**Case study on correlation & regression**

The data, I am working with contains details from the early days of the Italian clothing company Benetton. Each row in the table shows Benetton’s sales for a year and the amount spent on advertising that year. In this case, our outcome of interest is sales—it is what we want to predict.

For that, first we need to categorise the variables as independent variable and dependent variable.

The variable with values that should be predicted is the dependent variable. Here, it is the Sales, and we take those values as Yi.

The variables with values that are used for predicting the outcome are independent variables. Here, we use Advertising Amount to predict the Sales. So, Advertising Amount is the independent variable and we take it as Xi.

Questions

**1. First, the manager of Benetton wants to check whether there is any relationship between the Sales figures and the amount spent on Advertising. So, how he should check that?**

He should check the relationship of the variables under consideration by calculating the correlation coefficient. Only if there is a high degree of correlation, we can predict the future Sales based on the Advertising Amount.

In Excel, it can be done using CORREL function, which basically calculates the Karl Pearson’s correlation coefficient, which is obtained as 0.988, indicating very high positive correlation. This means, as Advertising Amount increases, Sales also increases.

**2. The manager of the clothing company wants to test whether his model is predicting the exact or nearby figures to the sales figures already given? What will he do?**

For Prediction, we need to build a prediction model.

By using Linear Regression Model, in Least Square Error Method,

The regression equation is **= b0+b1Xi**

Where, b1 =

b0 =

are the expected values of Yi. So calculating for the available values of Xi and comparing these expected values with the actual value, the manager can know whether his model is predicting the nearby or exact figures of the actual value.

**3. What will be the squared error between the actual values for Sales and predicted values for Sales?**

The Sum of the squared error between the actual values for Sales and predicted values for Sales is denoted as SSE. The main agenda of the model is to reduce this SSE.

Here, SSE = = 18803 Million Euros. (approx.)

Whereas, SST = = 807485 Million Euros. (approx.)

And, SSR = = SST – SSE = 788681 Million Euros. (approx.)

This means the total sum of squares of the error before applying the model is SST. By applying the regression model, we have reduced the SST to SSE. The error reduced by the regression model is SSR.

**4. If the advertising was done for an amount of 60 million Euros for the 10th year, what will be the Sales figure for the 10th year?**

Substituting the new Advertising Amount value (**60 million Euros**) in the regression equation, **= b0+b1Xi**, we get the predicted Sales Amount as around 1573 Million Euros.

He can also get an inference about the accuracy of his model by calculating the R square. Here, R square is 0.9767, which means 97.67% of the variation in the Sales is explained by the Advertising Amount. That is the model is pretty much accurate in prediction. So, Higher the R square value, Better the model we build.