Coefficient of Determination and Correlation

Coefficient of Determination

The coefficient of determination can be thought of as a percent. It gives you an idea of how many data points fall within the results of the line formed by the regression equation. The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted. If the coefficient is 0.80, then 80% of the points should fall within the regression line. Values of 1 or 0 would indicate the regression line represents all or none of the data, respectively. A higher coefficient is an indicator of a better goodness of fit for the observations.

The coefficient of determination (denoted by R2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
- With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
- An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R2 of 0.20 means that 20 percent is predictable; and so on.

The value of R^2 shows whether the model would be a good fit for the given data set. On the context of analysis, for any given per cent of the variation, it(good fit) would be different. For instance, in a few fields like rocket science, R^2 is expected to be nearer to 100 %. But R^2 =

0(minimum theoretical value), which might not be true as R^2 is always greater than 0(by Linear Regression).

The value of R^2 increases after adding a new variable predictor. Note that it might not be associated with the result or outcome. The R^2 which was adjusted will include the same information as the original one. The number of predictor variables in the model gets penalized. When in a multiple linear regression model, new predictors are added, it would increase R^2. Only an increase in R^2 which is greater than the expected (chance alone), will increase the adjusted R^2.

Coefficient of Determination Formula

The formula of correlation coefficient is given below:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

Where,
r = Correlation coefficient
x = Values in first set of data
y = Values in second set of data
n = Total number of values.

The procedure of finding the Coefficient of Determination

Step 1: Find the correlation coefficient, r (it may be given to you in the question). Example, r = **0.543**.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2\,][\, n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Step 2: Square the correlation coefficient.

0.543^2 = **.295**

Step 3: Convert the correlation coefficient to a percentage.

.295 = **29.5%**

Properties of Coefficient of Determination

- It helps to get the ratio of how a variable which can be predicted from the other one, varies.
- If we want to check how clear it is to make predictions from the data given, we can determine the same by this measurement.
- It helps to find Explained variation / Total Variation
- It also lets us know the strength of the association(linear) between the variables.
- If the value of r^2 gets close to 1, The values of y become close to the regression line and similarly if it goes close to 0, the values get away from the regression line.
- It helps in determining the strength of association between different variables.

Merits

The following are the chief points of merit that go in favor of the Karl Pearson's method of correlation:

- This method not only indicates the presence, or absence of correlation between any two variables but also, determines the exact extent, or degree to which they are correlated.
- Under this method, we can also ascertain the direction of the correlation i.e. whether the correlation between the two variables is positive, or negative.
- This method enables us in estimating the value of a dependent variable with reference to a particular value of an independent variable through regression equations.

- This method has a lot of algebraic properties for which the calculation of co-efficient of correlation, and a host of other related factors viz. co-efficient of determination, are made easy.

Demerits

Despite the above points of merits, this method also suffers from the following demerits:

- It is comparatively difficult to calculate as its computation involves intricate algebraic methods of calculations.
- It is very much affected by the values of the extreme items.
- It is based on a large number of assumptions viz. linear relationship, cause and effect relationship etc. which may not always hold good.
- It is very much likely to be misinterpreted particularly in case of homogeneous data.
- In comparison to the other methods, it takes much time to arrive at the results.
- It is subject to probable error which its propounded himself admits, and therefore, it is always advisable to compute it probable error while interpreting its results.

Coefficient of Correlation

The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no linear relationship between the movements of the two variables.

Correlation statistics can be used in finance and investing. For example, a correlation coefficient could be calculated to determine the level of correlation between the price of crude oil and the stock price of an oil-producing company, such as Exxon Mobil Corporation. Since oil companies earn greater profits as oil prices rise, the correlation between the two variables is highly positive.

- The coefficient of determination, r 2, is useful because it gives the proportion of the variance (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 the ratio of the explained variation to the total variation.
- The coefficient of determination is such that $0 < r\ 2 < 1$, and denotes the strength of the linear association between x and y.
- The coefficient of determination represents the percent of the data that is the closest to the line of best fit. For example, if r = 0.922, then r 2 = 0.850, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.
- 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 variation. The further the line is away from the points, the less it is able to explain.