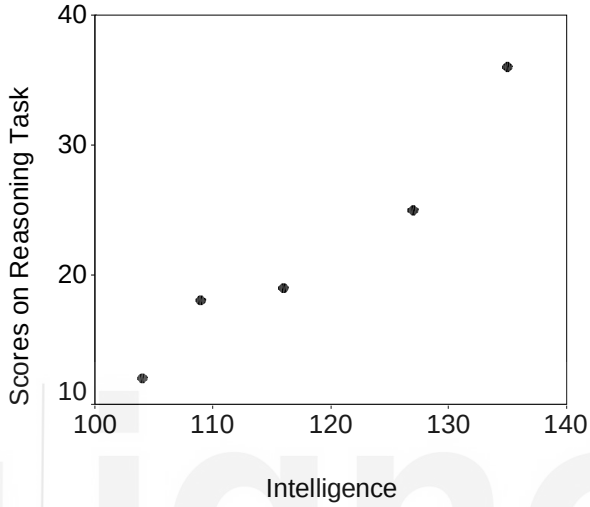


Fig. 1: Scatter diagram depicting relationship between intelligence and score on reasoning task

The graph shown above is scatterplot representing the relationship between intelligence and the scores on reasoning task. We have plotted intelligence on x-axis because it is a cause of the performance on the reasoning task. The scores on reasoning have started from 100 instead of zero simply because the smallest score on intelligence is 104 which is far away from zero. We have also started the range of reasoning scores from 10 since the lowest score on reasoning is 12. Then we have plotted the pair of scores. For example, subject A has score of 104 on intelligence and 12 on reasoning so we get x,y pair of 104,12. We have plotted this pair on the point of intersection between these two scores in the graph by a dot. This is the lowest dot at the left side of the graph. You can try to practice the scatter by using the data given in the practice.

1.2.2 Correlation: Linear and Non-Linear Relationship

The relationship between two variables can be of various types. Broadly, they can be classified as linear and nonlinear relationships. In this section we shall try to understand the linear and nonlinear relationships.

Linear Relationship

One of the basic forms of relationship is linear relationship. *Linear* relationship can be expressed as a relationship between two variables that can be plotted as a *straight* line. The linear relationship can be expressed in the following equation (eq. 1.1):

$$Y = á + â\ X \qquad \text{(eq. 1.1)}$$

In the equation 1.1,

- Y is a dependent variable (variable on y-axis),

- á (alpha) is a constant or Y intercept of straight line,
- â (beta) is slope of the line and
- X is independent variable (variable on x-axis).

We again plot scatter with the line that best fits for the data shown in table 1. So you can understand the linearity of the relationship. Figure 2 shows the scatter of the same data. In addition, it shows the line which is best fit line for the data. This line is plotted by using the method of least squares. We will learn more about it later (Unit 4). Figure 2 shows that there is a linear relationship between two variables, intelligence and Scores on Reasoning Task. The graph also shows the straight line relationship indicating linear relation.

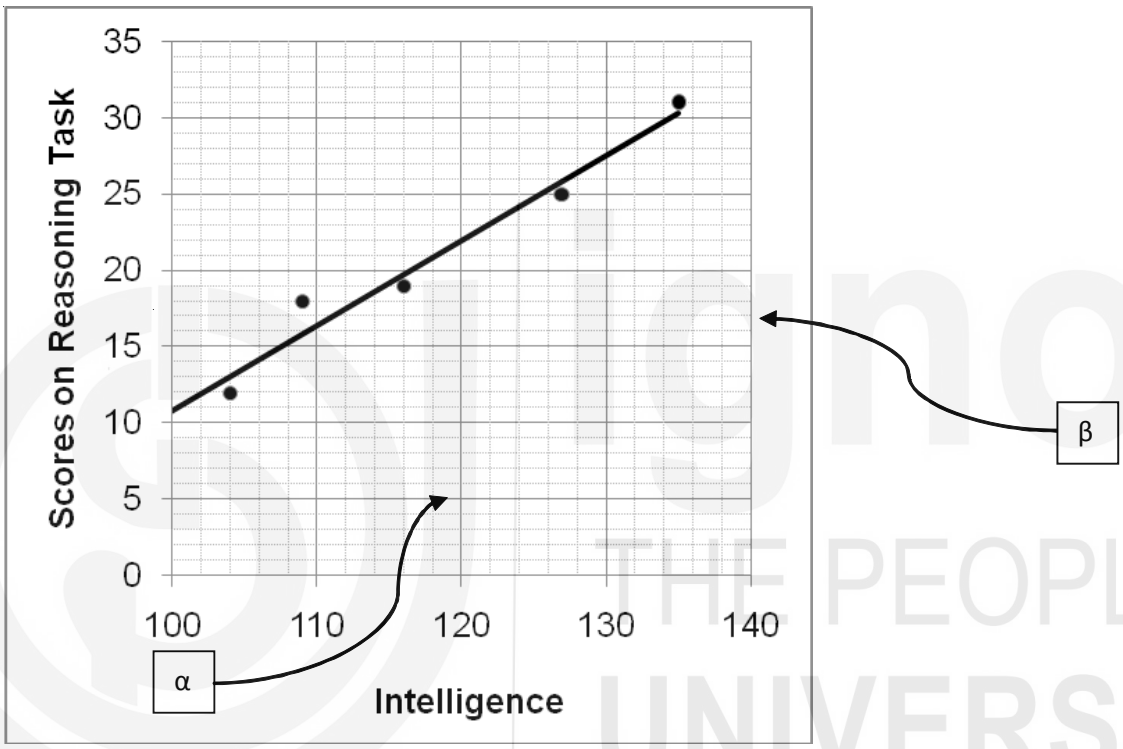


Fig. 2: Scatter showing linearity of the relationship between Intelligence and Scores on Reasoning Task

Non-linear Relationship

There are other forms of relationships as well. They are called as curvilinear or non-linear relationships. The Yorkes-Dodson Law, Steven's Power Law in Psychophysics, etc. are good examples of non-linear relationships. The relationship between stress and performance is popularly known as Yorkes-Dodson Law. It suggests that the performance is poor when the stress is too little or too much. It improves when the stress is moderate. Figure 3 shows this relationship. The *non-linear* relationships, cannot be plotted as a *straight line*.

The performance is poor at extremes and improves with moderate stress. This is one type of curvilinear relationship.

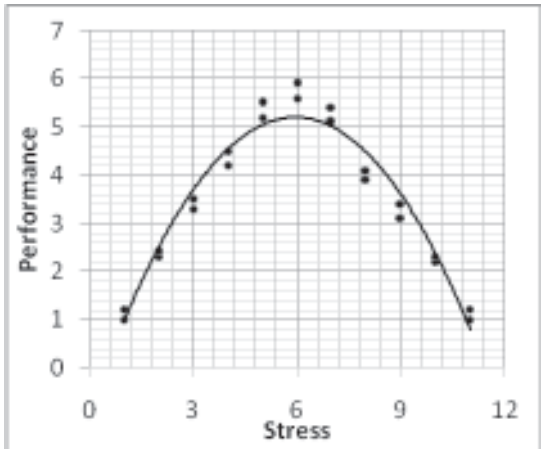


Fig. 3: Typical relationship between stress and performance

The curvilinear relationships are of various types (cubic, quadratic, polynomial, exponential, etc.). The point we need to note is that *relationships can be of various types*. This block discussed only *linear* relationships. Other forms of relationship are *not* discussed. The types of correlation presented in this block represent linear relationships. Pearson's product-moment correlation, Spearman's rho, etc. are linear correlations.

The Stevens' Power Law states that $r = cs^b$ where, r is sensation, s is stimulus, c and b are constants and coefficients, respectively. This is obviously a non-linear relationship between stimulus and sensation. Although, a reader who can recall some basic mathematics of 10^{th} grade can easily understand that by taking the log of both sides, the equation can be converted into linear equation.

1.2.3 Direction of Correlation: Positive and Negative

The direction of the relationship is an important aspect of the description of relationship. If the two variables are correlated then the relationship is either positive or negative. The absence of relationship indicates "zero correlation". Let's look at the positive, negative and zero correlation.

Positive Correlation

The positive correlation indicates that as the values of one variable increases the values of other variable also increase. Consequently, as the values of one variable decreases, the values of other variable also decrease. This means that both the variables move in the samer direction. For example,

a) As the *intelligence* (IQ) increases the *marks* obtained increases.

b) As *income* increases, the *expenditure* increases.

The figure 4 shows *scatterplot* of the positive relationship. You will realise that the higher scores on X axis are associated with higher score on Y axis and lower scores on X axis are generally associated with lower score on Y axis. In the 'a' example, higher scores on *intelligence* are associated with the higher score on *marks obtained*. Similarly, as the scores on *intelligence* drops down, the *marks obtained* has also dropped down.

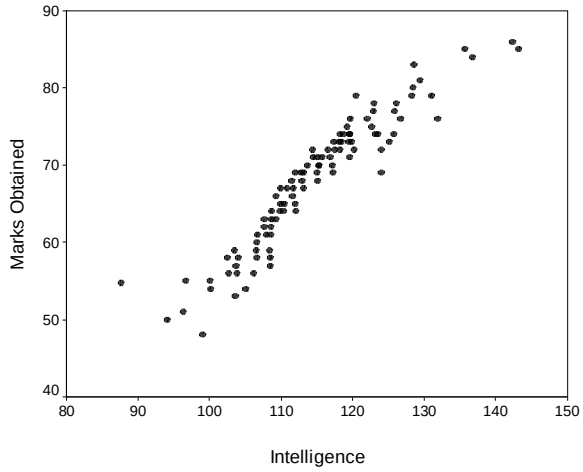


Fig. 4: Positive correlation: Scatter showing the positive correlation between *intelligence* and *marks obtained.*

Negative Correlation

The Negative correlation indicates that as the values of one variable increases, the values of the other variable decrease. Consequently, as the values of one variable decreases, the values of the other variable increase. This means that two variables move in the opposite direction. For example,

a) As the *intelligence* (IQ) increases the *errors on reasoning task* decreases.

b) As *hope* increases, *depression* decreases.

Figure 5 shows *scatterplot* of the negative relationship. You will realise that the higher scores on X axis are associated with lower scores on Y axis and lower scores on X axis are generally associated with higher score on Y axis.

In the 'a' example, higher scores on *intelligence* are associated with the lower score on *errors on reasoning task*. Similarly, as the scores on *intelligence* drops down, the *errors on reasoning task* have gone up.

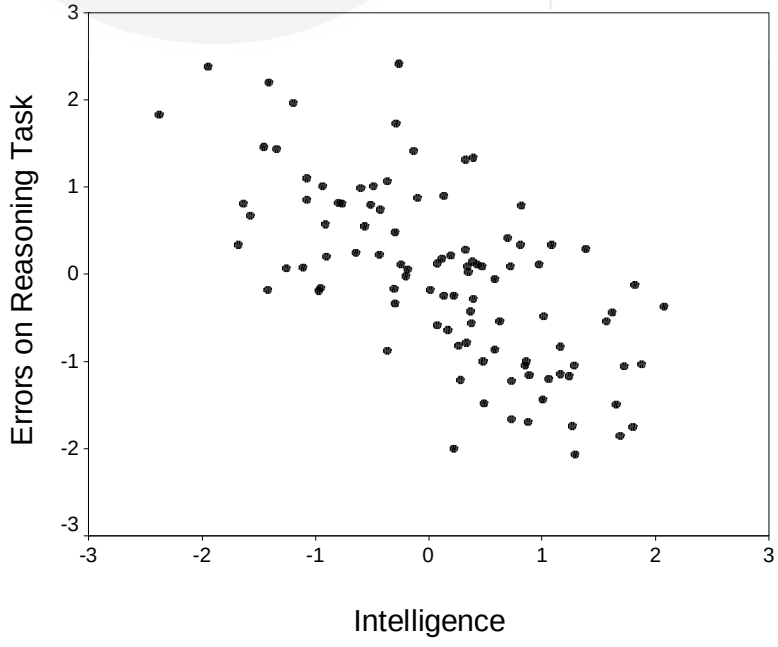


Fig. 5: Negative correlation: Scatter showing the negative correlation between *intelligence* and *errors on reasoning task*

No Relationship

Until now you have learned about the positive and negative correlations. Apart from positive and negative correlations, it is also possible that there is no relationship between *x* and *y*. That is the two variables do not share any relationship. If they do not share any relationship (that is, technically the correlation coefficient is zero), then, obviously, the direction of the correlation is neither positive nor negative. It is often called as zero correlation or no correlation.

(Please note that 'zero order correlation' is a different term than 'zero correlation' which we will discuss afterwards).

For example, guess the relationship between shoe size and intelligence?

This sounds an erratic question because there is no reason for any relationship between them. So there is no relationship between these two variables.

The data of one hundred individuals is plotted in Figure 6. It shows the scatterplot for no relationship.

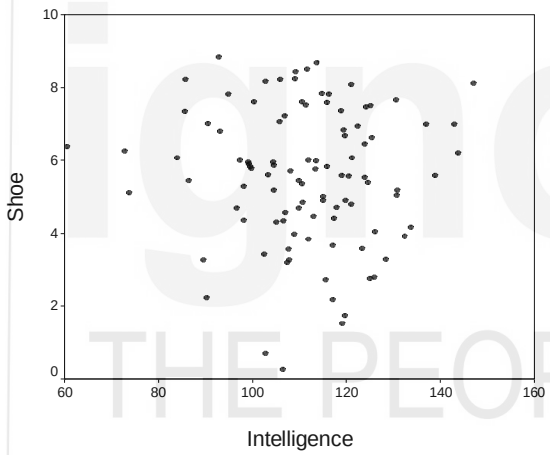


Fig. 6: Scatter between shoe size and intelligence of the individual

1.2.4 Correlation: The Strength of Relationship

You have so far learnt the direction of relationship between two variables. Any curious reader will ask a question "how strong is the relationship between the two variables?" For example, if you correlate intelligence with scores on reasoning, and creativity, what kind of relationship will you expect?

Obviously, the relationship between intelligence and reasoning as well as the relationship between intelligence and creativity are positive. At the same time the correlation coefficient (described in the following section) is higher for intelligence and reasoning than for intelligence and creativity, and therefore we realise that the relationship between intelligence and reasoning is stronger than relationship between intelligence and creativity. The strength of relationship between the two variables is an important information to interpret the relationship.

Correlation Coefficient

The correlation between any two variables is expressed in terms of a number, usually called as correlation coefficient. The correlation coefficient is denoted by various symbols depending on the type of correlation. The most common is '*r*' (small 'r') indicating the Pearson's product-moment correlation coefficient.

The representation of correlation between X and Y is r_{xy}.

The range of the correlation coefficient is from –1.00 to + 1.00.

It may take any value between these numbers including, for example, – 0.72, – 0.61, – 0.35, + 0.02, + 0.31, + 0.98, etc.

If the correlation coefficient is 1, then relationship between the two variables is perfect.

This will happen if the correlation coefficient is – 1 or + 1.

As the correlation coefficient moves nearer to + 1 or – 1, the strength of relationship between the two variables increases.

If the correlation coefficient moves away from the + 1 or – 1, then the strength of relationship between two variables decreases (that is, it becomes weak).

So correlation coefficient of + 0.87 (and similarly – 0.82, – 0.87, etc.) shows strong association between the two variables. Whereas, correlation coefficient of + 0.24 or – 0.24 will indicate weak relationship. Figure 7 indicates the range of correlation coefficient.

Fig. 7: The Range of Correlation Coefficient.

You can understand the strength of association as the common variance between two correlated variables. The correlation coefficient is NOT percentage.

So correlation of 0.30 does NOT mean it is 30% variance.

The shared variance between two correlated variables can be calculated. Let me explain this point. See, every variable has variance. We denote it as S_x^2 (variance of X). Similarly, Y also has its own variance (S_y^2). In the previous block you have learned to compute them. From the complete variance of X, it shares some variance with Y. It is called covariance.

The Figure 8 shown below explains the concept of shared variance. The circle X indicates the variance of X. Similarly, the circle Y indicates the variance of Y. The overlapping part of X and Y, indicated by shaded lines, shows the shared variance between X and Y. One can compute the shared variance.

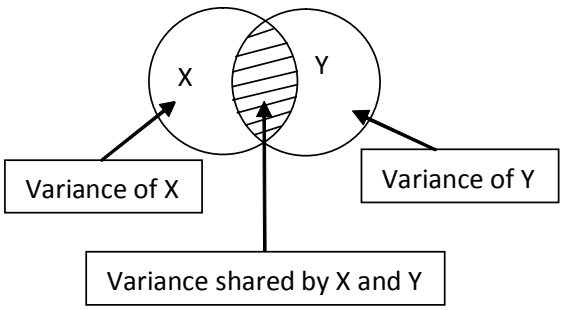


Fig. 8: Covariance indicates the degree to which X shares variance with Y

To calculate the percentage of shared variance between X and Y (common variance), one needs to square the correlation coefficient (*r*). The formula is given below:

Percentage of common variance between X and Y = $r_{xy}^2 \times 100$ (eq. 1.2)

For instance, if the correlation between X and Y is 0.50 then the percent of variation shared by X and Y can be calculated by using equation 1.2 as follows.

Percentage of common variance between X and Y = $r_{xy}^2 \times 100 = 0.50^2 \times 100 = 0.25 \times 100 = 25\%$

It indicates that, if the correlation between X and Y is 0.50 then 25% of the variance is shared by the two variables, X and Y. You would note that this formula is applicable to negative correlations as well. For instance, if $r_{xy} = -0.81$, then shared variance is:

Percentage of common variance between X and Y = $\times 100 = -0.81^2 \times 100 = 0.6561 \times 100 = 65.61\%$

1.2.5 Measurements of Correlation

Correlation coefficient can be calculated by various ways. The correlation coefficient is a description of association between two variables in the sample. So it is a descriptive statistics. Various ways to compute correlation simply indicate the degree of association between variables *in the sample*. The distributional assumptions are *not* required to compute correlation as a descriptive statistics. So it is not a parametric or nonparametric statistics.

The calculated sample correlation coefficient can be used to estimate population correlation coefficient.

The sample correlation coefficient is usually denoted by symbol '*r*'.

The population correlation coefficient is denoted by symbol '*ñ*'.

It is Greek letter *rho(ñ)*, pronounced as *row* (Spearman's correlation coefficient is also symbolised as *rho*.

This may create some confusion among the readers. Therefore, I shall use symbol r_s for Spearman's *rho* as a sample statistics and $\tilde{n}_s$ to indicate the population value of the Spearman's *rho*.

Henceforth, I shall also clearly mention the meaning with which *ñ* is used in this block.

- When the population correlation coefficient is estimated from sample correlation coefficient.
- then the correlation coefficient becomes an inferential statistic.
- Inference about population correlation (*ñ*) is drawn from sample statistics (*r*).
- The population correlation (*ñ*) is always unknown.
- What is known is sample correlation (*r*).
- The population indices are called as parameters and the sample indices are called as statistics.
- So *ñ* is a parameter and *r* is a statistics.

While inferring a parameter from sample, certain distributional assumptions are required. From this, you can understand that the descriptive use of the correlation coefficient does not require any distributional assumptions.

The most popular way to compute correlation is 'Pearson's Product Moment Correlation (*r*)'. This correlation coefficient can be computed when the data on both the variables is on at least equal interval scale or ratio scale.

Apart from Pearson's correlation there are various other ways to compute correlation. Spearman's Rank Order Correlation or Spearman's *rho* (r_s) is useful correlation coefficient when the data is in rank order.

Similarly, Kendall's *tau* (ô) is a useful correlation coefficient for rank-order data.

Biserial, Point Biserial, Tetrachoric, and Phi coefficient, are the correlations that are useful under special circumstances.

Apart from these, multiple correlations, part correlation and partial correlation are useful ways to understand the associations (Please note that the last three require more than two variables).

1.2.6 Correlation and Causality

The correlation does not necessarily imply causality. But, if the correlation between two variables is high then it might indicate the causality. If X and Y are correlated, then there are three different ways in which the relationship between two variables can be understood in terms of causality.

1) X is a cause of Y.

2) Y is cause of X.

3) Both, X and Y are caused by another variable Z.

However, causality can be inferred from the correlations.

Regression analysis, path analysis, structural equation modeling, are some examples where correlations are employed in order to understand causality.

1.3 PEARSON'S PRODUCT MOMENT COEFFICIENT OF CORRELATION

The Person's correlation coefficient was developed by Karl Pearson in 1886. Person was a editor of "Biometrica" which is a leading journal in statistics. Pearson was a close associate of psychologist Sir Francis Galton. The Pearson's correlation coefficient is usually calculated for two continuous variables. If either or both the variables are not continuous, then other statistical procedures are to be used. Some of them are equivalent to Pearson's correlation and others are not. We shall learn about these procedure after learning the Pearson's Correlation coefficient.

1.3.1 Variance and Covariance: Building Blocks of Correlations

Understanding product moment correlation coefficient requires understanding of mean, variance and covariance. We shall understand them once again in order to understand correlation.

Mean : Mean of variable X (symbolised as $\bar{X}$) is sum of scores ($\sum_{i=1}^{n} X_i$) divided by number of observations (n). The mean is calculated in following way.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(eq. 1.3)}$$

You have learned this in the first block. We will need to use this as a basic element to compute correlation.

Variance

The variance of a variable X (symbolised as S_X^2) is the sum of squares of the deviations of each X score from the mean of X ($\sum (X - \bar{X})^2$)divided by number of observations (*n*).

$$S_x^2 = \frac{\sum (X - \bar{X})^2}{n} \qquad \text{(eq. 1.4)}$$

You have already learned that standard deviation of variable X , symbolised as , is square root of variance of X, symbolised as .

Covariance

The covariance between X and Y (or S_{XY}) can be stated as

$$Cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} \qquad \text{(eq. 1.5)}$$

Covariance is a number that indicates the association between two variables. To compute covariance, deviation of each score on X from its mean ( ) and deviation of each score on Y from its mean ($\bar{y}$) is initially calculated.

Then products of these deviations are obtained.

Then, these products are summated.

This sum gives us the numerator for covariance.

Divide this sum by number of observations (*n*). The resulting number is covariance.

1.3.2 Equations for Pearson's Product Moment Coefficient of Correlation

Having revised the concepts, we shall now learn to compute the Pearson's Correlation Coefficient.

Formula

Since we have already learned to compute the covariance, the simplest way to define Pearson's correlation is…

$$r = \frac{Cov_{XY}}{S_X S_Y} \qquad \text{(eq. 1.6)}$$

Where,

the Cov_{XY} is covariance between X and Y,

S_X is standard deviation of X

S_Y is standard deviation of Y.

Since, it can be shown that Cov_{XY} is always smaller than or equal to $S_X S_Y$, the maximum value of correlation coefficient is bound to be 1.

The sign of Pearson's *r* depends on the sign of Cov_{XY}.

If the Cov_{XY} is negative, then *r* will be negative and

if Cov_{XY} is positive then *r* will be a positive value.

The denominator of this formula ($S_X S_Y$) is always positive. This is the reason for a – 1 to + 1 range of correlation coefficient. By substituting covariance equation (eq. 1.5) for covariance we can rewrite equation 1.6 as

$$r = \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{n}}{S_X S_Y} \qquad \text{(eq. 1.7)}$$

By following a simple rule, a÷ b ÷ c = a ÷ (b × c), we can rewrite equation 1.7 as follows.

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n S_X S_Y} \qquad \text{(eq. 1.8)}$$

1.3.3 Numerical Example

Now we shall use this formula to compute Pearson's correlation coefficient. For this purpose we will use the following data. The cognitive theory of depression argues that hopelessness is associated with depression. Aron Beck developed instruments to measure depression and hopelessness. The BHS (Beck Hopelessness Scale) and the BDI (Beck Depression Inventory) are measures of hopelessness and depression, respectively.

Let's take a hypothetical data of 10 individuals on whom these scales were administered. (In reality, such a small data is not sufficient to make sense of correlation; roughly, at least a data of 50 to 100 observations is required). We can hypothesize that the correlation between hopelessness and depression will be positive. This hypothetical data is given below in table 2.

Table 2: Hypothetical data of 10 subjects on BHS and BDI

Subject	BHS (X)	BDI (Y)	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
1	11	13	0	1	0	1	0
2	13	16	2	4	4	16	8
3	16	14	5	2	25	4	10
4	9	10	-2	-2	4	4	4
5	6	8	-5	-4	25	16	20
6	17	16	6	4	36	16	24
7	7	9	-4	-3	16	9	12
8	12	12	1	0	1	0	0
9	5	7	-6	-5	36	25	30
10	14	15	3	3	9	9	9
n = 10	$\sum X$ =110	$\sum Y$ =120			$\sum(X-\bar{X})^2$ = 156	$\sum(Y-\bar{Y})^2$ = 100	$\sum(X-\bar{X})(Y-\bar{Y})$ = 117
	$\bar{X}=11$	$\bar{Y}=12$					

$$S_X = \sqrt{\sum (X-\bar{X})^2 / n} = 4.16$$

$$S_Y = \sqrt{\sum (Y-\bar{Y})^2 / n} = 3.33$$

$$r = \sum (X-\bar{X})(Y-\bar{Y}) / nS_X S_Y = 117 / (10)(4.16)(3.33) = +0.937$$

Step 1. You need scores of subjects on two variables. We have scores on ten subjects on two variables, BHS and BDI.

Step 2. Then list the pairs of scores on two variables in two columns. The order will not make any difference. Remember, same individuals' two scores should be kept together. Label one variable as X and other as Y. We label BHS as X and BDI as Y.

Step 3. Compute the mean of variable X and variable Y. It was found to be 11 and 12 respectively.

Step 4. Compute the deviation of each X score from its mean ( ) and each Y score from its own mean ($\bar{Y}$). This is shown in the column labeled as $X-\bar{X}$ and $Y-\bar{Y}$. As you have learned earlier, the sum of these columns has to be zero.

Step 5. Compute the square of $x-\bar{x}$ and $y-\bar{y}$.

This is shown in next two columns labelled as $(x-\bar{x})^2$ and $(y-\bar{y})^2$.

Step 6. Then compute the sum of these squared deviations of X and Y. The sum of squared deviations for X is 156 and for Y it is 100.

Step 7. Divide them by *n* to obtain the standard deviations for X and Y. The S_x was found to be 4.16. Similarly, the S_y was found to be 3.33.

Step 8. Compute the cross-product of the deviations of X and Y. These cross-products are shown in the last column labeled as $(x - \bar{x})(y - \bar{y})$.

Step 9. Then obtain the sum of these cross-products. It was found to be 117. Now, we have all the elements required for computing *r*.

Step 10. Use the formula of *r* to compute correlation. The sum of the cross-product of deviations is numerator and n, S_x, S_y, are denominators. Compute *r*. the value of *r* is 0.937 in this example.

1.3.4 Significance Testing of Pearson's Correlation Coefficient

Statistical significance testing is testing the hypothesis about the population parameter from sample statistics. When the Pearson's Correlation coefficient is computed as an index of description of relationship between two variables in the sample, the significance testing is not required. The interpretation of correlation from the value and direction is enough.

However, when correlation is computed as an estimate of population correlation, obviously, statistical significance testing is required.

"Whether the obtained sample value of Pearson's correlation coefficient is greater than the value that can be obtained by chance?" is the question answered by statistical significance testing about correlation coefficient.

Different values of correlation can be obtained between any two variables, X and Y, for different samples of different sizes belonging to the same population.

The researcher is not merely interested in knowing the finding in the specific sample on which the data are obtained. But they are interested in estimating the population value of the correlation.

Testing the significance of correlation coefficient is a complex issue. It is because of the distribution of the correlation coefficient. The *t*-distribution and z-distribution are used to test statistical significance of *r*.

The population correlation between X and Y is denoted by $\tilde{n}_{xy}$. The sample correlation is r_{xy}.

As you have learned, we need to write a null hypothesis (H_O) and alternative hypothesis (H_A) for this purpose.

The typical null hypothesis states that population correlation coefficient between X and Y ($\tilde{n}_{xy}$) is zero.

$H_O : \tilde{n}_{xy} = 0$

$H_A: \tilde{n}_{xy} \neq 0$

If we reject the H_O then we accept the alternative (H_A) that the population correlation coefficient is other than zero. It implies that the finding obtained on the data is not a sample-specific error.

Sir Ronald Fisher has developed a method of using *t*-distribution for testing this null hypothesis.

The degrees of freedom (*df*) for this purpose are $n - 2$. Here *n* refers to number of observations.

We can use Appendix C in a statistic book for testing the significance of correlation coefficient. Appendix C provides critical values of correlation coefficients for various degrees of freedom. Let's learn how to use the Appendix C. We shall continue with the example of BHS and BDI.

The correlation between BHS and BDI is +.937 obtained on 10 individuals. We decide to do statistical significance testing at 0.05 level of significance, so our á = .05.

We also decided to apply two-tailed test.

The two-tailed test is used if alternative hypothesis is non-directional, i.e. it does not indicate the direction of correlation coefficient (meaning, it can be positive or negative) and one-tail test is used when alternative is directional (it states that correlation is either positive or negative).

Let us write the null hypothesis and alternative hypothesis:

Null hypothesis

$$H_O : \rho_{\text{BHS BDI}} = 0$$
$$H_A : \rho_{\text{BHS BDI}} \neq 0$$

Now we will calculate the degree of freedom for this example.

$df = n - 2 = 10 - 2 = 8$ (eq. 1.9)

So the *df* for this example are 8. Now look at Appendix C. Look down the leftmost *df* column till you reach *df* = 8. Then look across to find correlation coefficient from column of two-tailed test at level of significance of 0.05. You will reach the critical value of *r*:

$r_{\text{critical}} = 0.632$

Because the obtained (i.e., calculated) correlation value of + 0.937 is greater than critical (i.e., tabled) value, we reject the null hypothesis that there is no correlation between BHS and BDI in the population.

So we accept that there is correlation between BHS and BDI in the population. This method is used regardless of the sign of the correlation coefficient.

We use the absolute value (ignore the sign) of correlation while doing a two-tailed test of significance. The sign is considered while testing one-tailed hypothesis.

For example, if the H_A: $\tilde{n} > 0$, which is a directional hypothesis, then any correlation that is negative will be considered as insignificant.

1.3.5 Adjusted *r*

The Pearson's correlation coefficient (*r*) calculated on the sample is not an unbiased estimate of population coefficient ($\tilde{n}$). When the number of observations (sample size) are small the sample correlation is a biased estimate of population correlation. In order to reduce this bias, the calculated correlation coefficient is adjusted. This is called as adjusted correlation coefficient (r_{adj}).

$$r_{\text{adj}} = \sqrt{1 - \frac{(1 - r^2)(n - 1)}{n - 2}}$$

Where,

r_{adj}= adjusted *r*

r^{2-} = the square of Pearson's Correlation Coefficient obtained on sample,

n = sample size

In case of our data, presented in table 1.2, the correlation between BHS and BDI is +.937 obtained on the sample of 10. The adjusted r can be calculated as follows

$$r_{adj} = \sqrt{1 - \frac{(1 - .937^2)(10 - 1)}{10 - 2}} = \sqrt{1 - \frac{(.1220)(9)}{8}} = \sqrt{1 - 0.1373} = .929$$

The r_{adj} is found to be 0.929. This coefficient is unbiased estimate of population correlation coefficient.

1.3.6 Assumptions for Significance Testing

One may recall that simple descriptive use of correlation coefficient does not involve any assumption about the distribution of either of the variables. However, using correlation as an inferential statistics requires assumptions about X and Y. These assumptions are as follows. Since we are using *t*-distribution, the assumptions would be similar to *t*.

Assumptions:

Independence among the pairs of score

This assumption implies that the scores of any two observations (subjects in case of most of psychological data) are not influenced by each other. Each pair of observation is independent. This is assured when different subjects provides different pairs of observation.

The population of X and the population of Y follow normal distribution and the population pair of scores of X and Y has a normal bivariate distribution.

This assumption states that the population distribution of both the variables (X and Y) is normal. This also means that the pair of scores follows bivariate normal distribution. This assumption can be tested by using statistical tests for normality.

It should be remembered that the *r* is a robust statistics. It implies that some violation of assumption would not influence the distributional properties of *t* and the probability judgments associated with the population correlation.

1.3.7 Ramifications in the Interpretation of Pearson's r

The interpretation of the correlation coefficient depends primarily on two things: direction and strength of relationship. We have already discussed them in detail and hence repetition is avoided.

Direction

If the correlation is positive, then the relationship between two variables is positive. It means that as there is an increase in one there is an increase in another, and as there is a decrement in one there is a decrease in another. When the direction of correlation is negative then the interpretation is vice-versa.

Strength

The strength can be calculated in terms of percentage. We have already learned this formula. So we can convert the correlation coefficient into percentage of common

variance explained and accordingly interpret. For example, if the correlation between X and Y is 0.78, then the common variance shared by X and Y is 60.84 percent.

Usually distinct psychological variables do not share much of the common variance. In fact, the reliability of psychological variables is an issue while interpreting the correlations. Cohen and Cohen have suggested that considering the unreliability of psychological variables, the smaller correlations should also be considered significant.

Although, direction and strength are key pointers while interpreting the correlation, there are finer aspects of interpretation to correlations.

They are :

- range,
- outliers,
- reliability of variables, and
- linearity

The above are some of the important aspects which all obscure the interpretation of correlation coefficient. Let's discuss them one by one.

1.3.8 Restricted Range

It is expected the variables in correlation analysis are measured with full range. For example, suppose we want to study the correlation between hours spent in studies and marks. We are suppose to take students who have varying degree of hours of studies, that is, we need to select students who have spent very little time in studies to the once who have spent great deal of time in studies. Then we will be able to obtain true value of the correlation coefficient.

But suppose we take a very restricted range then the value of the correlation is likely to reduce. Look at the following examples the figure 1.9a and 1.9b.

The figure 1.9a is based on a complete range.

The figure 1.9b is based on the data of students who have studied for longer durations.

The scatter shows that when the range was full, the correlation coefficient was showing positive and high correlation. When the range was restricted, the correlation has reduced drastically.

You can think of some such examples. Suppose, a sports teacher selects 10 students from a group of 100 students on basis of selection criterion, that is their athletic performance.

The actual performance of these ten selected students in the game was correlated with the selection criterion. A very low correlation was obtained between selection criterion and actual game performance. This would naturally mean that the selection criterion is not related with actual game performance. Is it true..? Why so…?

If you look at the edata, you will realise that the range of the scores on selection criterion is extremely restricted (because these ten students were only high scorers) and hence the relationship is weak. So note that whenever you interpret correlations, the range of the variables is large. Otherwise the interpretations will not be valid.

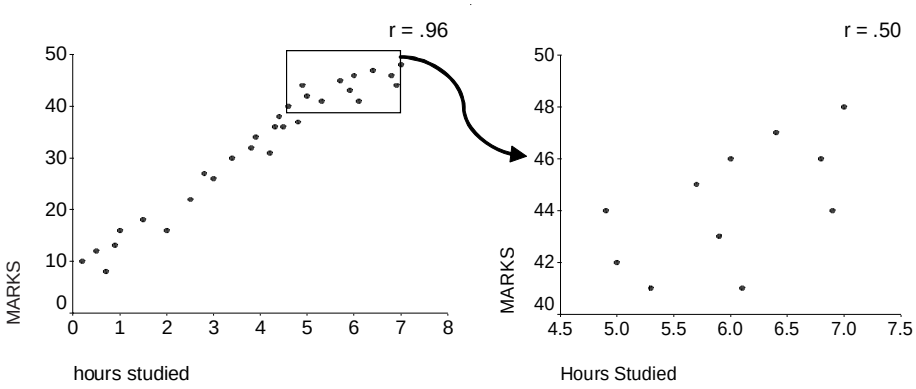


Fig. 1.9a: Scatter showing full range on both variables

Fig. 1.9b: Scatter with restricted range on hours studied

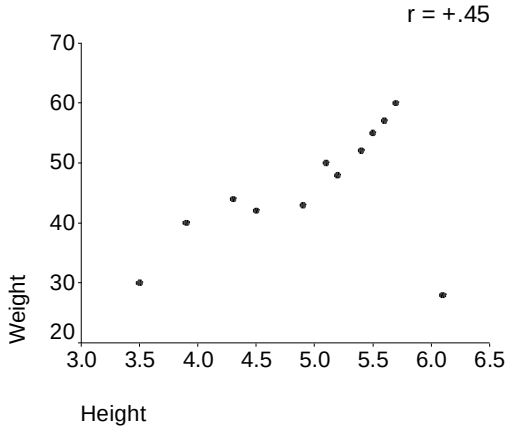
1.4 UNRELIABILITY OF MEASUREMENT

Psychological research involves the use of scales and tests. One of the psychometric property psychological instruments should poses is reliability. Reliability refers to consistency of a measurement. If the instrument is consistent, then the test has high reliability. But at times one of the variable or both the variables may have lower reliability. In this case, the correlation between two less reliable variable reduces. Generally, while interpreting the correlation, the reliability is assumed to be high. The general interpretations of correlations are not valid if the reliability is low. This reduction in the correlation can be adjusted for the reliability of the psychological test. More advanced procedures are available in the books of psychological testing and statistics. They involve calculating disattenuated correlations. Correlation between two variables that have less than perfect reliability is adjusted for unreliability. This is called as disattenuated correlation. If both variables were perfectly reliable then correlation between them is disattenuated correlation.

1.4.1 Outliers

Outliers are extreme score on one of the variables or both the variables. The presence of outliers has deterring impact on the correlation value. The strength and degree of the correlation are affected by the presence of outlier. Suppose you want to compute correlation between height and weight. They are known to correlate positively. Look at the figure below. One of the scores has low score on weight and high score on height (probably, some anorexia patient).

Figure 1.10. Impact of an outlier observation on correlation. Without the outlier, the correlation is 0.95. The presence of an outlier has drastically reduced a correlation coefficient to 0.45.



1.4.2 Curvilinearity

We have already discussed the issue of linearity of the relationship. The Pearson's product moment correlation is appropriate if the relationship between two variables is linear. The relationships are curvilinear then other techniques need to be used. If the degree of cruviliniarity is not very high, high score on both the variable go together, low scores go together, but the pattern is not linear then the useful option is Spearman's *rho*.

1.5 USING RAW SCORE METHOD FOR CALCULATING *r*

The method which we have learned to compute the correlation coefficient is called as deviation scores formula. Now we shall learn another method to calculate Pearson's correlation coefficient. It is called as raw score method. First we will understand how the two formulas are similar. Then we will solve a numerical example for the raw score method. We have learned following formula for calculating *r*.

1.5.1 Formulas for Raw Score

We have already learnt following formula of correlation (eq. 1.8). This is a deviation score formula.

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{nS_X S_Y}$$

The denominator of correlation formula can be written as

$$\sqrt{\sum (X - \bar{X})^2 (Y - \bar{Y})^2} \qquad \text{(eq. 1.10)}$$

Which is

$$\sqrt{(SS_X SS_Y)} \qquad \text{(eq. 1.11)}$$

We have already learnt that

$$SS_X = \sum (X - \bar{X})^2 = \sum X^2 - \frac{(\sum X)^2}{n} \qquad \text{(eq. 1.12)}$$

and

$$SS_Y = \sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} \qquad \text{(eq. 1.13)}$$

The numerator of the correlation formula can be written as

$$\sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n} \qquad \text{(eq. 1.14)}$$

So *r* can be calculated by following formula which is a raw score formula:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(SS_X SS_Y)}} \qquad \text{(eq. 1.15)}$$

1.5.2 Solved Numerical for Raw Score Formula

We shall solve the same numerical example by using the formulas shown above. Table 3 shows how to calculate Pearson's *r* by using raw score formula.

Table 1.3: Table showing the calculation of r by using raw score formula.

Subject	BHS (X)	BDI (Y)	X^2	Y^2	XY
1	11	13	100	676	260
2	13	16	64	529	184
3	16	14	81	529	207
4	9	10	169	676	338
5	6	8	121	576	264
6	17	16	196	900	420
7	7	9	256	729	432
8	12	12	144	729	324
9	5	7	225	841	435
10	14	15	144	625	300
Summation	110	120	1366	1540	1437
	$\bar{X} = 11$	$\bar{Y} = 12$			

$$SS_X = \sum X^2 - (\sum X)^2 / n = 1366 - (110)^2/10 = 156$$

$$SS_Y = \sum Y^2 - (\sum Y)^2 / n = 1540 - (120)^2 / 10 = 100$$

$$\sum (X - \bar{X})(Y - \bar{Y}) = \sum XY - \frac{(\sum X)(\sum Y)}{n} \qquad = 117$$

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(SS_X SS_Y)}} \qquad = 0.937$$

Readers might find one of the methods easier. There is nothing special about the methods. One should be able to correctly compute the value of correlation.

1.6 LET US SUM UP

In this unit we started with definition and meaning of correlation and followed it up with how the correlation could be depicted in graphical form and scatter diagram. We then learnt about linear and non-linear and curvilinear relationship amongst variables. We also learnt that the direction of relationship could be either in the positive or in the negative direction. It can also be no correlation in that it can be a zero correlation. Then we learnt the methods of measurement of correlation and

learnt how to use the formula for calculating the Pearson's r, that Pearson's Product Moment Coefficient of Correlation. We then discussed about the building blocks of correlation which included variance and co variance. We also learnt how to test the level of significance of a particular coefficient of correlation calculated by us. Then we learnt about the interpretation of the correlation coefficient and also its ramifications. Then we looked into the unreliability of a correlation and the causes for the same such as the inclusion of outliers etc.

1.7 UNIT END QUESTIONS

1) Problem:

Plot scatter diagram for the following data. Compute Pearson's correlation between x and y. Write the null hypothesis stating that population correlation is zero. Test the significance of the correlation coefficient.

X	Y
12	20
13	22
15	28
17	31
11	22
9	24
8	18
10	21
11	23
7	16

2) Plot scatter for following example. The data was collected on Perceived stress and anxiety on 10 subjects. Compute the Pearson's correlation between them State the null hypothesis. Test the null hypothesis using this hypothesis. Do the similar exercise after deleting a pair that clearly looks an outlier observation.

Perceived stress	Anxiety
9	12
8	11
7	9
4	5
8	9
4	6
6	8
14	2
7	11
11	9
9	11

3) Data showing scores on time taken to complete 200 meters race and duration of practice for 5 runners. Plot the scatter. Compute mean, variance, SD, and covariance. Compute correlation coefficient. Write the null hypothesis.

Time taken (in Seconds)	Duration of Practice (in months)
31	11
32	14
36	9
26	15
38	7

4) Data showing scores on dissatisfaction with work and scores on irritability measured by standardised test for thirteen individuals. Plot the scatter. Compute mean, variance, SD, and covariance. Compute correlation coefficient. Write the null hypothesis stating no relationship. Test the significance at 0.05 level of significance.

Dissatisfaction with work	Irritability scores
12	5
16	7
19	9
27	13
30	16
25	11
22	6
26	14
11	7
17	9
19	14
21	18
23	19

5) Check whether the following statements are true or false.

1)	Positive correlation means as X increases Y decreases.	True/False
2)	Negative correlation means as X decreases Y decreases.	True/False
3)	Generally, in a scatter, lower scores on X are paired with lower scores on Y for negative correlation.	True/False
4)	$-1.00 \leq$ Pearson's correlation $\leq +1.00$	True/False
5)	Generally, in a scatter, lower scores on X are paired with higher scores on Y in positive correlation.	True/False
6)	The scatter diagram cannot indicate the direction of the relationship.	True/False
7)	Percentage of shared variance by X and Y can be obtained by squaring the value of correlation.	True/False

Answers: 1) = False, 2) = False, 3) = False, 4) = True, 5) = False, 6) = False, 7) = True

Answer in brief.

6) What is correlation coefficient?

7) What is the range of correlation coefficient?

8) Is correlation coefficient a percentage?

9) How to calculate common variance from correlation coefficient?

10) What is the percentage of variance shared by X and Y if the $r_{xy} = 0.77$?

11) What is the percentage of variance shared by X and Y if the $r_{xy} = -0.56$?

Answers:

A number expressing the relationship between two variables.

The range of correlation coefficient is from –1.00 to + 1.00.

No. Correlation is not a percentage. But it can be converted into percentage of variance shared.

Common variance is calculated from correlation coefficient by using a formula: $r_{xy}^2 \times 100$.

59.29%

31.36%

1.8 SUGGESTED READINGS

Aron, A., Aron, E. N., Coups, E.J. (2007). *Statistics for Psychology*. Delhi: Pearson Education.

Minium, E. W., King, B. M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education*. Singapore: John-Wiley.

Guilford, J. P., & Fructore, B. (1978). *Fundamental Statistics for Psychology and Education*. N.Y.: McGraw-Hill.

Wilcoxon, R. R. (1996). *Statistics for Social Sciences*. San Diego: Academic Press.

UNIT 2 OTHER TYPES OF CORRELATION (PHI-COEFFICIENT)

Structure

2.0 INTRODUCTION

We have learned about the correlation as a concept and also learned about the Pearson's coefficient of correlation. We understand that Pearson's correlation is based on certain assumptions, and if those assumptions are not followed or the data is not appropriate for the Pearson's correlation, then what has to be done ? This unit is answering this practical problem. When either the data type or the assumptions are not followed then the correlation techniques listed in this unit are useful. Out of them some are actually Pearson's correlations with different name and some are non-Pearson correlations. The rank data also poses some issues and hence this unit is also providing the answers to this problem. In this unit we shall learn about Special

Types of Pearson Correlation, Special Correlation of Non-Pearson Type, and correlations for rank-order data. The special types of Pearson correlation are Point-Biserial Correlation and Phi coefficient. The non-Pearson correlations are Biserial and Tetrachoric. The rank order correlations discussed are Spearman's *rho* and Kendall's *tau*.

2.1 OBJECTIVES

After completing this unit, you will be able to:

- describe and explain concept of special correlation;
- explain the concept of special correlation and describe and differentiate between their types;
- describe and explain concept of Point-Biserial and Phi coefficient;
- describe and explain concept of Biserial and Tetrachorich coefficient;
- compute and interpret Special correlations;
- test the significance and apply the correlation to the real data;
- explain concept of Spearman's rho and tau coefficient;
- compute and interpret rho and tau; and
- apply the correlation techniques to the real data.

2.2 SPECIAL TYPES OF CORRELATION

The correlation we have learned in the last unit is Pearson's product moment coefficient of correlation. The Pearson's *r* is one of the computational processes for calculating the correlation between two variables. Nevertheless, this is not the only way to calculate correlations. It is just one of the ways of calculating correlation coefficient.

This correlation can be calculated under various restrictions. The variables X and Y were assumed to be continuous variables. The distribution of these variables is expected to be normal. Some homogeneity among the variables is also expected. Linearity of the relationship is also required for computing the Pearson's *r*. There might be instances when one or more of these conditions are not met. In such cases, one needs to use alternative methods of correlations. Some of them are Pearson's correlation modified for specific kind of data. Others are non-Pearson correlations.

Let us take a quick note on distinction between measures of correlation and measures of association. Howell (2002) made this point quite clear. Measures of correlations are those where some sort of order can be assigned to each of the variable. Increment in scores either represent higher levels (or lower levels) of some quantified attribute. For example, number of friends, BHS hope scores, time taken to complete a task, etc. Measures of association are those statistical procedures that are utilised for variables that do not have a property of order. They are categorical variables, or nominal variables, for example association of gender (male and female) with ownership of residence (own and do not own). Both these variables are nominal variables and do not involve any order.

In this unit we shall learn about Special Types of Pearson Correlation, Special Correlation of Non-Pearson Type, and correlations for rank-order data. The special

types of Pearson correlation are Point-Biserial Correlation and Phi coefficient. The non-Pearson correlations are Biserial and Tetrachoric. The rank order correlations discussed are Spearman's *rho* and Kendall's *tau*.

2.3 POINT BISERIAL CORRELATION (r_{PB})

Some variables are dichotomous. The dichotomous variable is the one that can be divided into two sharply distinguished or mutually exclusive categories. Some examples are, male-female, rural-urban, Indian-American, diagnosed with illness and not diagnosed with illness, Experimental group and Control Group, etc. These are the truly dichotomous variables for which no underlying continuous distribution can be assumed. Now if we want to correlate these variables, then applying Pearson's formula have problems because of lack of continuity. Pearson's correlation requires continuous variables.

Suppose we are correlating gender, then male will be given a score of 0, and females will be given a score of 1 (or vice versa; indeed you can give a score of 5 to male and score of 11 to female and it won't make any difference for the correlation calculated).

Point Biserial Correlation (r_{pb}) is Pearson's Product moment correlation between one truly dichotomous variable and other continuous variable. Algebraically, the $r_{pb} = r$. So we can calculate r_{pb} in a similar way.

2.3.1 Calculation of r_{pb}

Let's look at the following data. It is a data of 20 subjects, out of which 9 are male and 11 are females. Their marks in the final examination are also provided. We want to correlate marks in the final examination with sex of the subject. The marks obtained in the final examination are a continuous variable whereas sex is truly dichotomous variable, taking two values male or female. We are using value of 0 for male subject and value of 1 for female subjects. The correlation appropriate for this purpose is Point-Biserial correlation (r_{pb}).

Table 1: Data showing the gender and mark for 20 subjects

Subject	Sex (male) X	Marks (Y)	Subject	Sex (Female (X)	Marks (Y)
1	0	46	11	1	58
2	0	74	12	1	69
3	0	58	13	1	76
4	0	67	14	1	78
5	0	62	15	1	65
6	0	71	16	1	69
7	0	54	17	1	59
8	0	63	18	1	53
9	0	53	19	1	73
10	1	67	20	1	81

$\text{Mean}_{sex} = 0.55$ $\text{Mean}_{marks} = 64.8$ $\text{Mean Marks}_{male} = 60.88$

$S_{sex} = 0.497$ $S_{marks} = 9.17$ $\text{Mean Marks}_{female} = 68$

$Cov_{XY} = 1.76$

$$r = \frac{Cov_{XY}}{S_X S_Y} = \frac{1.76}{0.497 \times 9.17} = 0.386$$

The Pearson's correlation (point biserial correlation) between sex and marks obtained is 0.386. The sign is positive. The sign is arbitrary and need to be interpreted depending on the coding of the dichotomous group. The interpretation of the sign is the group that is coded as 1 has a higher mean than the group that is coded as 0. The strength of correlation coefficient is calculated in a similar way. The correlation is 0.386, so the percentage of variance shared by both the variables is r^2 for Pearson's correlation. Same would hold true for point biserial correlation. The r_{pb}^2 is $0.386^2 = 0.149$. This means that 15% of information in marks is shared by sex.

2.3.2 Significance Testing of r_{pb}

The null hypothesis and alternative hypothesis for this purpose are as follows:

H_o: $\tilde{n} = 0$

H_A: $\tilde{n} \neq 0$

Since the r_{pb} is Pearson's correlation, the significance testing is also similar to it. the *t*- distribution is used for this purpose with $n - 2$ as *df*.

$$t = \frac{r_{pb}\sqrt{n-2}}{\sqrt{1-r_{pb}^2}} \qquad \text{(eq. 2.1)}$$

The t value for our data is 1.775. The $df = n - 2 = 20 - 2 = 18$. The value is not significant at 0.05 level. Hence we retain the null hypothesis.

2.4 PHI COEFFICIENT (ϕ)

The Pearson's correlation between one dichotomous variable and another continuous variable is called as point-biserial correlation. When both the variables are dichotomous, then the Pearson's correlation calculated is called as Phi Coefficient (ϕ).

For example, let us say that you have to compute correlation between gender and ownership of the property. The gender takes two levels, male and female. The ownership of property can be measured as either the person owns a property and the person do not own property. Now you have both the variables measured as dichotomous variables. Now if you compute the Pearson's correlation between these two variables is called as Phi Coefficient (ϕ). Both the variables take value of either of 0 or one. Look at the data given in the table below.

Table 2: Data and calculation for correlation between gender and ownership of property

X: Gender	0= Male 1 = Female											
Y: Ownership of Property	0=No ownership 1 = Ownership											
X	1	0	1	1	0	0	0	0	1	1	1	0
Y	0	1	0	1	1	1	0	1	0	0	1	1

Calculations

$\overline{X} = 0.5$ $S_x = 0.52$

$\overline{Y} = .58$ $S_y = 0.51$ Cov $_{XY}$ = – 0.13

$$r_{XY} = \phi_{XY} = \frac{Cov_{XY}}{S_X S_Y} = \frac{-0.13}{0.52 \times 0.51} = -.465$$

The value of ϕ coefficient is found to be – .465.

The relationship is negative, is function of the way we have assigned the number 0 and 1 to each of the variable. If we assign 0 to females and 1 to males, then we will get the same value of correlation with positive sign. Nevertheless, this does not mean that sign of the relationship cannot be interpreted. Once these numbers have been assigned, then we can interpret the sign. Male is 0 and female is 1; whereas 0 = no ownership and 1 is ownership.

The negative relation can be interpreted as follows: as we move from male to female we move negatively from no ownership to ownership. Meaning that male have more ownership than females. We can also calculate the proportion of variance shared by these two variables.

That is $r^2 = \phi^2 = -.465^2 = 0.216$ percent.

2.4.1 Significance Testing of Phi (ϕ)

The significance can be tested by using the Chi-Square ($\div^2$) distribution.

The ϕ can be converted into the $\div^2$ by obtaining a product of n and ϕ^2.

The Chi-Square of $n\phi^2$ will have $df = 1$.

The null and alternative hypothesis are as follows:

H_O: $\tilde{n} = 0$

H_A: $\tilde{n}$ ‘‘‘ 0

$\div^2 = n\phi^2 = 12 \times .216 = 2.59$ (eq. 2.2)

The value of the chi-square at 1 *df* is 3.84. the obtained value is less than the tabled value. So we accept the null hypothesis which states that the population correlation is zero.

One need to know that this is primarily because of the small sample size. If we take a larger sample, then the values would be significant. Quickly note the relationship between $\div^2$ and ϕ.

$$\phi = \sqrt{\frac{\chi^2}{n}} \qquad \text{(eq. 2.3)}$$

So one can compute the chi-square and then calculate the phi coefficient.

2.5 BISERIAL CORRELATION

The biserial correlation coefficient (r_b), is a measure of correlation. It is like the point-biserial correlation. But point-biserial correlation is computed while one of the variables is dichotomous and do not have any underlying continuity. If a variable has underlying continuity but measured dichotomously, then the biserial correlation can be calculated.

An example might be mood (happy-sad) and hope, which we have discussed in the first unit. Suppose we measure hope with BHS and measure mood by classifying those who have clinically low *vs*. normal mood. Actually, it is fair to assume that mood is a normally distributed variable.

But this variable is measured discretely and takes only two values, low mood (0) and normal mood (1).

Let's call continuous variable as Y and dichotomized variable as X. the values taken by X are 0 and 1.

So biserial correlation is a correlation coefficient between two continuous variables (X and Y), out of which one is measured dichotomously (X). The formula is very similar to the point-biserial but yet different:

$$r_b = \left[\frac{\overline{Y}_1 - \overline{Y}_0}{S_Y}\right]\left[\frac{P_0 P_1}{h}\right] \qquad \text{(eq. 2.4)}$$

where $\overline{Y}_0$ and $\overline{Y}_1$ are the Y score means for data pairs with an X score of 0 and 1, respectively, P_0 and P_1 are the proportions of data pairs with X scores of 0 and 1, respectively, and S_Y is the standard deviation for the Y data, and *h* is ordinate or the height of the standard normal distribution at the point which divides the proportions of P_0 and P_1.

The relationship between the point-biserial and the biserial correlation is as follows.

$$r_b = \frac{r_{pb}\sqrt{p_0 p_1}}{h} \qquad \text{(e.q 2.5)}$$

So once you compute the r_{pb}, its easy to compute the r_b.

2.6 TETRACHORIC CORRELATION (r_{TET})

Tetrachoric correlation is a correlation between two dichotomous variables that have underlying continuous distribution. If the two variables are measured in a more refined way, then the continuous distribution will result. For example, attitude to females and attitude towards liberalisation are two variables to be correlated. Now, we simply measure them as having positive or negative attitude. So we have 0 (negative attitude) and 1 (positive attitude) scores available on both the variables. Then the correlation between these two variables can be computed using Tetrachoric correlation (r_{tet}).

The correlation can be expressed as

$$r = \cos\theta \qquad \text{(eq. 2.6)}$$

Where, è is angle between the vector X and Y. Using this logic, r_{tet} can also be calculated.

$$r_{tet} = \cos\left[\frac{180^0}{1+\sqrt{\frac{ad}{bc}}}\right] \qquad \text{(eq. 2.6)}$$

Look at the following data summarised in table. 3.

Table 3: Data for Tetrachoric correlation.

		X variable: Attitude towards women		
		0 (Negative attitude)	1 (Positive attitude)	Sum of row
Attitude towards Liberalisation	0 (Negative attitude)	68 (a)	32 (b)	100
	1 (Positive attitude)	30 (c)	70 (d)	100
	Sum of columns	98	102	total =200

The table values are self explanatory. Out of 200 individuals, 68 have negative attitude towards both variables, 32 have negative attitude to liberalisation but positive attitude to women, and so on. The tetrachoric correlation can be computed as follows.

$$r_{tet} = \cos\left[\frac{180^0}{1+\sqrt{\frac{ad}{bc}}}\right] = \cos\left[\frac{180^0}{1+\sqrt{\frac{(68)(70)}{(30)(32)}}}\right] = \cos 55.784^0 = .722$$

So the tetrachoric correlation between attitude towards liberalisation and attitude towards women is positive.

2.7 RANK ORDER CORRELATIONS

We have learned about Pearson's correlation in earlier unit. The Pearson's correlation is calculated on continuous variables. Pearson's correlation is not advised under two circumstances: one, when the data are in the form of ranks and two, when the assumptions of Pearson's correlation are not followed by the data. In this condition, the application of Pearson's correlations is doubtful. Under such circumstances, rank-order correlations constitute one of the important options. The ordinal scale data is called as rank-order data. Now let us look at these two aspects, rank-order and assumption of Pearson's correlations, in greater detail.

2.7.1 Rank-Order Data

When the data is in rank-order format, then the correlation that can be computed is called as rank order correlations. The rank-order data present the ranks of the individuals or subjects. The observations are already in the rank order or the rank order is assigned to them. Marks obtained in the unit test will constitute a continuous data. But if only the merit list of the students is displayed then the data is called as rank order data. If the data is in terms of ranks, then Pearson's correlation need not be done. Spearman's rho constitutes a good option.

2.7.2 Assumptions Underlying Pearson's Correlation not Satisfied

The statistical significance testing of the Pearson's correlation requires some assumptions about the distributional properties of the variables. We have already delineated these assumptions in the earlier unit. When the assumptions are not followed by the data, then employing the Pearson's correlation is problematic. It should be noted that small violations of the assumptions does not influence the distributional properties and associated probability judgments. Hence it is called as a robust statistics. However, when the assumptions are seriously violated, then application of Pearson's correlation should no longer be considered as a choice. Under such circumstances, Rank order correlations should be preferred over Pearson's correlation.

It needs to be noted that rank-order correlations are applicable under the circumstances when the relationship between two variables is not linear but still it is a *monotonic* relationship. The *monotonic* relationship is one where values in the data are consistently increasing and never decreasing or consistently decreasing and never increasing. Hence, monotonic relationship implies that as X increases Y consistently increase or as X increases Y consistently decrease. In such cases, rank-order is a better option than Pearson's correlation coefficient.

However, some caution should be observed while doing so. A careful scrutiny of Figure 1 below indicates that, in reality, it is a power function. So actually a relationship between X and Y is not linear but curvilinear power function. So, indeed, curve-fitting is a best approach for such data than using the rank order correlation.

The rank-order can be used with this data since the curvilinear relationship shown in figure 1 is also a *monotonic* relationship. It must be kept in mind that all curvilinear relationships would not be monotonic relationships.

In the previous unit, we have discussed the issue of linearity in the section 1.1.2. I have exemplified the non-linear relationship with Yorkes- Dodson law. It states that relationship between stress and performance is non-linear relationship. But this relationship is NOT a monotonic relationship because initially Y increasers with the corresponding increase in X. But beyond the modal value of X, the scores on Y decrease. So this is not a monotonic relationship. Hence, rank-order correlations should not be Calculated for such a data.

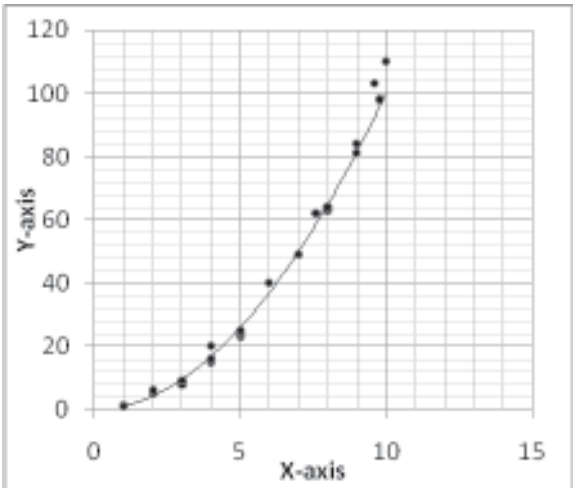


Fig. 1: The figure shows a monotonic relationship between X and Y

2.8 SPEARMAN'S RANK-ORDER CORRELATION OR SPEARMAN'S *RHO* (r_S)

A well-known psychologist and intelligence theorist, Charles Spearman (1904), developed a correlation procedure called in his honor as Spearman's rank-order correlation or Spearman's *rho* (r_s). It was developed to compute correlation when the data is presented on two variables for *n* subjects. It can also be calculated for data of *n* subjects evaluated by two judges for inter-judge agreement. It is suitable for the rank-order data. If the data on X or Y or on both the variables are in rank-order then Spearman's *rho* is applicable. It can also be used with continuous data when the assumptions of Pearson's assumptions are not satisfied. It is used to assess a monotonic relationship.

The range of Spearman's *rho* (r_s) is also from – 1.00 to + 1.00. Like Pearson's correlation, the interpretation of Spearman's *rho* is based on sign of the coefficient and the value of the coefficient.

If the sign of r_s is positive the relationship is positive, if the sign of r_s is negative then the relationship is negative. If the value of r_s is close to zero then relationship is weak, and as the value of r_s approaches to ± 1.00, the strength of relationship increases. When the value of r_s is zero then there is no relationship between X and Y. If r_s is ± 1.00, then the relationship between X and Y is perfect. Whatever the value of r_s may take, it does not directly imply causation. We have already discussed the correlation and causality in the previous unit.

2.8.1 Null and Alternative Hypothesis

The Spearman's *rho* can be computed as a descriptive statistics. We do not carry out statistical hypothesis testing for descriptive use of *rho*. If the r_s is computed as a statistic to estimate population correlation (parameter), then null and alternative hypothesis are required.

The null hypothesis states that

H_O: $\tilde{n}_s = 0$

It means that the value of Spearman's correlation coefficient between X and Y is zero in the population represented by sample.

The alternative hypothesis states that

H_A: $\tilde{n}_s \neq 0$

It means that the value of Spearman's *rho* between X and Y is not zero in the population represented by sample. This alternative hypothesis would require a two-tailed test.

Depending on the theory, the other alternatives could also be written. They are either

H_A: $\tilde{n}_s < 0$

or

H_A: $\tilde{n}_s > 0$.

The first alternative hypothesis, H_A, states that the population value of Spearman's *rho* is smaller than zero. The second H_A denotes that the population value of Spearman's *rho* is greater than zero. Remember, only one of them has to be tested and not both.

You can recall from earlier discussion that one-tailed test is required for this hypothesis.

2.8.2 Numerical Example: for Untied and Tied Ranks

Very obviously, the data on X and Y variables on are required to compute Spearman's *rho*. If the data are on continuous variables then it need to be converted into a rank-orders. The computational formula of Spearman's *rho* (r_s) is as follows:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} \qquad \text{(eq. 2.7)}$$

Where,

r_s = Spearman's rank-order correlation

D = difference between the pair of ranks of X and Y

n = the number of pairs of ranks

Steps:

Let's solve an example. We have to appear for entrance examination after the under-graduate studies. We are interested in correlating the undergraduate marks and performance in the entrance test. We have a data of 10 individuals. But we only have ranks of these individuals in undergraduate examination, and merit list of the entrance performance. We want to find the correlation between rank in undergraduate examination and rank in entrance. The data are provided in table 4 and 5. Since this is a rank order data, we can carry out the Spearman's *rho*. (If the data on one or both variable were continuous, we need to transfer this data into ranks for computing the Spearman's *rho*.)

Table 4: Data for Spearman's rho.

Students	Rank in Undergraduate Examination (X)	Rank in entrance test (Y)
A	1	4
B	5	6
C	3	2
D	6	7
E	9	10
F	2	1
G	4	3
H	10	9
I	8	8
J	7	5

The steps for computation of r_s are given below:

Step 1: List the names/serial number of subjects (students, in this case) in column 1.

Step 2: Write the scores of each subject on X variable (undergraduate examination) in the column labeled as X (column 2), and write the scores of each subject on Y variable (Entrance test) in the column labeled as Y (column 3). We will skip this step because we do not have original scores in undergraduate examination and entrance test.

Step 3: Rank the scores of X variable in ascending order. Give rank 1 to the lowest score, 2 to the next lowest score, and so on. In case of our data, the scores are already ranked.

Step 4: Rank the scores of Y variable in ascending order. Give rank 1 to the lowest score, 2 to the next lowest score, and so on. This column is labeled as R_Y (Column 5). Do cross-check your ranking by calculating the sum of ranks. In case of our data, the scores are already ranked.

Step 5: Now find out D, where $D = R_X - R_Y$ (Column 6).

Step 6: Square each value of D and enter it in the next column labeled as D^2 (Column7). Obtain the sum of the D^2. This is written at the end of the column D^2. This $\sum D^2$ is 18 for this example.

Step 7: Use the equation 4.1 (given below) to compute the correlation between rank in undergraduate examination and rank in entrance test.

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} \qquad \text{(eq. 2.8)}$$

Table 5: Table showing the data on rank obtained in undergraduate examination and ranks in entrance examination. It also shows the computation of Spearman's *rho*.

Students	Rank in Undergraduate Examination (X)	Rank in entrance test (Y)	R_X	R_Y	$D = R_X - R_Y$	D^2
A	1	4	1	4	-3	9
B	5	6	5	6	-1	1
C	3	2	3	2	1	1
D	6	7	6	7	-1	1
E	9	10	9	10	-1	1
F	2	1	2	1	1	1
G	4	3	4	3	1	1
H	10	9	10	9	1	1
I	8	8	8	8	0	0
J	7	5	7	5	2	4
$n = 10$						$\sum D^2=20$

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\times 20}{10(10^2-1)} = 1 - \frac{180}{990} = 1 - 0.1818 = 0.818$$

Now the Spearman's *rho* has been computed for this example. The value of *rho* is 0.818. This value is positive value. It shows that the correlation between the ranks in undergraduate examination and the ranks in entrance test is positive. It indicates that the relationship between them is positively monotonic. The value of the correlation coefficient is very close to 1.00 which indicates that the strength association between the two set of ranks is very high. The tied ranks were not employed in this example since it was the first example. Now I shall introduce you to the problem of tied ranks.

Interesting point need to be noted about the relationship between Pearson's correlation

and Spearman's *rho*. The Pearson's correlation on ranks of X and Y (i.e., R_X and R_Y) is equal the Spearman's *rho* on X and Y. That's the relationship between Pearson's *r* and Spearman's *rho*. The Spearman's *rho* can be considered as a special case of Pearson's *r*.

2.8.3 Spearman's *rho* with Tied Ranks

The ranks are called as *tied ranks* when two or more subjects have the same score on a variable. We usually get larger than the actual value of Spearman's *rho* if we employ the formula in the equation 2.1 for the data with the tied ranks. So the formula in equation 2.1 is not appropriate for tied ranks. A correction is required in this formula in order to calculate correct value of Spearman's *rho*. The recommended procedure of correction for tied ranks is computationally tedious. So we shall use a computationally more efficient procedure. The easier procedure of correction actually uses Pearson's formula on the ranks. The formula and the steps are as follows:

$$r = r_s = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}} \qquad \text{(eq. 2.10)}$$

Where,

r_s = Spearman's *rho*

X = ranks of variable X

Y = rank on variable Y

n = number of pairs

Look at the example we have solved for Pearson's correlation. It is an example of relationship between BHS and BDI. The data is different than the one we have used in the earlier unit. We shall solve this example with Spearman's *rho*.

2.8.4 Steps for r_s with Tied Ranks

If the data are not in ranks, then convert it into rank-order. In this example, we have assigned ranks to X and Y (column 2 and 3) in column 4, and 5.

Appropriately rank the ties (Cross-check the ranking by using sum of ranks check). This is the basic information for the Spearman's *rho*.

Compute the square of rank of X and rank of Y for all the observations. It is in columns 6 and 7.

Multiply the rank of X by rank of Y for each observation. It is provided in column 8.

Obtain sum of all the columns. Now all the basic data for the computation is available.

Enter this data into the formula shown in the equation 2.2 and calculate r_s.

Table 6: Spearman's *rho* for tied ranks

Subject	BHS (X)	BDI (Y)	Rank X	Rank Y	(Rank X)^2	(Rank Y)^2	(Rank X) (Rank Y)
1	7	8	3.5	2.5	12.25	6.25	8.75
2	11	16	6.5	9.5	42.25	90.25	61.75
3	16	14	9	7	81	49	63
4	9	12	5	5.5	25	30.25	27.5
5	6	8	2	2.5	4	6.25	5
6	17	16	10	9.5	100	90.25	95
7	7	9	3.5	4	12.25	16	14
8	11	12	6.5	5.5	42.25	30.25	35.75
9	5	7	1	1	1	1	1
10	14	15	8	8	64	64	64
Sum			55	55	384	383.5	375.75

$$r_s = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]\left[\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right]}} = \frac{375.75 - \frac{(55)(55)}{10}}{\sqrt{\left[384 - \frac{55^2}{10}\right]\left[383.5 - \frac{55^2}{10}\right]}} = \frac{73.25}{81.2496} = 0.902$$

The Spearman's *rho* for this example is 0.902. Since this is a positive value, the relationship between them is also positive. This value is rather near to 1.00. So the strength of association between the ranks of BDI and BHS are very high. This is a simpler way to calculate the Spearman's *rho* with tied ranks. Now, we shall look at the issue of significance testing of the Spearman's *rho*.

2.8.5 Significance Testing of Spearman's *rho*

Once the statistics of Spearman's *rho* is calculated, then the significance of Spearman's *rho* need to be found out. The null hypothesis tested is

H_O: $\tilde{n}_s = 0$

It states that the value Spearman's *rho* between X and Y is zero in the population represented by sample.

The alternative hypothesis is

H_A: $\tilde{n}_s \neq 0$

It states that the value Spearman's *rho* between X and Y is not zero in the population represented by sample. This alternative hypothesis requires a two-tailed test. We have already discussed about writing a directional alternative which requires one-tailed test.

We need to refer to Appendix D for significance testing. The appendix in statistics book, provides critical values for one-tailed as well as two-tailed tests. Let us use the table for the purpose of hypothesis testing for the first example of correlation between ranks in undergraduate examination and entrance test (table 2).

The obtained Spearman's *rho* is 0.818 on the sample of 10 individuals. For $n = 10$, and two-tailed level of significance of 0.05, the critical value of $r_s = 0.648$. The critical value of $r_s = 0.794$ at the two-tailed significance level of 0.01.

The obtained value of 0.818 is larger than the critical value at 0.01. So the obtained correlation is significant at 0.01 level (two-tailed). We reject the null hypothesis and accept the alternative hypothesis. It indicates that the value of the Spearman's *rho* is not zero in the population represented by the sample.

For the second example (table 3), the obtained r_s value is 0.902 on the sample of 10 individuals. For $n = 10$, the critical value is 0.794 at the two-tailed significance level of 0.01. The obtained value of 0.902 is larger than the critical value at 0.01. So the obtained correlation is significant at 0.01 level (two-tailed). Hence, we reject the null hypothesis and accept the alternative hypothesis.

When the sample size is greater than ten, then the *t*-distribution can be used for computing the significance with $df = n - 2$. Following equation is used for this purpose.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \qquad \text{(eq. 2.10)}$$

For the example shown in table 2, the *t*-value is computed using equation 2.11.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} = \frac{0.818\sqrt{10-2}}{\sqrt{1-0.818^2}} = 4.022 \qquad \text{(eq.2.11)}$$

At the $df = 10 - 2 = 8$, the critical *t*-value at 0.01 (two-tailed) is 3.355. The obtained *t*-value is larger than the critical *t*-value. Hence, we reject the null hypothesis and accept the alternative hypothesis.

2.9 KENDALL'S *TAU* (τ)

Kendall's *tau* is another useful measure of correlation. It is as an alternative to Spearman's *rho* (r_s).

This correlation procedure was developed by Kendall (1938). Kendall's *tau* is based on an analysis of two sets of ranks, X and Y. Kendall's *tau* is symbolised as ô, which is a lowercase Greek letter *tau*. The parameter (population value) is symbolised as ô and the statistics computed on the sample is symbolised as. The range of *tau* is from – 1.00 to + 1.00. The interpretation of *tau* is based on the sign and the value of coefficient. The *tau* value closer to ±1.00 indicates stronger relationship. Positive value of *tau* indicates positive relationship and vice versa. It should be noted that Kendall's Concordance Coefficient is a different statistics and should not be confused with Kendall's *tau*.

2.9.1 Null and Alternative Hypothesis

When the Kendall's *tau* is computed as a descriptive statistics, statistical hypothesis testing is not required. If the sample statistic $\tilde{\tau}$ is computed to estimate population correlation ô, then null and alternative hypothesis are required.

The null hypothesis states that

H_O: ô = 0

It stated that the value Kendall's *tau* between X and Y is zero in the population represented by sample.

The alternative hypothesis states that

H_A: $\hat{o} \neq 0$

It states that the value Kendall's *tau* between X and Y is not zero in the population represented by sample. This alternative hypothesis requires a two-tailed test.

Depending on the theory, the other alternatives could be written. They are either

1) H_A: $\hat{o} < 0$ or

2) H_A: $\hat{o} > 0$.

The first H_A denotes that the population value of Kendall's *tau* is smaller than zero.

The second H_A denotes that the population value of Kendall's *tau* is greater than zero. Remember, only one of them has to be tested and not both. One-tailed test is required for these hypotheses.

2.9.2 Logic of τ and Computation

The *tau* is based on concordance and discordance among two sets of ranks. For example, table 4.4 shows ranks of four subjects on variables X and Y as R_X and R_Y. In order to obtain concordant and discordant pairs, we need to order one of the variables according to the ranks, from lowest to highest (we have ordered X in this fashion).

Take a pair of ranks for two subjects A (1,1) and B (2,3) on X and Y.

Now, if sign or the direction of $R_X - R_X$ for subject A and B is similar to the sign or direction of $R_Y - R_Y$ for subject A and B, then the pair of ranks is said to be concordant (i.e., in agreement).

In case of subject A and B, the $R_X - R_X$ is (1 – 2 = – 1) and $R_Y - R_Y$ is also (1 – 3 = – 2). The sign or direction of A and B pair is in agreement. So pair A and B is called as concordant pair.

Look at second example of B and C pair. The $R_X - R_X$ is (2 – 3 = – 1) and $R_Y - R_Y$ is also (3 – 2 = +1). The sign or the direction of B and C pair is not in agreement. This pair is called as discordant pair.

Table 7: Small data example for *tau* on four subjects

Subject	R_X	R_Y
A	1	1
B	2	3
C	3	2
D	4	4

How many such pair we need to evaluate? They will be $n(n-1)/2 = (4 \times 3)/2 = 6$, so six pairs. Here is an illustration: AB, AC, AD, BC, BD, and CD. Once we know the concordant and discordant pairs, then we can calculate by using following equation.

$$\tilde{\tau} = \frac{n_C - n_D}{\left[\frac{n(n-1)}{2}\right]} \qquad \text{(eq. 2.13)}$$

Where,

$\tilde{\tau}$ = value of ô obtained on sample

n_C = number of concordant pairs

n_D = number of discordant pairs

n = number of subjects

Now, I illustrate a method to obtain the number of concordant (n_C) and discordant (n_D) pairs for this small data in the table above. We shall also learn a computationally easy method later.

Step 1. First, Ranks of X are placed in second row in the ascending order.

Step 2. Accordingly ranks of Y are arranged in the third row.

Step 3. Then the ranks of Y are entered diagonally.

Step 4. Start with the first element in the diagonal which is 1 (row 4).

Step 5. Now move across the row.

Step 6. Compare it (1) with each column element of Y. If it is smaller then enter C in the intersection. If it is larger, then enter D in the intersection. For example, 1 is smaller than 3 (column 3) so C is entered.

Step 7. In the next row (row 5), 3 is in the diagonal which is greater than 2 (column 4) of Y, so D is entered in the intersection.

Step 8. Then "C and "D are computed for each row.

Step 9. The n_C is obtained from $\sum\sum$C (i.e., 5) and

Step 10. n_D is obtained from $\sum\sum$D (i.e., 1).

Step 11. These values are entered in the equation 4.4 to obtain.

Table 8. Computation of concordant and discordant pairs.

Subjects	A	B	C	D	∑C	∑D
Rank of X	1	2	3	4		
Rank of Y	1	3	2	4		
	1	C	C	C	3	0
		3	D	C	1	1
			2	C	1	0
				4	0	0
					∑∑C = 5	∑∑D = 1

$$\tilde{\tau} = \frac{n_C - n_D}{\left[\frac{n(n-1)}{2}\right]} = \frac{5-1}{\left[\frac{4(4-1)}{2}\right]} = \frac{4}{6} = 0.667$$

2.9.3 Computational Alternative for τ

This procedure of computing the *tau* is tedious. I suggest an easier alternative. Suppose, we want to correlate rank in practice sessions and rank in sports competitions. We also know the ranks of the sportspersons on both variables. The data are given below for 10 sportspersons.

Table 9: Data of 10 subjects on X (rank in practice session) and Y (ranks in sports competition)

	Subjects being ranked									
	A	B	C	D	E	F	G	H	I	J
Practice session (Ranks on X)	1	2	3	4	5	6	7	8	9	10
Sports competition (Ranks on Y)	2	1	5	3	4	6	10	8	7	9

First we arrange the ranks of the students in ascending order (in increasing order; begin from 1 for lowest score) according to one variable, X in this case. Then we arrange the ranks of Y as per the ranks of X. I have drawn the lines to connect the comparable ranking of X with Y. Please note that lines are not drawn if the subject gets the same rank on both the variables. Now we calculate number of inversions. Number of inversions is number of intersection of the lines. We have five intersections of the lines.

So the following equation can be used to compute $\tilde{\tau}$

$$\tilde{\tau} = 1 - \frac{2(n_s)}{\frac{n(n-1)}{2}} \qquad \text{(eq. 2.14)}$$

Where

$\tilde{\tau}$ = sample value of ô

n_s = number of inversions

n = number of subjects

$$\tilde{\tau} = 1 - \frac{2(n_s)}{\frac{n(n-1)}{2}} = 1 - \frac{2(5)}{\frac{10(10-1)}{2}} = 1 - \frac{10}{45} = 1 - 0.222 = 0.778$$

The value of Kendall's *tau* for this data is 0.778. The value is positive. So the relationship between X and Y is positive. This means as the rank on time taken increases the rank on subject increases. Interpretation of *tau* is straightforward. For example, if the $\tilde{\tau}$ is 0.778, then it can be interpreted as follows: if the pair of subjects is sampled at random, then the probability that their order on two variables (X and Y) is similar is 0.778 higher than the probability that it would be in reverse order. The calculation of *tau* need to be modified for tied ranks. Those modifications are not discussed here.

2.9.4 Significance Testing of ô

The statistical significance testing of Kendall's *tau* is carried out by using either Appendix E and referring to the critical value provided in the Appendix E. The other way is to use the z transformation. The z can be calculated by using following equation

$$z = \frac{\tilde{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} \qquad \text{(eq. 2.15)}$$

You will realise that the denominator is the standard error of *tau*. Once the Z is calculated, you can refer to Appendix A for finding out the probability.

For our example in table 4, the value of $\tilde{\tau} = 0.664$ for the $n = 4$. The Appendix E provides the critical value of 1.00 at two-tailed significance level of 0.05. The obtained value is smaller than the critical value. So it is not statistically significant. Hence, we retain the null hypothesis which states H_O: ô = 0. So we accept this hypothesis. It implies that the underlying population represented by the sample has no relationship between X and Y.

For example in table 6, the obtained value of *tau* is 0.778 with the $n = 10$. From the Appendix E, for the $n = 10$, the critical value of *tau* is 0.644 at two-tailed 0.01 level of significance. The value obtained is 0.778 which is higher than the critical value of 0.664. So the obtained value of *tau* is significant at 0.01 level. Hence, we reject the null hypothesis H_O: ô = 0 and accept the alternative hypothesis H_A: ô ≠ 0. It implies that the value of *tau* in the population represented by sample is other than zero. So there exists a positive relationship between practice ranks and sports competition ranks.

Other way of testing significance is to convert the obtained value of the tau into z. Then use the z distribution for testing the significance of the tau. For this purpose, following formula can be used.

$$z = \frac{\tilde{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} = \frac{0.778}{\sqrt{\frac{2(2\times10+5)}{9\times10(10-1)}}} = 3.313$$

The z table (normal distribution table) in the Appendix A has a value of z = 1.96 at 0.05 level and 2.58 at 0.01 level. The obtained value of z = 3.313 is far greater than these values. So we reject the null hypothesis at 0.01 level of significance.

Kendall's *tau* is said to better alternative to Spearman's *rho* under the conditions of tie ranks. The *tau* is also supposed to do better than Pearson's *r* under the conditions of extreme non-normality. This holds true only under the conditions of very extreme cases. Otherwise, Pearson's *r* is still a coefficient of choice.

2.10 LET US SUM UP

In this unit, we have learned the specific types of correlations that can be used under circumstances that are special. These correlations are either Pearsons correlations with different names or non-Pearson correlations. We have also learned to compute the values as well as test the significances of these correlations. We have also learned the coorelations that can be calculated for the ordinal data. They are Spearman's rho and tau. Indeed we also got to know that Spearman's rho can be considered as a special case of Pearson's correlation. The tau is useful under ties ranks. This will help you to handle the correlation data of various types.

2.11 UNIT END QUESTIONS

1) What are the special types of correlations and why are they to be used?

2) Discuss the point biserial correlation and indicate its advantages.

3) Calculate point biserial for the following data:

Subject	Sex (male) X	Marks (Y)	Subject	Sex (Female (X)	Marks (Y)
1	0	30	11	1	38
2	0	56	12	1	69
3	0	68	13	1	78
4	0	48	14	1	58
5	0	52	15	1	55
6	0	80	16	1	89
7	0	78	17	1	82
8	0	72	18	1	85
9	0	55	19	1	73
10	0	48	20	1	62

4) How will you do the significance of testing for point biserial correlation

5) When do we use Phi Coefficient?

6) Calculate phi coefficient for the following data

X: Gender 0= Male
1 = Female

Y: Ownership of Property 0=No ownership
1 = Ownership

X	1	0	1	1	0	1	1	0	0	1	1	0
Y	1	1	0	0	1	0	0	1	1	0	1	1

7) What is biserial correlation? When do we use biserial correlation?

8) Discuss the use of Tetrachoric correlation.

9) What are the important assumptions of rank order correlation?

10) Discuss in detail Spearman's Rank Correlation and compare it with Kendall's tau.

11) Calculate Rho for the following data and test the significance of Rho

Students	Marks in history	Marks in English
A	50	60
B	45	48
C	63	72
D	65	76
E	48	58
F	59	60
G	62	68

12) Discuss Kendall's Tau.

13) Discuss the significance testing of Tau.

14) Calculate Tau for the following data

	A	B	C	D	E	F	G	H	I	J
Practice session (Ranks on X)	1	2	3	4	5	6	7	8	9	10
Sports competition (Ranks on Y)	5	1	2	4	4	10	6	7	9	8

2.12 SUGGESTED READINGS

Garrett, H.E. (19). *Statistics In Psychology And Education*. Goyal Publishing House, New Delhi.

Guilford, J.P.(1956). *Fundamental Statistics in Psychology and Education.* McGraw Hill Book company Inc. New York.

UNIT 3 PARTIAL AND MULTIPLE CORRELATIONS

Structure

3.0 INTRODUCTION

While learning about correlation, we understood that it indicates relationship between two variables. Indeed, there are correlation coefficients that involve more than two variables. It sounds unusual and you would wonder how to do it? Under what circumstance it can be done? Let me give you two examples. The first is about the correlation between cholesterol level and bank balance for adults. Let us say that we find a positive correlation between these two factors. That is, as the bank balance increases, cholesterol level also increases. But this is not a correct relationship as Cholesterol level can also increase as age increases. Also as age increases, the bank balance may also increase because a person can save from his salary over the years. Thus there is age factor which influences both cholesterol level and bank balance. Suppose we want to know only the correlation between cholesterol and bank balance without the age influence, we could take persons from the same age group and thus control age, but if this is not possible we can statistically control the age factor and thus remove its influence on both cholesterol and bank balance. This if done is called partial correlation. That is, we can use partial and part correlation for doing the same. Sometimes in psychology we have certain factors which are influenced by large number of variables. For instance academic achievement will be affected by intelligence, work habit, extra coaching, socio economic status, etc. To find out the correlation between academic achievement with various other factors ad mentioned above can be done by Multiple Correlation. In this unit we will be learning about partial, part and multiple correlation.

3.1 OBJECTIVES

After completing this unit, you will be able to:

- Describe and explain concept of partial correlation;

- Explain, the difference between partial and semipartial correlation;
- Describe and explain concept of multiple correlation;
- Compute and interpret partial and semipartial correlations;
- Test the significance and apply the correlation to the real data;
- Compute and interpret multiple correlation; and
- Apply the correlation techniques to the real data.

3.2 PARTIAL CORRELATION (r_P)

Two variables, A and B, are closely related. The correlation between them is partialled out, or controlled for the influence of one or more variables is called as partial correlation. So when it is assumed that some other variable is influencing the correlation between A and B, then the influence of this variable(s) is partialled out for both A and B. Hence it can be considered as a correlation between two sets of residuals. Here we discuss a simple case of correlation between A and B is partialled out for C. This can be represented as $r_{AB.C}$ which is read as correlation between A and B partialled out for C. the correlation between A and B can be partialled out for more variables as well.

3.2.1 Formula and Example

For example, the researcher is interested in computing the correlation between anxiety and academic achievement controlled from intelligence. Then correlation between academic achievement (A) and anxiety (B) will be controlled for Intelligence (C). This can be represented as: $r_{\text{Academic Achievement(A) Anxiety (B) . Intelligence (C)}}$. To calculate the partial correlation (r_p) we will need a data on all three variables. The computational formula is as follows:

$$r_p = r_{AB.C} = \frac{r_{AB} - r_{AC} r_{BC}}{\sqrt{(1 - r_{AC}^2)(1 - r_{BC}^2)}} \quad \text{(eq. 3.1)}$$

Look at the data of academic achievement, anxiety and intelligence. Here, the academic achievement test, the anxiety scale and intelligence test is administered on ten students. The data for ten students is provided for the three variables in the table below.

Table 3.1: Data of academic achievement, anxiety and intelligence for 10 subjects

Subject	Academic Achievement	Anxiety	Intelligence
1	15	6	25
2	18	3	29
3	13	8	27
4	14	6	24
5	19	2	30
6	11	3	21
7	17	4	26
8	20	4	31
9	10	5	20
10	16	7	25

In order to compute the partial correlation between the academic achievement and anxiety partialled out for Intelligence, we first need to compute the Pearson's Product moment correlation coefficient between all three variables. We have already learned to compute it in the first Unit of this Block. So I do not again explain it here.

The correlation between anxiety (B) and academic achievement (A) is – 0.369.

The correlation between intelligence (C) and academic achievement (A) is 0.918.

The correlation between anxiety (B) and intelligence (C) is – 0.245.

Give the correlations, we can now calculate the partial correlation .

$$r_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1-r_{AC}^2)(1-r_{BC}^2)}} = \frac{-0.369 - (0.918 \times -0.245)}{\sqrt{(1-0.918^2)(1-(-0.245^2))}} = \frac{-.1441}{0.499} = -0.375$$

(eq.3.2)

The partial correlation between the two variables, academic achievement and anxiety controlled for intelligence, is -0.375. You will realise that the correlation between academic achievement and anxiety is -0.369. Whereas, after partialling out for the effect of intelligence, the correlation between them has almost remained unchanged. While computing this correlation, the effect of intelligence on both the variables, academic achievement and anxiety, was removed.

The following figure explains the relationship between them.

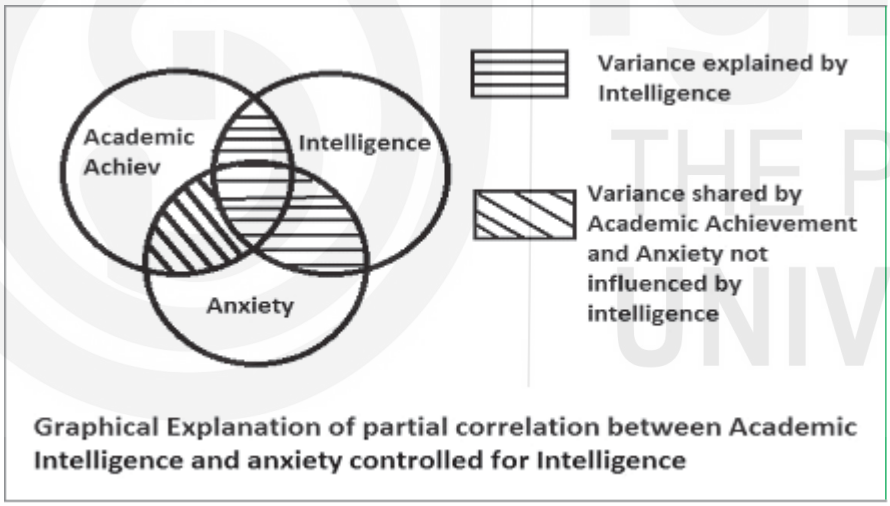


Fig. 3.1: Venn diagram explaining the partial correlation

Significance testing of the partial correlation

We can test the significance of the partial correlation for the null hypothesis

$H_0 : \tilde{n}_P = 0$

and the alternative hypothesis

$H_0: \tilde{n}_P = 0$

Where, the $\tilde{n}_p$ denote the population partial correlation coefficient. The *t*-distribution is used for this purpose. Following formula is used to calculate the *t*-value.

$$t = \frac{r_P\sqrt{n-v}}{\sqrt{1-r_P^2}} \qquad \text{(eq. 3.3)}$$

Where,

r_p = partial correlation computed on sample, $r_{AB.C}$

n = sample size,

v = total number of variables employed in the analysis.

The significance of the r_p is tested at the df = n – v.

In the present example, we can employ significance testing as follows:

$$t = \frac{r_P\sqrt{n-v}}{\sqrt{1-r_P^2}} = \frac{-.375\sqrt{10-3}}{\sqrt{1-(-.375^2)}} = \frac{-0.992}{0.927} = 1.69$$

We test the significance of this value at the df = 7 in the table for t-distribution in the appendix. You will realise that at the df = 7, the table provides the critical value of 2.36 at 0.05 level of significance. The obtained value of 1.69 is smaller than this value. So we accept the null hypothesis stating that $H_0 : \tilde{n}_P = 0$.

Large sample example:

Now we take a relatively large sample example. A counseling psychologist is interested in understanding the relationship between practice of study skills and marks obtained. But she is skeptical about the effectiveness of the study skills. She believes that they can be effective because they are good cognitive techniques or they can be effective simply because the subjects believes that the study skills are going to help them. The first is attribute of the skills while second is placebo effect. She wanted to test this hypothesis. So, along with measuring the hours spent in practicing the study skills and marks obtained, she also took measures on belief that study skill training is useful. She collected the data on 100 students. The obtained correlations are as follows.

The correlation between practice of study skills (A) and unit test marks (B) is 0.69

The correlation between practice of study skills (A) and belief about usefulness of study skills (C) is 0.46

The correlation between marks in unit test (B) and belief about usefulness of study skills (C) is 0.39

$$r_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1-r_{AC}^2)(1-r_{BC}^2)}} = \frac{0.69-(0.46\times 0.39)}{\sqrt{(1-0.46^2)(1-0.39^2))}} = \frac{.51}{0.82} = 0.625$$

The partial correlation between practice of study skills (A) and unit test marks (B) is 0.625. Let's test the null hypothesis about the partial correlation for a null hypothesis which states that $H_0 : \tilde{n}_P = 0$.

$$t = \frac{r_P\sqrt{n-v}}{\sqrt{1-r_P^2}} = \frac{.625\sqrt{100-3}}{\sqrt{1-.625^2}} = \frac{1.65}{0.781} = 2.12$$

The t value is significant at 0.05 level. So we reject the null hypothesis and accept that there is a partial correlation between A and B. This means that the partial correlation between practice of study skills (A) and unit test marks (B) is non-zero at population. We can conclude that the correlation between practice of study skills (A) and unit test marks (B) still exists even after controlled for the belief in the usefulness of the study skills. So the skepticism of our researcher is unwarranted.

3.2.2 Alternative Use of Partial Correlation

Suppose you have one variable which is dichotomous. These variables take two values. Some examples are, male and female, experimental and control group, patients and normal, Indians and Americans, etc. Now these two groups were measured on two variables, X and Y. You want to correlate these two variables. But you are also interested in testing whether these groups influence the correlation between the two variables. This can be done by using partial correlations. Look at the following data. This data is for male and female subjects on two variables, neuroticism and intolerance to ambiguity.

Table 3.2: Table showing gender wise data for IOA and N.

Male		**Female**	
IOA	N	IOA	N
12	22	27	20
17	28	25	15
7	24	20	18
12	32	19	12
14	30	26	18
11	27	23	13
13	29	24	20
10	17	22	9
21	34	21	19

If you compute the correlation between Intolerance of Ambiguity and neuroticism for the entire sample of male and female for 20 subjects. It is – 0.462. This is against the expectation.

This is a surprising finding which states that the as the neuroticism increases the intolerance to ambiguous situations decreases. What might be the reason for such correlation? If we examine the mean of these two variables across gender, then you will realise that the trend of mean is reversed.

If you calculate the Pearson's correlations separately for each gender, then they are well in the expected line (0.64 for males and 0.41 for females).

The partial correlations can help us in order to solve this problem. Here, we calculate the Pearson's product moment correlation between IOA and N partialled out for sex. This will be the correlation between neuroticism and intolerance of ambiguity from which the influence of sex is removed.

$$r_{AB.C} = \frac{r_{AB} - r_{AC}r_{BC}}{\sqrt{(1-r_{AC}^2)(1-r_{BC}^2)}} = \frac{-0.462 - (0.837 \times -0.782)}{\sqrt{(1-0.837^2)(1-(-0.782^2))}} = \frac{.193}{0.341} = 0.566$$

The correlation partialled out for sex is 0.57. Let's test the significance of this correlation.

$$t = \frac{r_P\sqrt{n-v}}{\sqrt{1-r_P^2}} = \frac{.566\sqrt{18-3}}{\sqrt{1-.566^2}} = \frac{2.194}{0.824} = 2.66$$

The tabled value form the appendix at *df* = 15 for 0.05 level is 2.13 and for 0.01 level is 2.95. The obtained t-value is significant at 0.05 level. So we reject the null

hypothesis which stated that population partial correlation, between IOA and N partialled out for sex is zero.

Partial correlation as Pearson's Correlation between Errors

Partial Correlation can also be understood as a Pearson's correlation between two errors.

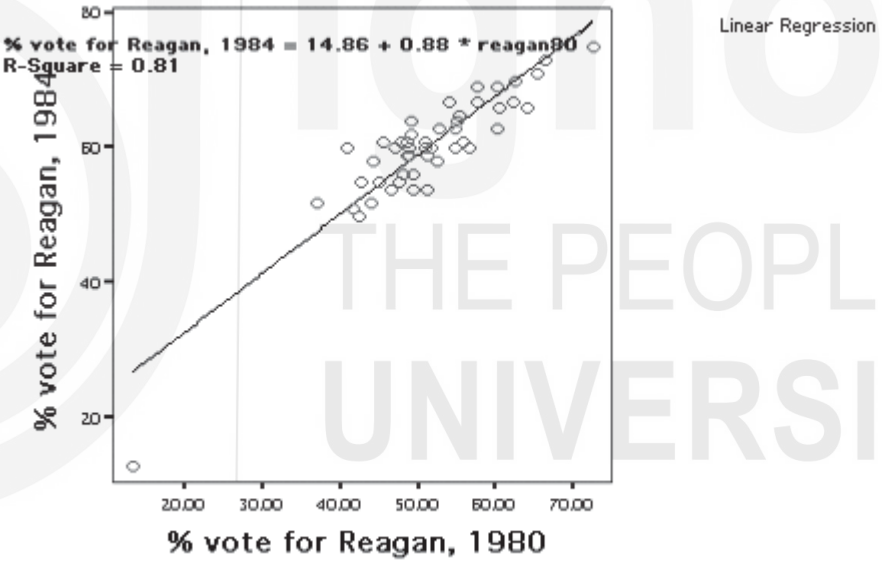
Before you proceed you need to know what is regression equation

3.3 LINEAR REGRESSION

Regression goes one step beyond correlation in identifying the relationship between two variables. It creates an equation so that values can be predicted within the range framed by the data. That is if you know X you can predict Y and if you know Y you can predict X. This is done by an equation called regression equation.

When we have a scatter plot you have learnt that the correlation between X and Y are scattered in the graph and we can draw a straight line covering the entire data. This line is called the regression line.

Here is the line and the regression equation superimposed on the scatterplot**:**



Source: http://janda.org/c10/Lectures/topic04/L25-Modeling.htm

From this line, you can predict X from Y that is % votes in1984 if known, you can find out the % of votes in 1980. Similarly if you know % of votes in 1980 you can know % of votes in 1984.

The regression line seen in the above diagram is close to the scatterplots. That is the predicted values need to be as close as possible to the data. Such a line is called the best fitting line or Regression line. There are certain guidelines for regression lines:

1) Use regression lines when there is a significant correlation to predict values.

2) Do not use if there is not a significant correlation.

3) Stay within the range of the data. For example, if the data is from 10 to 60, do not predict a value for 400.

4) Do not make predictions for a population based on another population's regression line.

The *y* variable is often termed the **criterion variable** and the *x* variable the **predictor variable**. The slope is often called the **regression coefficient** and the intercept the **regression constant**. The slope can also be expressed compactly as $ß_1 = r \times s_y/s_x$.

Normally we then predict values for *y* based on values of *x*. This still does not mean that *y* is caused by *x*. It is still imperative for the researcher to understand the variables under study and the context they operate under before making such an interpretation. Of course, simple algebra also allows one to calculate *x* values for a given value of *y*.

To obtain regression equation we use the following equation:

$$\beta = \{N * \sum xy\} - \{\sum y^2 * \sum y\} / \{(N * \sum x^2) - (\sum y^2)\}$$

Regression equation can also be written including error component 'å'

The regression equation can be written as

$$Y = \alpha + \beta X + \varepsilon \qquad \text{(eq. 4.8)}$$

Where,

Y = dependent variable or criterion variable

á = the population parameter for the y-intercept of the regression line, or regression coefficient (r=óy/ óx)

â = population slope of the regression line or regression coefficient (r*óx/ óy)

å = the error in the equation or residual

The value of á and â are not known, since they are values at the level of population. The population level value is called the parameter. It is virtually impossible to calculate parameter. So we have to estimate it. The two parameters estimated are á and â. The estimator of the á is 'a' and the estimator for â is 'b'. So at the sample level equation can be written as

$$Y = a + bX + e \qquad \text{(eq. 4.9)}$$

Where,

Y = the scores on Y variable

X = scores on X variable

a = the Y-intercept of the regression line for the sample or regression constant in sample

b = the slope of the regression line or regression coefficient in sample

e = error in prediction of the scores on Y variable, or residual

Let us take an example and demonstrate

Example: Write the regression line for the following points:

x	*y*
1	4
3	2
4	1
5	0
8	0

Solution 1: $\sum x = 21; \sum y = 7; \sum x^2 = 115; \sum y^2 = 21; \sum xy = 14$

Thus $ß_0 = [7*115 – 21*14] \div [5 * 115 - 21^2] = 511 \div 134 = 3.81$ and $ß_1 = [5*14 – 21*7] \div [5 * 115 - 21^2] = -77 \div 134 = -0.575$.

Thus the regression equation for this example is $y = -0.575x + 3.81$.

Thus if you have x , then you can find or predict y.

If you have y you can predict x.

Let's continue with the first example.

It was relationship between anxiety and academic achievement. This relationship was controlled for (partialled out for) intelligence.

In this case we can write two linear regression equations and solve them by using ordinary least-squares (OLS). They are as follows:

Academic Achievement = $a_1 + b_1 \times$ Intelligence + e_1

Where, 'a_1' is a y intercept of the regression line;

'b_1' is the slope of the line;

'e_1' is the error in the prediction of academic achievement using intelligence.

Anxiety = $a_2 + b_2 \times$ Intelligence + $å_2$

Where, 'a_2' is a y intercept of the regression line;

'b_2' is the slope of the line;

'e_2' is the error in the prediction of academic achievement using intelligence.

Now we have e_1 and e_2. They are residuals of each of the variables after intelligence explain variation in them. Meaning, e_1 is the remaining variance in academic achievement once the variance accounted for intelligence is removed. Similarly, e_2 is the variance left in the anxiety once the variance accounted for the intelligence is removed.

Now, the partial correlation can be defined as the Pearson's correlation between e_1 and e_2.

$$r_p = e_1 e_2 \qquad \text{(eq. 3.4)}$$

You will realise that this correlation is the correlation of academic achievement and anxiety, from which a linear influence of intelligence has been removed. That is called as partial correlation.

3.4 PART CORRELATION (SEMIPARTIAL CORRELATION) r_{SP}

The Part correlation is also known as semi-partial correlation (r_{sp}). Semipartial correlation or part correlation are correlation between two variables, one of which is partialled for a third variable.

In partial correlations ($r_p = r_{AB.C}$) the effect of the third variable (C) is partialled out from BOTH the variables (A and B).

In semipartial correlations ($r_{sp} = r_{A(B.C)}$), as the name suggests, the effect of third variable (C) was partialled out from only one variable (B) and NOT from both the variables.

Let's continue with the earlier example. The example was about the correlation between anxiety (A) and academic achievement (B).

In the earlier example of partial correlation, we have partialled the effect of intelligence (C) from both academic achievement and anxiety.

One may argue that the academic achievement is r=the only variable that relates to intelligence.

So we need to partial out the effect of the intelligence only from academic achievement and not from anxiety.

Now, we correlate anxiety (A) as one variable and academic achievement partialled for intelligence (B.C) as another variable.

If we correlate these two then, the correlation of anxiety (A) with academic achievement partialled for intelligence (B.C) is called as semipartial correlation ($r_{A(B.C)}$).

In fact, if there are three variables, then total six semipartial correlations can be computed. They are $r_{A(B.C)}$, $r_{A(C.B)}$, $r_{B(A.C)}$, $r_{B(C.A)}$, $r_{C(A.B)}$, and $r_{C(B.A)}$.

Formula:

In order to compute the semipartial correlation coefficient, following formula can be used.

$$r_{SP} = r_{A(B.C)} = \frac{r_{AB} - r_{AC} r_{BC}}{\sqrt{1 - r_{BC}^2}} \qquad \text{(eq. 3.5)}$$

Where,

$r_{A(B.C)}$ is a semipartial correlation of A with the B after linear relationship that C has with B is removed

r_{AB} Pearson's product moment correlation between A and B

r_{AC} Pearson's product moment correlation between A and C

r_{BC} Pearson's product moment correlation between B and C

Example:

Let's take the data from the earlier example of academic achievement, anxiety and intelligence. The data table 3.1 is as follows.

Subject	Academic Achievement	Anxiety	Intelligence
1	15	6	25
2	18	3	29
3	13	8	27
4	14	6	24
5	19	2	30
6	11	3	21
7	17	4	26
8	20	4	31
9	10	5	20
10	16	7	25

The correlation between anxiety (A) and academic achievement (B) is – 0.369.

The correlation between intelligence (C) and academic achievement (B) is 0.918.

The correlation between anxiety (A) and intelligence (C) is – 0.245.

Given the correlations, we can now calculate the semipartial correlation (r_{SP}) as follows. We are not computing the correlation coefficients, simply because you have already learned to compute the correlations earlier. The formula for semipartial correlation is as follows:

$$r_{SP} = r_{A(B.C)} = \frac{r_{AB} - r_{AC} r_{BC}}{\sqrt{1 - r_{BC}^2}} \qquad \text{(eq. 3.6)}$$

Now we can calculate semipartial correlation by using this formula.

$$r_{AB.C} = \frac{r_{AB} - r_{AC} r_{BC}}{\sqrt{1 - r_{BC}^2}} = \frac{-0.369 - (-0.245 \times 0.918)}{\sqrt{1 - (0.918^2)}} = \frac{-.1441}{0.3966} = -0.363$$

The semipartial correlation between anxiety and academic achievement after the linear relationship between the academic achievement and intelligence is removed is -0.363.

The significance of the semipartial correlation can be tested by using t-distribution. The null hpothsis and the alternate hypothesis are as follows.

H_0: $\tilde{n}_{SP} = 0$

H_A: $\tilde{n}_{SP} \neq 0$

Where, the $\tilde{n}_{SP}$ is the semipartial correlation in the population. We test the null hypothesis whether the semipartial correlation in the population is zero. This can be done by using following formula

$$t = \frac{r_{SP}\sqrt{n - v}}{\sqrt{1 - r_{SP}^2}} \qquad \text{(eq. 3.7)}$$

Where,

t = students t-value

r_{SP} = semipartial correlation computed on sample,

n = sample size,

v = number of variables used in the analysis

The significance of this t-value is tested at the df = n – v. when three variables are involved then the df is n – 3.

For our example, the t-values can be computed as follows:

$$t = \frac{-0.363\sqrt{10-3}}{\sqrt{1-(-0.363^2)}} = -1.032$$

The obtained t-value is tested at df = n – v = 10 – 3 = 7.

The t-value at .05 level is 2.364. The obtained t-value is smaller than that. So we accept the null hypothesis that the population semipartial correlation is zero.

It has an interesting implication for our data. The correlation between anxiety and academic achievement is zero in the population if the linear relationship between academic achievement and intelligence is removed.

3.4.1 Semipartial Correlation: Alternative Understanding

Partial Correlation can also be understood as a Pearson's correlation between a variable and error (residual).

Let us continue with the first example. It was relationship between anxiety and academic achievement. This relationship was controlled for (partialled out for) intelligence and academic achievement. (So far you have not learned regression and you may not follow some of the points and equations. So you can revisit this discussion after learning regression.)

In this case we can write a linear regression equation and solve them by using ordinary least-squares (OLS). They are as follows:

Academic Achievement = $a_1 + b_1 \times$ Intelligence + e_1

Where, 'a_1' is a y intercept of the regression line;

'b_1' is the slope of the line;

'e_1' is the error in the prediction of academic achievement using intelligence.

Now we have e_1. It is a residuals of academic achievement after intelligence explain variation in academic achievement.

That is, e_1 is the remaining variance in academic achievement once the variance accounted for intelligence is removed.

Now, the semipartial correlation can be defined as the Pearson's correlation between anxiety and e_1.

You will realise that this correlation is the correlation of academic achievement and anxiety, from which a linear influence of intelligence on academic achievement has been removed.

That is called the semipartial correlation.

(Since you have not learned the regression equation you may not be able to understand this point. So revisit this point after learning regression.)

3.5 MULTIPLE CORRELATION COEFFICIENT (R)

The multiple correlation coefficient denoting a correlation of one variable with multiple other variables. The multiple correlation coefficient is denoted as $R_{A.BCD\ldots k}$ which denotes that A is correlated with B, C, D, up to *k* variables.

For example, we want to compute multiple correlation between A with B and C then it is expressed as $R_{A.BC}$. In this case we create a linear combination of the B and C which is correlated with A.

We continue with the same example which we have discussed for partial and semipartial correlations. This example has academic achievement, anxiety and intelligence as three variables. The correlation between academic achievement with the linear combination of anxiety and intelligence is multiple correlation.

This denotes the proportion of variance in academic achievement explained by intelligence and anxiety. We denote this as

$R_{\text{(Academic Achievement. Intelligence, Anxiety),}}$ which is a multiple correlation.

Often, it is used in the context of regression, where academic achievement is a criterion variable and intelligence and anxiety are called as predictors.

We are not using regression equation since you have not learned it. The Multiple R can be calculated for two predictor variable as follows.

$$R_{A\,.\,BC} = \sqrt{\frac{r_{AB}^2 + r_{AC}^2 - 2r_{AB}r_{AC}r_{BC}}{1 - r_{BC}^2}} \qquad \text{(eq. 3.7)}$$

Where,

$R_{A\,.\,BC}$ = is multiple correlation between A and linear combination of B and C.

r_{AB} = is correlation between A and B

r_{AC} = is correlation between A and C

r_{BC} = is correlation between B and C

Example

We shall continue with the earlier data.

The data table 3.1 is as follows.

Subject	Academic Achievement	Anxiety	Intelligence
1	15	6	25
2	18	3	29
3	13	8	27
4	14	6	24
5	19	2	30
6	11	3	21
7	17	4	26
8	20	4	31
9	10	5	20
10	16	7	25

The correlation between anxiety (A) and academic achievement (B) is – 0.369.

The correlation between intelligence (C) and academic achievement (B) is 0.918.

The correlation between anxiety (A) and intelligence (C) is – 0.245.

The multiple correlation can be calculated as follows.

$$R_{A.BC} = \sqrt{\frac{r_{AB}^2 + r_{AC}^2 - 2r_{AB}r_{AC}r_{BC}}{1 - r_{BC}^2}}$$

$$= \sqrt{\frac{-0.369^2 + 0.918^2 - 2 \times -.369 \times .918 \times -.245}{1 - (-0.245)^2}}$$

$$= \sqrt{\frac{0.813}{0.94}}$$

$$= 0.929$$

This means that the multiple correlation between academic achievement and the linear combination of intelligence and anxiety is 0.929 or 0.93. We have earlier learned that the square of the correlation coefficient can be understood as percentage of variance explained.

The R^2 is then percentage of variance in academic achievement explained by the linear combination of intelligence and anxiety. In this example the R^2 is 0.929^2 which is 0.865. The linear combination of intelligence and anxiety explain 86.5 percent variance in the academic achievement.

We have already converted the R into the R^2 value. The R^2 is the value obtained on a sample. The population value of the R^2 is denoted as P^2. The R^2 is an estimator of the P^2.

But there is a problem in estimating the P^2 value from the R^2 value.

The R^2 is not an unbiased estimator of the P^2.

So we need to adjust the value of the R^2 in order to make it unbiased estimator.

Following formula is used for this purpose.

Let $\tilde{R}^2$ denote an adjusted R^2. Then $\tilde{R}^2$ can be computed as follows:

$$\tilde{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1} \qquad \text{(eq. 3.8)}$$

Where,

$\tilde{R}^2$ = adjusted value of R^2

k = number of predicted variables (or the variable for which a linear combination is created)

n = sample size

For our example the $\tilde{R}^2$ value need to be computed.

$$\tilde{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

$$\tilde{R}^2 = 1 - \frac{(1-0.865)(10-1)}{10-2-1}$$

$$\tilde{R}^2 = 1 - \frac{1.217}{7} = 0.826$$

So the unbiased estimator of the R^2 the adjusted value, $\tilde{R}^2$, is 0.826 which is smaller than the value of R^2. It is usual to get a smaller adjusted value.

The significance testing of the R:

This can be used for the purpose of the significance testing. The null hypothesis and the alternative hypothesis employed for this purpose are

$$H_0 : P^2 = 0$$
$$H_A : P^2 \neq 0$$

The null hypothesis denotes that the population R^2 is zero whereas the alternative hypothesis denotes that the population R^2 is not zero.

The *F*-distribution is used for calculating the significance of the R^2 as follows:

$$F = \frac{(n-k-1)R^2}{k(1-R^2)} \qquad \text{(eq. 3.9)}$$

The value of R^2 can be adjusted R^2 ($\tilde{R}^2$ = .825) value of the R^2 value (.865).

When the sample size is small, it is recommended that $\tilde{R}^2$ value be used. As the sample size increase the difference between the resulting F values reduce considerably. Since our sample is obviously small, we will use unbiased estimator.

$$F = \frac{(n-k-1)R^2}{k(1-R^2)}$$

$$F = \frac{(10-2-1)\times 0.826}{2\times(1-0.826)} = \frac{5.783}{0.348} = 16.635$$

The degrees of freedom employed in the significance testing of this F value are $df_{num} = k$ and $df_{denominator}$ = n – k – 1.

For our example the degrees of freedom are as follows:

$df_{num} = k = 2$

$df_{denominator}$ = n – k – 1 = 10 – 2 – 1 = 7

The tabled value for the $F_{(2,7)} = 4.737$ at 0.05 level of significance and the $F_{(2,7)} = 9.547$ at 0.01 level of significance. The calculated value of the F 16.635 is greater than the critical value of F. so we reject the null hypothesis and accept the alternative hypothesis which stated that the P^2 is not zero at less than 0.01 level.

You could have computed the F value using the R^2 instead of adjusted R^2 ($\tilde{R}^2$).

Often the statistical packages report the significance of R^2 and not significance of adjusted R^2 ($\tilde{R}^2$).

It is the judgment of the researcher to use either of them. In the same example if R^2 value is substituted for the adjusted R^2 ($\tilde{R}^2$) value then the F is 22.387 that is significant at .01 level.

3.6 LET US SUM UP

In this unit we have learned about the interesting procedures of computing the correlations. Especially, when we are interested in controlling for one or more variable. The multiple correlations provide us with an opportunity to calculate correlations between a variable and a linear combination of other variable. You practice them by solving some of the example given below, and you will understand the use of them.

3.7 UNIT END QUESTIONS

Problems

1) A clinical psychologist was interested in testing the relationship between health and stress. But she realised that coping skills will have an influence on this relationship so she administered General Health Questionnaire, Stress Scale and a coping scale. The data was collected on 15 individuals. Calculate the multiple correlation for this problem and test the significance. The data is as follows:

Health	Stress	Coping
9	5	18
10	5	21
8	7	17
7	4	16
8	7	22
9	6	19
12	8	25
11	3	17
10	5	20
14	6	22
7	7	18
9	9	22
6	7	17
16	3	20
14	8	26

In addition, answer the following questions:

Which correlation coefficient she should compute, if she wants to control the relationship between stress and health for the effect of coping?

Write a null and alternative hypothesis for partial correlation.

Calculate the partial correlation between stress and health controlled for the effect of coping.

Test the significance of the relationship.

Which correlation coefficient she should compute, if she wants to control the relationship between stress and health for the effect of coping only on stress?

Write a null and alternative hypothesis for part correlation.

Calculate the part correlation between stress and health controlled for the effect of coping on stress and test the significance.

Which correlation coefficient she should compute, if she wants to know the relationship between health (Y) and linear combination of stress (X_1) and coping (X_2)?

Answer:

a) Partial correlation; b) H_0: $r_p = 0$ and H_A : $r_p \neq 0$; c) –.77; d) significant at 0.01 level; e) part (semipartial); f) H_0: $r_{sp} = 0$ and H_A : $r_{sp} \neq 0$; g) –.62, $p < .05$. h) Multiple correlation; i) $R = 0.86$, $p < .01$.

2) A social psychologist was interested in testing the relationship between attitude towards women (ATW) and openness to values (OV). But she realised education (EDU) will influence this relationship so she administered attitude to women scale, openness to values scale and also recorded the years spent in formal education. The data was collected on 10 individuals.

Calculate the multiple correlation for this problem and test the significance.

The data is as follows:

ATW	OV	Edu
2	7	14
4	10	13
8	14	11
7	13	9
8	9	5
9	10	14
1	6	5
0	9	6
6	12	11
5	10	12

In addition, answer the following questions:

1) Which correlation coefficient she should compute, if she wants to control the relationship between ATW and OV for the effect of EDU?

2) Write a null and alternative hypothesis for partial correlation.

3) Calculate the partial correlation between ATW and OV controlled for the effect of EDU.

4) Test the significance of the relationship.

5) Which correlation coefficient she should compute, if she wants to control the relationship between ATW and OV for the effect of EDU only on OV?

6) Write a null and alternative hypothesis for part correlation.

7) Calculate the part correlation between ATW and OV controlled for the effect of EDU on OV and test the significance.

8) Which correlation coefficient she should compute, if she wants to know the relationship between ATW (Y) and linear combination of OV (X_1) and EDU (X_2)?

9) Write a null and alternative hypothesis for multiple correlation.

Answer:

a) Partial correlation; b) H_0: $r_p = 0$ and H_A : $r_p \neq 0$; c) .64; d) insignificant (the n is small); e) part (semipartial); f) H_0: $r_{sp} = 0$ and H_A : $r_{sp} \neq 0$; g) 0.61, p >.05. h) Multiple correlation; i) R = 0.67, p >.05.

3.8 SUGGESTED READINGS

Garrett, H.E. (19). *Statistics In Psychology And Education.* Goyal Publishing House, New Delhi.

Guilford, J.P.(1956). *Fundamental Statistics in Psychology and Education.* McGraw Hill Book company Inc. New York.

UNIT 4 BIVARIATE AND MULTIPLE REGRESSION

Structure

4.0 INTRODUCTION

Psychologists, as other scientists, are also interested in prediction. Since our domain of enquiry relates with human behaviour, our predictions are associated with human behaviour. We are interested in knowing how human beings will behave provided we have some information about them. It is not that we all the time depend on theories such as psychoanalysis, behaviourism or cognitive in order to predict human behaviour. There are also statistical methods which can help predict certain phenomenon of human behaviour. We would study in this unit the statistical methods that can be used for the purpose of prediction. These statistical methods are called Regression. We will first learn the concept of regression, then learn how to plot the relationship between variables, and learn to work out The Regression Equation. We will also deal with how far we can be accurate in predicting with the help of regression equation by the help of tests of significance. Finally we will be dealing with how to interpret regression and deal with also Multiple regression, that is, which variables influence a particular phenomenon.

4.1 OBJECTIVES

After completing this unit, you will be able to:

- Describe and explain concept of regression correlation;
- Explain, describe and differentiate between bivariate regression and multiple regresion;

- Describe and explain concept of multiple correlation;
- Develop a regression equation;
- Compute the a and b of bivariate regression by using OLS;
- Test the significance of regression;
- Interpret regression results;
- Apply the regression techniques to the real data;
- Explain Multiple regression; and
- Use Multiple regression in real data.

4.2 BIVARIATE AND MULTIPLE REGRESSION

We always see that the meteorology department predicts the rain, the economists predict the outcome of a particular policy, financial experts predict the share market, the election experts predict the outcome of voting and so on. Similarly using statistical method, psychologists too can predict certain human behaviours. For example, one can predict the examination marks after writing the examination by checking what questions we have attempted and how we have faired in it and what marks we can expect for each question and so on.

Most of us are interested in this exercise of prediction. On many occasions, we also predict the behaviour of our friends, colleagues and family. Predicting and trying to speculate about what might happen in future is integral part of human curiosity. While there are many theories of psychology and personality that would help us predict behaviours, one can also predict a certain phenomenon in terms of statistics. This method is called regression in statistics.

The simplest form of the regression is simple linear regression (at times also called as bivariate regression). Carl Frederick Gauss discovered a method of least squares (1809) and later on developed Gauss-Markov theorem (1821). Sir Fransis Galton contributed to the method of regression and also gave the name.

Let us see what this is all about. Let us say we have data on two variables (Y and X), and we create an equation, called regression equation, which later on helps us in predicting the score of one variable (Y) by simply using the scores on another variable (X). Let us learn about the utility of the regression analysis, how to do it, how to test the significance and issues surrounding it.

Regression analysis tries to predict Y variable from X variable. In the general form, it tries to predict Y from a X_1, X_2, …, X_k, where k is number of predictor variables. Initially we will learn about two variable prediction, one of which is a predictor and the other one will be predicted. Then we will look at the general form of Regression.

Just think of the variables that can be used in prediction in psychology. Look at the following statements. (See the box below)

- Stress leads to health deterioration.
- Openness increases creativity.
- Extraversion increases social acceptance.

- Social support influences coping with mental health problems.
- Stigma about mental illness decides the help seeking behaviour.
- Parental intelligence leads to child's intelligence
- Attitude to job and attrition depends on affective commitment to the organisation.

What do you see in common in all these statements?

All the statements above have two variables.

One of the variables can potentially predict the other variable.

4.2.1 Predicting One Variable from Another

Let us now consider the problem of prediction. How to predict Y from X.

The Y variable is called the dependent variable. It is also called as criterion variable. It is the variable that has to be predicted.

The X variable is a called as independent variable. It is also called as predictor variable (Please note that in experimental psychology we define independent variable as the variable that is manipulated by the experimenter, whereas in regression the term is used less strictly. In Regression, the independent variable is not manipulated by the researcher or experimenter.)

If X is predicting Y, then typically it is said that 'X is regressed on Y'.

Let's identify the X and Y in our statements given in the box.

In the first statement Stress (X) lead to the health (Y) deterioration

In the second statement, Openness (X) increases the creativity (Y).

In the third statement, Extroversion (X) increases the social acceptance (Y).

In fourth statement, Social support (X) influences the coping with mental health problems (Y).

In the fifth statement, Stigma about mental illness (X) decided the help seeking behaviour (Y).

In the sixth statement, Parental intelligence (X) leads to child's intelligence (Y).

In the last statement, Attitude (Y) to job and attrition depends on affective commitment(X) to the organisation

Before we learn how to do the regression, we shall quickly browse through the basic concepts in regression analysis.

4.2.2 Plotting the Relationship

We have already learned to plot the scatter plots. We shall try to plot a scatter and try to understand regression graphically.

The perfect relationship

Look at the following example. You have data of five swimmers on two variables, hours of practice per day (X) and time taken (Y).

Swimmer	hours of practice per day	Time taken (in seconds)
A	1	50
B	2	45
C	3	40
D	4	35
E	5	30

Now plot the relationship between them as a scatter. You know how to do that. We have now tried to draw a line that passes through all the data points in the scatter. And we have successfully done it.

Looking at figure 4.1 you realise that as the number of hours spent in practice increase the time taken is reducing. There is a perfect linear relationship between them. This means that you can draw a line on the scatter that passes through all the data points on the scatter.

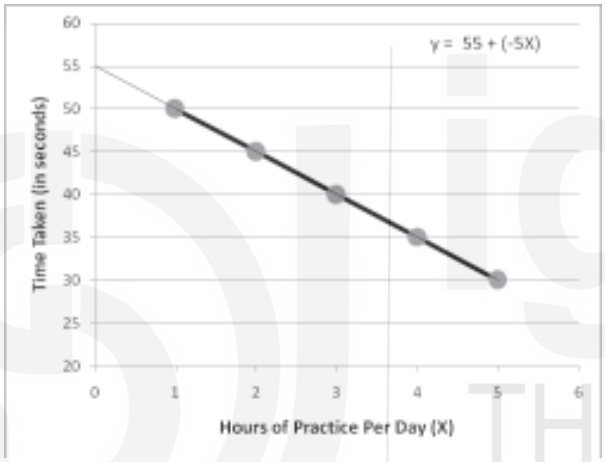


Fig. 4.1: Figure showing the data between number of hours spent in practice and time take.

For this data, the slope of the line can be calculated by using a simple technique.

$$Slope = \frac{Y_2 - Y_1}{X_2 - X_1} \qquad \text{(eq. 4.1)}$$

Where Y_2 and Y_1 are any two points on Y axis and X_2 and X_1 are corresponding two points on X axis.

For example, take $Y_2 = 45$ and $Y_1 = 40$ and corresponding X_2 and X_1 are 2 and 3. The slope is

$$Slope = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{45 - 40}{2 - 3} = \frac{5}{-1} = -5 \qquad \text{(eq. 4.2)}$$

The slope of the line is – 5.

The point at which the line passes through the Y axis (the Y intercept of the line) is 55.

Now, if we ask about the unknown score, 6 hours of practice per day, then the predicted X score is 25 seconds (which is very close the world record).

How have we obtained it? we have solved it for a equation of straight line. That equation is

$$Y = a + bX \quad \text{(eq. 4.3)}$$

Where a = point where the line passes the Y axis and

b = is a slope of the line.

We have $a = 55$ and $b = -5$. So for X = 6 the Y will be

(eq. 4.4)

The Imperfect Relationship.

But the problem is the real data will not be so systematic and all data points in scatter will not fall on a straight line.

Look at the following example of the stigma and visits to mental health professionals. The Table 4.2 shown below display the data of stigma and number of appointments missed to mental health professional.

Table 4.2: Data of stigma and number of appointments missed to mental health professional

Patient	Stigma scores	Number of appointments missed
1	60	5
2	50	2
3	70	9
4	73	6
5	64	9
6	68	4
7	56	3
8	54	8
9	49	3
10	66	11

This data was obtained from ten patients who are suffering due to mental illness. The data was collected on King, Show and others (2007) Stigma scale and the data were obtained on number of visits missed by the patients. The data is plotted in the scatter plot below.

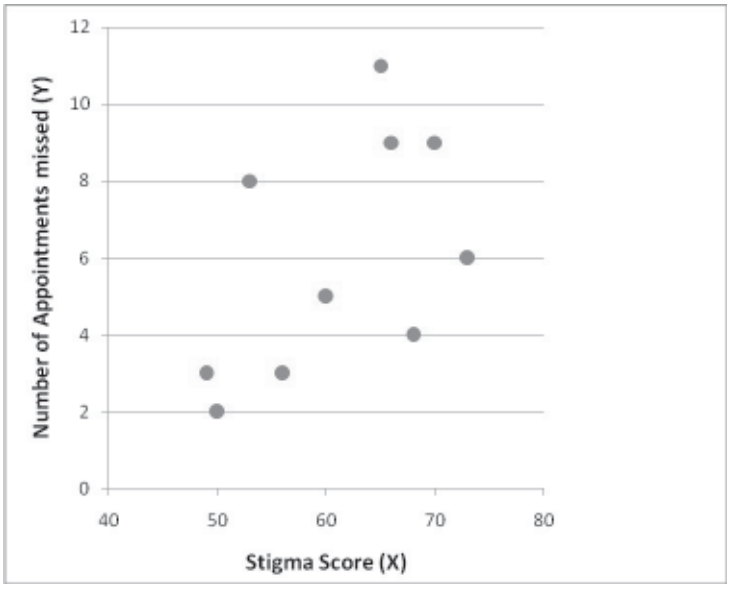


Fig. 4.2: Scatter showing the relationship between stigma and number of appointments missed

Now you will realise that it is not possible to draw a straight line that passes through all the data points. Then how to know the relationship between X and Y and then predict the scores of Y from scores of X. How to draw the straight line for this data? This is a problem one would face with real data. The linear regression analysis solves this problem.

4.2.3 Mean, Variance and Covariance: Building Blocks of Regression

In order to understand the building blocks of regression we must describe some of the terms such as (i) the mean, (ii) variance, and (iii) covariance which are presented in the following section.

i) Mean

Mean of variable X (symbolised as $\bar{X}$) is sum of scores ($\sum_{i=1}^{n} X_i$) divided by number of observations (n). The mean is calculated in following way.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(eq. 4.5)}$$

You have learned this in the first block. We will need to use this as a basic element to compute correlation.

ii) Variance

The variance of a variable X (symbolised as S_X^2) is the sum of squares of the deviations of each X score from the mean of X ($\sum (X - \bar{X})^2$) divided by number of observations (n).

$$^{2}_{X} = \frac{\sum (X - \bar{X})^2}{n} \qquad \text{(eq. 4.6)}$$

You have already learned that standard deviation of variable X , symbolised as S_X , is square root of variance of X, symbolised as .

iii) Covariance

The covariance between X and Y (Cov_{XY} or S_{XY}) can be stated as

$$Cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} \qquad \text{(eq. 4.7)}$$

Covariance is a number that indicates the association between two variables. To compute covariance, deviation of each score on X from its mean ($\bar{X}$) and deviation of each score on Y from its mean ($\bar{Y}$) is initially calculated.

Then products of these deviations are obtained.

Then, these products are summated.

This sum gives us the numerator for covariance.

Divide this sum by number of observations (n).

The resulting number is covariance.

4.2.4 The Regression Equation

The regression equation can be written as

$$Y = \alpha + \beta X + \varepsilon \quad \text{(eq. 4.8)}$$

Where,

Y = dependent variable or criterion variable

á = the population parameter for the y-intercept of the regression line, or regression coefficient (r=óy/ óx)

â = population slope of the regression line or regression coefficient (r*óx/ óy)

å = the error in the equation or residual

The value of á and â are not known, since they are values at the level of population. The population level value is called the parameter. It is virtually impossible to calculate parameter. So we have to estimate it. The two parameters estimated are á and â. The estimator of the á is 'a' and the estimator for â is 'b'. So at the sample level equation can be written as

$$Y = a + bX + e \quad \text{(eq. 4.9)}$$

Where,

Y = the scores on Y variable

X = scores on X variable

a = the Y-intercept of the regression line for the sample or regression constant in sample

b = the slope of the regression line or regression coefficient in sample

e = error in prediction of the scores on Y variable, or residual

$$\hat{Y} = a + bX \quad \text{(eq. 4.10)}$$

Where, Y = predicted value of Y in sample. This value is not an actual value but the value of Y that is predicted using the equation $\hat{Y} = a + bX$. So we can write error as by substituting the in the earlier equation.

$$S_x^2 \quad \text{(eq. 4.11)}$$

$$Y - \hat{Y} = e \quad \text{(eq. 4.12)}$$

This is a useful expression. We shall use it while computing the statistical significance of the regression and will also be useful for understanding the least squares.

4.2.5 Ordinary Least Squares (OLS)

Just recall the data between the stigma scores and number of appointments missed by the person. Now, if we have to draw the straight line that will explain the relationship between the stigma scores and number of appointments missed, then there will be many such lines possible. Out of them, which line we should consider as the best fit line?

It is not possible to draw a straight line that will pass through all the points. And many lines are possible with the earlier equation $Y = a + bX + e$

This problem is solved by the method of least squares or ordinary least squares (OLS).

One easy way to judging how good the line is, is to know how close various values of $\hat{Y}$ are to corresponding values of Y, which means to check how close predicted value ($\hat{Y}$) is to the actual value of the Y.

These predicted values are computed by using the various values of X in the data.

But how to decide what is the best fit?

One logical solution to this problem is to look at the error term, e. the e is defined as

$$Y - \hat{Y} = e$$

Which means,

$$Y - (a + bX) = e \qquad \text{(eq. 4.13)}$$

The $Y - \hat{Y}$ is the error in prediction of the Y.

This error is called as an obtained residual of the regression.

The best line is the one that minimises this residual.

Some of the predicted values of Y will be higher than the actual value of Y and some would be lower, and hence the sum of residual will be zero.

In order to take care of this problem, the summation is not done over the $Y - \hat{Y}$ Instead the $\sum\left(Y - \hat{Y}\right)^2$ is summated. An attempt to *minimise the sum of the squared errors* — minimise the $\sum\left(Y - \hat{Y}\right)^2$ is made, this is called as least squares.

Calculation of a and b

The values for *a* and *b* that minimises the sum of the squared errors — minimise the $\sum\left(Y - \hat{Y}\right)^2$ need to be calculated. The *b* can be calculated as follows.

$$b = \frac{Cov_{XY}}{S_X^2} \qquad \text{(eq. 4.14)}$$

Where,

Cov_{XY} = covariance between X and Y. This is given by the formula $\sum$(X-X)(Y-Y) / N S_X^2 = variance of X

The *b* is covariance of X and Y divided by the variance of X. it can be rewritten as

$$b = \frac{\frac{\sum(X - \bar{X})(Y - \bar{Y})}{n}}{S_X^2} \qquad \text{(eq. 4.15)}$$

$$b = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{nS_X^2} \qquad \text{(eq. 4.16)}$$

The *a* can be calculated as follows by using our earlier equation.

$$\bar{Y} = a + b\bar{X} \qquad \text{(eq. 4.17)}$$

$$a = \bar{Y} - b\bar{X} \qquad \text{(eq. 4.18)}$$

Once we know how to calculate *a* and *b*, then we can solve the problem of regression. Let's now solve the example we have started with. The example was about the predicting the number of appointments missed by the patient (Y) by using the Stigma scale scores (X). The data is as follows:

Table 4.3: Table showing the computation of *a* and *b*.

Patient	Stigma scores	Number of appointments missed	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	60	5	-1	-1	1	1	1
2	50	2	-11	-4	121	16	44
3	70	9	9	3	81	9	27
4	73	6	12	0	144	0	0
5	64	9	3	3	9	9	9
6	68	4	7	-2	49	4	-14
7	56	3	-5	-3	25	9	15
8	54	8	-7	2	49	4	-14
9	49	3	-12	-3	144	9	36
10	66	11	5	5	25	25	25
X	$\sum X$ =610	$\sum Y$ =60			$\sum (X - \bar{X})^2$ = 648	$\sum (Y - \bar{Y})^2$ = 86	$\sum (X - \bar{X})(Y - \bar{Y})$ = 129
n = 10	$\bar{X} = 61$	$\bar{Y} = 6$					

$$S_X = \sqrt{\sum (X - \bar{X})^2 / n} = 8.50$$

$$S_Y = \sqrt{\sum (Y - \bar{Y})^2 / n} = 2.93$$

$$Cov_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n} = \frac{129}{10} = 12.9$$

$$b = \frac{Cov_{XY}}{S_X^2} = \frac{12.9}{8.50^2} = \frac{12.9}{64.8} = 0.1991$$

$$a = \bar{Y} - b\bar{X} = 6 - (0.1991 \times 61) = -6.144$$

Step 1. You need scores of subjects on two variables. We have scores on ten subjects on two variables, the Stigma scores (X) and number of appointments missed (Y).

Then list the pairs of scores on two variables in two columns.

The order will not make any difference.

Remember, same individuals' two scores should be kept together.

Label the predictor variable as X and criterion as Y.

Step 2. Compute the mean of variable X and variable Y. It was found to be 61 and 6 respectively.

Step 3. Compute the deviation of each X score from its mean ($\bar{X}$) and each Y score from its own mean ($\bar{Y}$). This is shown in the column labeled as $X - \bar{X}$ and $Y - \bar{Y}$. As you have learned earlier, the sum of these columns has to be zero.

Step 4. Compute the square of $X - \bar{X}$ and $Y - \bar{Y}$. This is shown in next two columns labeled as $(X - \bar{X})^2$ and $(Y - \bar{Y})^2$. Then compute the sum of these squared deviations of X and Y.

The sum of squared deviations for X is 648 and for Y it is 86.

Divide them by *n* to obtain the standard deviations for X and Y. The was found to be 8.49. Similarly, the S_y was found to be 3.09.

Step 5. Compute the cross-product of the deviations of X and Y. These cross-products are shown in the last column labeled as $(X - \bar{X})$ and $(Y - \bar{Y})$. Then obtain the sum of these cross-products. It was found to be 129. Now, we have all the elements required for computing *b*.

Step 6. Compute the covariance between X and Y, which turned out to be 12.9.

Step 7. Compute the *b* value by dividing the covariance XY (Cov_{XY}) by the variance of . We compute S_x^2 by taking *n* as a denominator the S_x^2 value is 64.8. The b is found to be 0.1991. Now we can easily compute the *a* which is

$$a = \bar{Y} - b\bar{X} = 6 - (0.1991 \times 61) = -6.144 .$$

Once the *a* and *b* are computed, we can write the regression equation to get the predicted values of Y as follows:

$$\hat{Y} = a + bX \qquad \text{(eq. 4.19)}$$

$$\hat{Y} = -6.144 + (0.1991 \times X)$$

Now we can compute the predicted values for each of the X value. For example the predicted value for the first X value (60) is as follows:

$$5.80 = -6.144 + (0.1991 \times 60)$$

In this way you can compute the predicted Y value for each of the X score. Now you realise that this value is not Y value but the predicted Y value obtained from X.

Now look at the table below. It gives the X, Y and Predicted Y values.

Table 4.4: Table showing the computation of the significance for the *b*, the slope of the line

Ss	Stigma scores (X)	Number of appointments missed (Y)	Predicted value of Y	Residual $Y-\hat{Y}=e$	Residual $(Y-\hat{Y})^2=e$	Variance explained $\hat{Y}-\overline{Y}$	Variance explained Squared $(\hat{Y}-\overline{Y})^2$
1	60	5	5.80	-0.80	0.64	-0.20	0.04
2	50	2	3.81	-1.81	3.28	-2.19	4.80
3	70	9	7.79	1.21	1.46	1.79	3.21
4	73	6	8.39	-2.39	5.71	2.39	5.71
5	64	9	6.60	2.40	5.77	0.60	0.36
6	68	4	7.39	-3.39	11.52	1.39	1.94
7	56	3	5.00	-2.00	4.02	-1.00	0.99
8	54	8	4.61	3.39	11.52	-1.39	1.94
9	49	3	3.61	-0.61	0.37	-2.39	5.71
10	66	11	7.00	4.00	16.04	1.00	0.99
Sum	610	60	60	0	60.32	0	25.68

With the availability of residual, we can obtain the sum of squared residual. The sum of squared residual is 60.32. This is the minimum value that can be obtained if a straight line is drawn for the relationship between X and Y.

There is no other line than can give value as small as this.

So this line is considered as a best fit line.

The mean of Y is 6. So we can now obtain an interesting expression. This expression is $\hat{Y}-\overline{Y}$.

This will provide us the amount of variance in Y explained by the predicted value of Y which is $\hat{Y}$.

The sum of this difference is bound to be zero. So we square the difference.

The sum of square of the difference between predicted value of Y and mean of Y is given below:

$$\sum(\hat{Y}-\overline{Y})^2 ,$$

This is the amount of variance explained in the Y by the predicted value of the Y. This can be expressed as follows:

Total Variance in Y = Variance Explained by Regresion + Residual variance

(eq. 4.20)

This can be written as

$$Y-\overline{Y}=(\hat{Y}-\overline{Y})+(Y-\hat{Y}) \quad \text{(eq. 4.21)}$$

$$\sum Y-\overline{Y}=\sum(\hat{Y}-\overline{Y})+\sum(Y-\hat{Y}) \quad \text{(eq. 4.22)}$$

Since the summation of these differences are zero, we square the difference. The equation can be rewritten as

$$\sum (Y-\bar{Y})^2 = \sum (\hat{Y}-\bar{Y})^2 + \sum (Y-\hat{Y})^2 \qquad \text{(eq. 4.23)}$$

Where,

$\sum (Y-\bar{Y})^2$ = Total variance in Y. Total sum of squares (SS_T).

$\sum (\hat{Y}-\bar{Y})^2$ = Variance in Y explained by X. Sum of squares explained ($SS_{Regression}$).

$\sum (Y-\hat{Y})^2$ = variance in Y not explained by X. Residual sum of squares ($SS_{Residual}$).

$$SS_{Total} = SS_{Regresion} + SS_{Residual} \qquad \text{(eq. 4.24)}$$

Look at the figure below. You will understand the division of SS_{Total} into $SS_{Regression}$ and $SS_{Residual}$.

It shows that the distance between the $\bar{Y}$ and Y is total deviation of that Y value from $\bar{Y}$. This is shown as $(Y-\bar{Y})$.

From this total deviation or variation, the explained variation is distance between $\bar{Y}$ and the predicted Y value. This is shown as $\hat{Y}-\bar{Y}$.

This is explained by the regression line. The distance that regression equation fails to explain is between Y and predicted value of Y. This distance is residual or remaining variance that regression equation cannot explain. This is shown as $Y-\hat{Y}$.

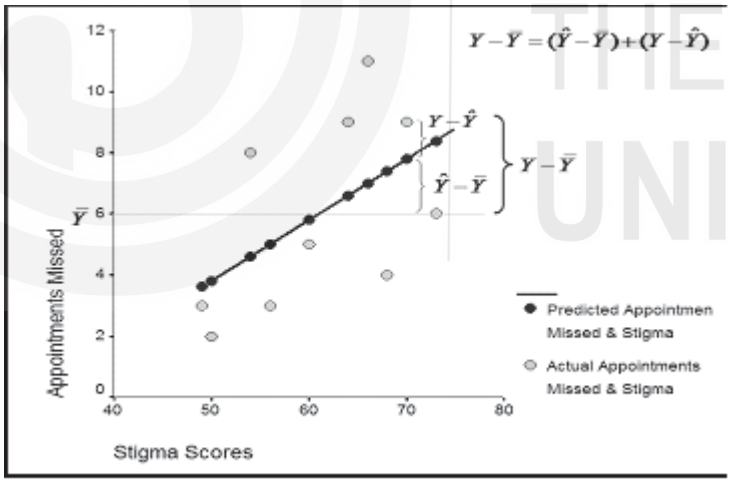


Fig. 4.3: The figure showing the scatter of X and Y, the regression line, and also explains the variance explained, residual and total.

4.2.6 Significance Testing of b

The *F*-distribution is employed to test the significance of the *b*.

The slope or regression coefficient obtained on sample is an estimator of the population slope or population regression coefficient called as â.

We already have completed the basics for computing the *F*-distribution.

$$F = \frac{S^2_{Between}}{S^2_{Within}} \qquad \text{(eq. 4.25)}$$

In case of regression, the same formula is used. The sum of squares total, sum of squares regression, and sum of squares residual have already been computed. We will use them now. Look at the table below.

Table 4.5: Table showing the computation of significance of *b*.

Source	Sum of Squares	*df*	S^2	F
Regression	$\sum(\hat{Y}-\bar{Y})^2$	k	$\frac{\sum(\hat{Y}-\bar{Y})^2}{k}$	$\frac{S^2_{Regresion}}{S^2_{Residual}}$
Residual	$\sum(Y-\hat{Y})^2$	$n-k-1$	$\frac{\sum(Y-\hat{Y})^2}{n-k-1}$	
Total	$\sum(Y-\bar{Y})^2$	$n-1$		

Where, n = sample size, and k = number of independent variables.

The null and the alternative hypothesis tested are as follows:

The F is computed for our example.

Table 4.6: Table showing the computation of the *F*-statistics for the data.

Source	Sum of Squares	*df*	S^2	F
Regression	25.68	1	25.68	$\frac{25.68}{7.54}=3.41$
Residual	60.32	8	7.54	
Total	86	9		

The *F*-value needs to be tested for its significance. The *F*-value at numerator *df* = 1, and denominator *df* = 8 at 0.05 level is 5.31. The obtained value of the *F* is smaller than the tabled value of the F. This means that we need to accept the null hypothesis which states that the $\hat{a} = 0$.

This might look surprising for some of you. But one thing we need to understand is the fact that the sample size (*n*) for this example is very small. Given that small *n*, the ability to reject the false null hypothesis is not so good and that's the reason we are accepting this null hypothesis.

4.2.7 Accuracy of Prediction

The present example has not turned out to be significant. But we will continue to discuss the issues in regression. How accurate we are in predicting Y from X is one of the important issues. We will look at various measures that tell us about the accuracy of prediction. We will continue to use this example considering it as significant even when it is not.

Standard Error of Measurement:

The standard error of estimate provides us an estimate of the error in the estimation. It can be calculated as follows:

$$s_{Y.X} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{SS_{Residual}}{df}} \qquad \text{(eq. 4.26)}$$

The standard error in our example can be computed using the formula as follows:

$$s_{Y.X} = \sqrt{\frac{SS_{Residual}}{df}} = \sqrt{\frac{60.32}{8}} = 7.54$$

Percentage of Variance Explained: r^2

The r^2 can be used as a measure of amount of variance X explained in Y. This shows the proportion of variance explained from total variance. Look at the following equation.

$$r^2 = \frac{SS_{Regression}}{SS_{Total}} \qquad \text{(eq. 4.27)}$$

$$r^2 = \frac{\sum(\hat{Y} - \bar{Y})}{\sum(Y - \bar{Y})} \qquad \text{(eq. 4.28)}$$

$$r^2 = \frac{\sum(\hat{Y} - \bar{Y})}{\sum(Y - \bar{Y})} = \frac{25.68}{86} = 0.299$$

Which means that 29.9 percent variance in Y is explained by X. This 'explained variance' around 30 percent is a good amount of variance considering the unreliability of psychological variables.

Indeed, the square root of the r^2, will give us the correlation between the X and Y.

Proportional Improvement in Prediction

The Proportional Improvement in Prediction (PIP) is one of the measure of accuracy. It is calculated as follows:

$$\text{PIP} = 1 - \sqrt{(1 - r^2)} \qquad \text{(eq. 4.29)}$$

In case of our example,

$$\text{PIP} = 1 - \sqrt{(1 - r^2)} = 1 - \sqrt{(1 - .299} = 0.162$$

The PIP value for our example is 0.162. So the proportional improvement in prediction is .162.

4.2.8 Assumption Underlying Regression

Some of the important assumptions for doing the regression analysis are as follows:

i) Independence among the pairs of score.

   This assumption implies that the scores of any two observations (subjects in case of most of psychological data) are not influenced by each other. Each pair of observation is independent. This is assured when different subjects provides different pairs of observation.

ii) The variance of the error terms is constant for each value of X.

iii) The relationship between X and Y is linear.

iv) The error terms follow the normal distribution with a mean zero and variance one.

v) Independence of Error Terms. The error terms are independent. They are uncorrelated.

vi) The population of X and the population of Y follow normal distribution and the population pair of scores of X and Y has a normal bivariate distribution.

This assumption(vi) states that the population distribution of both the variables (X and Y) is normal. This also means that the pair of scores follows bivariate normal distribution. This assumption can be tested by using statistical tests for normality.

4.2.9 Interpretation of Regression

The linear regression analysis provides us with lot of information about the data. This information need to be carefully interpreted. The intercept (á), the slope (â), the r^2, the *F*-value, need to be interpreted. Let us take these one by one .

The Intercept (á)

The intercept of regression line is the point at which regression line passes through the y-axis. This point is called as *a* in the sample and á in the population.

One straightforward interpretation of the *a* is it is a regression constant. It is that value, which we need to add into *b*X in order to get the predicted value of Y.

The other way of understanding the intercept is, intercept of regression line is that value of Y when the X value is zero. This interpretation looks intuitive. The correctness of this interpretation depends on whether we have sufficient X values near zero. In our example, the X was Stigma scores. The lowest value of the stigma scores was 49. Obviously we do not have any scores of X that are near zero. So the interpretation is unwarranted. This is due to two reasons.

i) We have not taken the complete range of the X values since we are studying the group of patients.

ii) The real zero X value is almost near impossible and the value of intercept 6.144, which is a Y-value when X is zero, is defiantly not possible.

iii) Nobody would miss -6 appointments. The best is not a single appointment is missed. So this interpretation is not applicable.

Slope (â):

The slope parameter is most important part of regression. The slope is called as regression coefficient. This also has straightforward interpretation.

The slope is that change in the Y when X changes by one unit.

So *rate of change* interpretation is common interpretation of the slope.

In our example, the slope value is 0.1991. This means that if the score on Stigma scale increases by one unit, the number of appointments missed will change by a value of .20.

This would also mean that as the score increases by 5 scale points on the Stigma scale, the person is likely to miss one appointment.

The r^2.:

The r^2 is the value that gives us the percentage of the variance X explains in Y.

The value is .299 in our example.

This means that roughly 30 percent variance in Y can be explained by X.

This can also be understood as proportional reduction in error.

Since we obtain r^2 by dividing the $SS_{Total} - SS_{Error}$ by the $SS_{Total.}$

This tells us about how much of the error is reduced.

The F ratio

The *F*-statistics is computed to test a null hypothesis â = 0.

If the *F*-value is statistically significant then the null hypothesis â = 0 is rejected.

Otherwise one has to accept the null hypothesis. If the null hypothesis is accepted, then there is no need to do the rest of the statistics.

This clearly means that X cannot linearly predict the Y.

However, in our example, the sample size is too small.

So the power of the statistical test is also small.

4.3 STANDARDISED REGRESSION ANALYSIS AND STANDARDISED COEFFICIENTS

We have learned about doing the regression analysis with X and Y. In previous chapters we have also learned to calculate the Z-scores. The Z is a standard score of a variable. To remind you, the Z can be calculated for each of the variable with the formula given below:

$$Z = \frac{X - \bar{X}}{S}$$

The mean of the Z is zero and the standard deviation is one.

Now, instead of predicting Y from X, we calculate the Z scores for both X and Y.

They will be denoted as Z_X and Z_Y.

Now we carry out the regression on standard variables than on unstandardised variables.

The regression equation will be

$$Z_Y = a + bZ_X + e \qquad \text{(eq. 4.30)}$$

Now, the intercept term is completely redundant in this equation because when we take the standard variable (that is Z) then the Y-intercept of the regression line is by default becomes *zero*. so the equation reduces to

$$Z_Y = bZ_X + e \qquad \text{(eq. 4.31)}$$

The beta value obtained in this regression equation is quite interesting.

Let us recall the correlation coefficient.

The correlation coefficient $r = Cov_{XY} / S_X S_Y$.

Now, with both the variable being standardised, the S_X, S_Y and S_X^2, will all be equal to one.

The slope for regression is calculated as $b = Cov_{XY} / S_X^2$.

Now you will realise that the slope (b) is equal to the r, correlation coefficient.

4.4 MULTIPLE REGRESSION

When we have multiple predictors than a single predictor variable, the regression carried out is called as multiple regression.

So we have a dependent variable and a set of independent variables. Suppose we have X_1, X_2, X_3, … up to X_k as k independent variables, and Y as a dependent variable, then the regression equation for sample can be written as:

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_k K_k \qquad \text{(eq. 4.32)}$$

The same equation for the population can be written as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k K_k \qquad \text{(eq. 4.33)}$$

Look at the following data. The data is about three variables, number of appointments missed, stigma scores, and the distance between the hospital and home.

Generally, one would expect that if the stigma is high, then the appointments would be missed. Similarly if the hospital is far away, then the appointments may be missed.

Table 4.7: Table of the data for appointments missed, stigma scores and distance of the hospital from home for 10 patients

Appointments Missed (Y)	Stigma Scores (X_1)	Distance of the hospital (X_2)
2	40	2
3	43	5
4	45	4
5	46	7
6	60	9
7	63	5
8	69	2
9	54	8
11	70	6
11	62	9

The equation for which we carry out the regression analysis is as follows:

$$\text{Appointmnets Missed } (Y) = a + b_1 \text{ Stigma Score} + b_2 \text{ Distance from Home} + e$$

We will solve the numerical for this problem. I shall directly provide you with the answer.

The Multiple R^2 for this problem is 0.81. which means that 81 percent information in appointments missed is explained by these two variables.

The adjusted value for the same is .76.

The value of intercept is –7.88.

The slope for stigma is 0.22 and

The slope for distance is 0.40.

The results of significance testing are as follows:

Table 4.8: Table showing the significance testing and the ANOVA summary

Source	Sum of Squares	*Df*	Mean Square	F	Sig.
Regression	73.279	2	36.640	14.981	.003
Residual	17.121	7	2.446		
Total	90.400	9			

The obtained F-value tells us that the overall model we have tested for is turning out to be significant. We can actually test the significance of each of the b separately. When that is done, the b of stigma turned out to be significant ($t = 4.61$, $p < .01$) but the distance did not ($t = 1.93$, $p > .05$).

Here too the size of the sample appears to be the problem leading to non significant results.

The multiple regression equation can be solved hierarchically or directly.

When the equation is solved directly, all the predictors are entered into the equation simultaneously.

When the equation is solved hierarchically, then the predictors are entered one after another depending on the theory or simply depending on their statistical ability to predict the Y.

The multiple regression is very useful technique in psychological research.

4.5 LET US SUM UP

Now we know how to solve the problem of prediction in psychological research. We can develop suitable regression equation and test it against the data. We can test the predictability, amount of information in dependent explained by independent, etc. this technique is very informative.

When we do regression, it does not mean that the causality appears in the equation. It is not a function of statistics. It has to come from theory.

We now know how to set up a multiple regression equation. Though we do not know how to do the calculations, we can understand the results of multiple regression.

4.6 UNIT END QUESTIONS

Given below are some problems with Answers

1) A researcher was interested in predicting marks obtained in the first year of the college form the marks obtained in the high school. He collected data of 15 individuals which is given below. Find out the Independent Variable and Dependent Variable.

   Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

School marks	College marks
67	65
45	50
65	60
60	71
55	54
53	49
59	58
64	69
67	75
69	73
70	64
58	66
63	62
71	65
74	78

2) A researcher was interested in predicting general satisfaction of people form perceived social support. She collected data of 10 individuals which is given below. Find out the IV and DV, Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

Satisfaction with Life	Perceived Social Support
7	7
6	6
5	6
8	3
9	6
7	4
6	4
3	2
11	9
8	5

3) A researcher was interested in predicting stage performance form social anxiety. She collected data of 10 individuals which is given below. Find out the IV and DV, Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

Stage Performance	Social Anxiety
9	11
7	9
6	11
10	7
10	11
9	9
9	8
5	7
14	13
10	9

4) A researcher was interested in predicting attitude to working condition form affective commitment to job. She collected data of 12 individuals which is given below. Find out the IV and DV, Write regression equation, calculate *a* and *b*, plot the scatter and straight line, write null and alternative hypothesis, determine significance, and comment on the accuracy of the prediction.

Attitude to Work	Affective Commitment
5	10
7	13
4	8
5	9
7	14
9	16
3	10
2	6
8	16
7	13
6	9
9	8

Answers:

1) $r = .78$, $r^2 = .608$, $a = 9.27$, $b = .87$, $SS_{Regression} = 641.75$, $SS_{Residual} = 413.19$, $F = 20.19$.

2) $r = .64$, $r^2 = .41$, $a = 3.41$, $b = .69$, $SS_{Regression} = 17.98$, $SS_{Residual} = 26.02$, $F = 5.53$.

3) $r = .51$, $r^2 = .26$, $a = 2.7$, $b = .65$, $SS_{Regression} = 14.67$, $SS_{Residual} = 42.22$, $F = 2.78$.

4) $r = .67$, $r^2 = .45$, $a = .958$, $b = .458$, $SS_{Regression} = 25.21$, $SS_{Residual} = 30.79$, $F = 8.19$.

4.7 SUGGESTED READINGS

Garrett, H.E. (19). *Statistics In Psychology And Education*. Goyal Publishing House, New Delhi.

Guilford, J.P.(1956). *Fundamental Statistics in Psychology and Education.* McGraw Hill Book company Inc. New York.