1. Introduction

Welcome to the series of e-learning modules on Coefficient of Determination. In this module, we will see how to do the inference of relation between the variables using coefficient of determination and probable error, use of coefficient of determination in regression analysis, adjusted and generalised r square.

By the end of this session, you will be able to:

- Explain the probable error and coefficient of determination, which are used to interpret the coefficient of correlation found by product moment method
- Explain its uses in interpreting r and finding the best fit for the given data
- Explain the adjusted and generalised r square

In the previous module, we have learnt about the product moment correlation coefficient, which is used to measure the degree of correlation between the two variables. It also gives the direction of correlation between the two variables, which is positive or negative.
Once we find the correlation coefficient using the product moment method, then we need to interpret it. Mainly, we use the following two tools to interpret the correlation coefficient:

1. Probable error
2. Coefficient of determination

First, let us study about the probable error.

After the calculation of coefficient of correlation, the next thing is to find out the extent to which it is dependable. For this purpose, the probable error of the coefficient of correlation is calculated. If the probable error is added to and subtracted from the coefficient of correlation, it would give two such limits within which we can reasonably expect the value of coefficient of correlation to vary.

It means that if another set of samples from the same universe were selected based on random sampling, the coefficient between the two variables in this new sample would not fall outside the limits so established.

The formula for finding the probable error of the product moment correlation coefficient is, probable error is equal to zero point 6 seven 4 five into 1 minus r square divided by square root of n.
Where,
r is the coefficient of correlation
n is the number of pairs of observation

- If the value of r is less than the probable error, then there is no evidence for correlation.
- If the value of r is more than six times of the probable error, then it is a significant correlation.

- It is a measure of interpreting coefficient of correlation, which should be used only when the number of pairs of observations is large. If n is small, probable error may give misleading conclusions.

2. Need of Coefficient of Determination – Part 1

Now let us see why we need coefficient of determination.
It as an important and useful method of interpreting coefficient of correlation, which is the square of coefficient of correlation or r square.
Coefficient of Determination is equal to coefficient of correlation square.

The coefficient of determination, r square, is useful because it gives the proportion of the variance (or fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph. The coefficient of determination is such that 0 less than or equal to r square less than or equal to 1, and denotes the strength of the linear association between x and y.

However, if we calculate the Coefficient of Determination r square, which in this case would be plus zero point 8 square or plus zero point 6 four. It means 64 percent of the variations in supply are because of the coefficient of determination. Thus we can say that
Coefficient of determination is equal to explained variance divided by total variance.
In the above example, the total variance is unity and the explained variance is zero point 6 four.

The complement of coefficient of determination is known as coefficient of non-determination.
We know that total variance is equal to explained variance plus unexplained variance.
Sometimes a Coefficient of Correlation is interpreted by finding out the Unexplained Variance.
The ratio of unexplained variance to total variance is called the Coefficient of non-determination.

And is given by,
K square is equal to coefficient of non-determination is equal to unexplained variance divided by total variance is equal to 1 minus r square.
The square root of k square is called the coefficient of alienation and represented by K and is given by square root of 1 minus r square.

3. Need of Coefficient of Determination – Part 2

In the above example, the unexplained variance is zero point 3 six, which is 1 minus zero point 6 four. This indicates the extent to which the factors other than the independent variable are affecting the dependent variable. Thus 64 percent of the changes in supply are due to changes in price and 36 percent are due to other factors. Generally, coefficient of determination is used to interpret correlation.

The following table gives some values whose interpretation is also given below the table.

Table 1

r	r2	r	r2
0.3	0.09	0.7	0.49
0.4	0.16	0.8	0.64
0.5	0.25	0.9	0.81
0.6	0.36	1	1

The table gives the values of r corresponding r square values, that is zero point 3 square is equal to zero point zero 9, zero point 4 square is equal to zero point 1 six, zero point 5 square is equal to zero point 2 five and so on.
Observe that, the above table shows that r square is always less than r unless r happens to be unity in which case r and r square are equal. It should be observed that if r is equal to zero point four in two variables and zero point eight in the other two variables it does not mean that correlation in the second set is twice as strong as it is in the first.

In these two sets, the value of r^2 would be 0.16 and 0.64 respectively, which shows that the relationship in the second set is four times as strong as it is in the first set.
In the second set, 64% of the total variables are explained while in the first set, only 16% of the total variations are due to changes in the independent variable.

The coefficient of determination (denoted by r square) is a key output of regression analysis. The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variations. Further, the line is away from the points, the less it is able to explain.

Every sample has some variation in it (unless all the values are identical, and that is unlikely to happen). The total variation is made up of two parts, the part that can be explained by the

regression equation and the part that cannot be explained by the regression equation.

That is summation y minus y bar whole square is equal to summation y dash minus y bar whole square plus summation y minus y dash whole square, which is same as total variance is equal to explained variance plus unexplained variance.

- The coefficient of determination r square is one of the measures of how well the least squares equation Y is equal to a plus b into x performs as a predictor of y.
- The higher the r square the more useful the model is.
- r square takes values between 0 and 1.
- Essentially, r square tells us how much better we can do in predicting y by using the model and computing y cap than by just using the mean y bar as a predictor.
- Note that, when we use the model and compute y cap the prediction depends on x because, y cap is equal to a plus b into x. Thus, we act as if x contains information about y
- If we just use y bar to predict y, then we are saying that x does not contribute information about y and thus our predictions of y do not depend on x.

Hence, we can define r square as follows:
The coefficient of determination r square is a measure of the proportion of variance of a predicted outcome. With a value of 0 to 1, the coefficient of determination is calculated as the square of the correlation coefficient r between the sample and predicted data.
The coefficient of determination shows how well a regression model fits the data. Its value represents the percentage of variation that can be explained by the regression equation.

A value of 1 means every point on the regression line fits the data that is the dependent variable can be predicted without error from the independent variable.
A value of zero point 5 means only half of the variation is explained by the regression.

Once we find the coefficient of determination, let us see how to interpret it.
In regression, the *r square* coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An *r square* of 1.0 indicates that the regression line perfectly fits the data.

Values of r^2 outside the range 0 to 1 can occur, where it is used to measure the agreement between observed and modelled values and where the "modelled" values are not obtained by linear regression and depending on which formulation of *r square* is used.

4. Concept of Adjusted Coefficient of Determination

Now, let us discuss the concept of adjusted coefficient of determination that is adjusted r square.
In many (but not all) instances where *r* square is used, the predictors are calculated by ordinary least squares regression that is by minimizing *sum of squares of error.* In this case, r-squared increases as we increase the number of variables in the model (*r square* will not decrease).

This illustrates a drawback to one possible use of *r square*, where one might try to include more variables in the model until "there is no more improvement". This leads to the alternative approach of looking at the adjusted *r square*. The explanation of this statistic is almost the same as *r square* but it penalizes the statistic, as extra variables are included in the model.

For cases other than fitting by ordinary least squares, the r^2 statistic can be calculated as above and may still be a useful measure.
If fitting is by weighted least squares or generalized least square, alternative versions of r^2 can be calculated appropriate to those statistical frameworks, while the "raw" r^2 may still be useful if it is more easily interpreted.
Values for r^2 can be calculated for any type of predictive model, which need not have a statistical basis.

Now, let us discuss the generalized r square.
Nagelkerke generalizes the definition of the coefficient of determination as:

- A generalized coefficient of determination should be consistent with the classical coefficient of determination when both can be computed
- Its value should also be maximized by the maximum likelihood estimation of a model
- It should be at least asymptotically, independent of the sample size
- Its interpretation should be the proportion of the variation explained by the model
- It should be between 0 and 1, with 0 denoting that model does not explain any variation and 1 denoting that it perfectly explains the observed variation
- It should not have any unit

r square does not indicate whether:

- The independent variables are a true cause of the changes in the dependent variable
- Omitted-variable bias exists
- The correct regression was used
- The most appropriate set of independent variables has been chosen
- There is co-linearity present in the data on the explanatory variables
- The model might be improved by using transformed versions of the existing set of independent variables

5. Illustrations

Now, let us consider the following illustrations.
Illustration 1:
From the data given below, calculate product moment correlation coefficient between heights of fathers and their sons.
The table gives the heights of seven fathers and their sons in inches.

Table 2

Height of fathers (inches)	Height of sons (inches)
70	72
71	69
65	67
66	68
67	66
68	69
69	72

Since given data is raw data, we know that the formula for calculating the product moment coefficient of correlation is,
r X Y is equal to n into summation x i into y i minus summation x i into summation y i whole divided by square root of n into summation x i square minus summation x i whole square into n into summation y i square minus summation y i whole square.

Hence, we find the following table to find the sum of required variables, where we take x i as the weights of fathers and y i as the ages of their sons.

Table 3

x	y	x2	y2	xy
70	72	4900	5184	5040
71	69	5041	4761	4899
65	67	4225	4489	4355
66	68	4356	4624	4488
67	66	4489	4356	4422
68	69	4624	4761	4692
69	72	4761	5184	4968
476	483	32396	33359	32864

The first and second columns in the table are written as it is.
The third column is obtained by squaring the numbers of the 1st column. That is 70 square is equal to 4 thousand 900, 71 square is equal to 5 thousand 41 and so on.
The 4th column is obtained by squaring the numbers of the 2nd column. That is, 72 square is equal to 5 thousand 184, 69 square is equal to 4 thousand 761 and so on.
The last column is obtained by multiplying the numbers of 1st and 2nd column. That is 70 into 72 is equal to 5 thousand 40, 71 into 69 is equal to 4 thousand 899 and so on. Once we find all the values in the different columns, we find the totals of all the columns, which are written in bold numbers in the column.

Therefore, r X Y is equal to 7 into 32 thousand 864 minus 476 into 483 whole divided by square root of 7 into 32 thousand 396 minus 476 square into 7 into 33 thousand 359 minus 483 square
Is equal to zero point 6 six eight.
To interpret the data, we use coefficient of determination.
r square is equal to zero point 6 six 8 square is equal to zero point 4 four 6 two.
Hence, only 44 point 62 percent of variation in variable Y is explained by x. Therefore, there is moderate positive relation between the height of father and son.

Find out Karl Pearson's correlation coefficient between age and playing habit of the following students.
The table gives the ages, number of students in that age group and number of players out of those.

Table 4

Age (in years)	No. of Students	Regular Players
15	250	200
16	200	150
17	150	100
18	120	48
19	100	30
20	80	12

Since given data is raw data, we know that the formula for calculating the product moment coefficient of correlation is,
r X Y is equal to n into summation x i into y i minus summation x i into summation y i whole divided by square root of n into summation x i square minus summation x i whole square into n into summation y i square minus summation y i whole square.

Hence, we find the following table to find the sum of required variables, where we take x as

the ages in year of the players. Since we have given number of students in particular age group and total number of students in that age group, we find number of players in each age group per same number of persons. Hence, we divide number of players by number of students and multiply by 100, which give the number of players per 100 students and is denoted by y.

Table 5

Age (x) (in years)	players per 100 (y)	x2	y2	xy
15	80	225	6400	1200
16	75	256	5625	1200
17	67	289	4489	1139
18	40	324	1600	720
19	30	361	900	570
20	15	400	225	300
105	307	1855	19239	5129

The 1st column, which gives ages of students, is written as it is in the problem, which is denoted by x.
The 2nd column is obtained as explained above. That is number of players per 100 players which is denoted by y. That is, 200 divided by 250 into 100 is equal to 80, 150 divided by 200 into 100 is equal to 75, 100 divided by 150 into 100 is equal to 66 point 6 seven, which is rounded off to 67 and so on.
3rd column is obtained by squaring the numbers of the first column. That is 15 square is equal to 225, 16 square is equal to 256 and so on.
The 4th column is obtained by squaring the numbers of the 2nd column. That is 80 square is equal to 6 thousand 400, 75 square is equal to 5 thousand 625 and so on. The last column is obtained by multiplying the numbers of the 1st and 2nd columns. That is 15 into 80 is equal to 1 thousand 200, 16 into 75 is equal to 1 thousand 200 and so on. Once we find all the elements in the different columns, we find the grand totals of each column, which are written in bold numbers in the column.

Now, let us substitute the different values in the formula. Here we have 6 pairs or numbers, hence n is equal to 6.
Therefore, r X Y is equal to 6 into 5 thousand 129 minus 105 into 307 divided by square root of 6 into 1 thousand 855 minus 105 square into
6 into 19 thousand 289 minus 307 square is equal to minus zero point 9 seven 9.
Again since n is small, we use coefficient of determination to interpret the data.
That is r square is equal to minus zero point 9 seven 9 square is equal to zero point 9 five 8 4.
That is there is is high negative correlation between age of the students and the playing habits of the students. That is they grow older, they play less or their playing habit reduces.

Here's a summary of our learning in this session, where we have understood:

- The probable error and coefficient of determination, which are used to interpret the coefficient of correlation found by product moment method
- The uses in interpreting r and finding the best fit for the given data
- The adjusted and generalised r square