04

Regression Analysis

Somy Kuriakose

Introduction

Correlation gives us the idea of the measure of magnitude and direction between correlated variables. Now it is natural to think of a method that helps us in estimating the value of one variable when the other is known. The fact that the variables x and y are correlated does not necessarily mean that x causes y or vice versa. Regression analysis is a statistical tool for the investigation of relationships between variables. It is a powerful technique used for predicting the unknown value of a variable from the known value of another variable. When there is only one independent variable then the relationship is expressed by a straight line. This procedure is called simple linear regression. More precisely, if X and Y are two related variables, then linear regression analysis helps us to predict the value of Y for a given value of X or vice verse. Multiple regression is an extension of bivariate regression in which several independent variables are combined to predict the dependent variable. In multilple regression analysis,the value of Y is predicted for given values of X_1, X_2, …, X_k.

Dependent and Independent Variables

By simple linear regression, we mean models with just one independent and one dependent variable. The variable whose value is to be predicted is known as the dependent variable and the one whose known value is used for prediction is known as the independent variable. Similarly for Multiple Regression the variable whose value is to be predicted is known as the dependent variable and the ones whose known values are used for prediction are known independent (exploratory) variables.

The Regression Model

The line of regression of Y on X is given by Y = a + bX where a and b are unknown constants known as intercept and slope of the equation. This is used to predict the unknown value of variable Y when value of variable X is known. The Simple Regression model is

$$\boldsymbol{Y = a + bX}$$

The coefficient of X in the line of regression of Y on X is called the regression coefficient of Y on X. It represents change in the value of dependent variable (Y) corresponding to unit change in the value of independent variable (X).

In general, the multiple regression equation of Y on X1, X2, ..., Xk is given by:

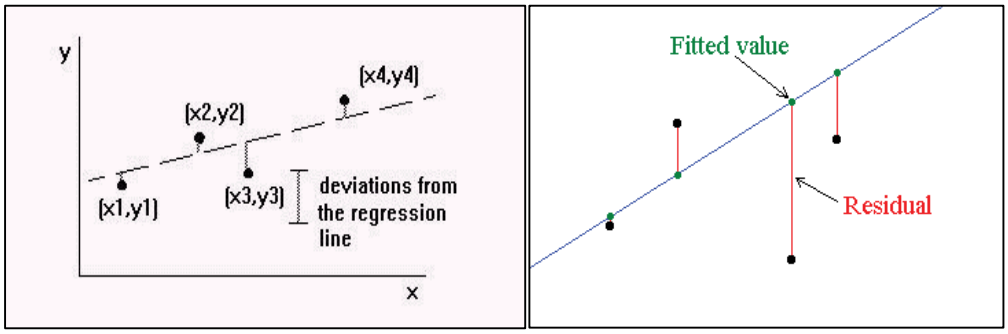
$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots\ldots\ldots + b_k X_k$$

Here b_0 is the intercept and b_1, b_2, b_3, ..., b_k are analogous to the slope in linear regression equation and are also called regression coefficients. They can be interpreted as the change in the value of dependent variable (Y) corresponding to unit change in the value of independent variable X_i.

Fitting of regression line

In scatter plot, we have seen that if the variables are highly correlated then the points (dots) lie in a narrow strip. If the strip is nearly straight, we can draw a straight line, such that all points are close to it from both sides. Such a line can be taken as an ideal representation of variation. This line is called the line of best fit if it minimizes the distances of all data points from it and also called as the line of regression. Now prediction is easy because all we need to do is to extend the line and read the value. Thus to obtain a line of regression, we need to have a line of best fit.

The problem of choosing the best straight line then comes down to finding the best values of a and b. By 'best' we mean the values of a and b that produce a line closest to all n observations. This means that we find the line that minimizes the distances of each observation to the line. Choose the a and b values that give the line such that the sum of squared deviations from the line is minimized. This method of estimation of parameters is called least square method. The best line is called the regression line, and the equation describing it is called the regression equation. The deviations from the line are also called residuals.



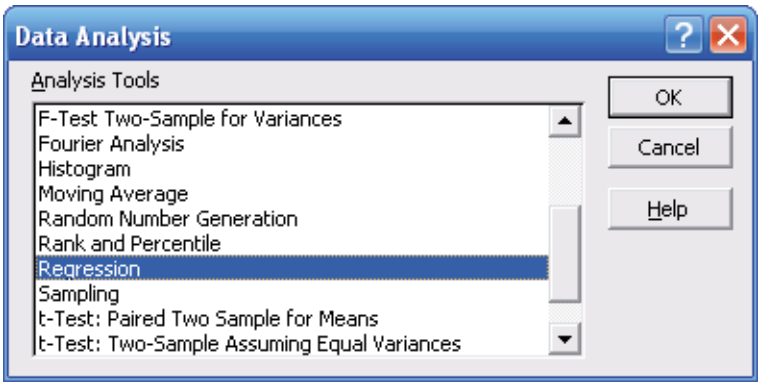
R^2 - coefficient of determination

Once a line of regression has been constructed, one can check how good it is (in terms of predictive ability) by examining the coefficient of determination (R^2), which is defined as the proportion of variance of the dependent variable that can be explained by the independent variables. The coefficient of determination is a measure of how well the regression equation y = a + bx performs as a predictor of y. R2 always lies between 0

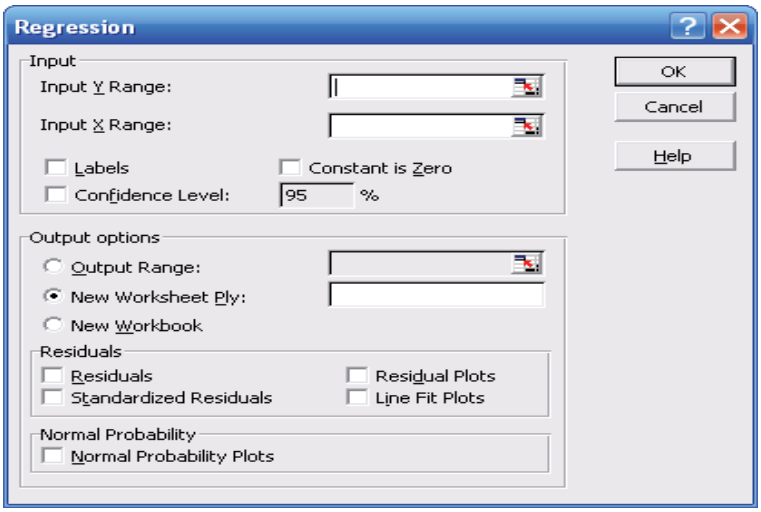
and 1. Higher values of this are generally taken to indicate a better model. The closer R2 is to 1, the better is the model and its prediction.

Regression Analysis using Microsoft Excel

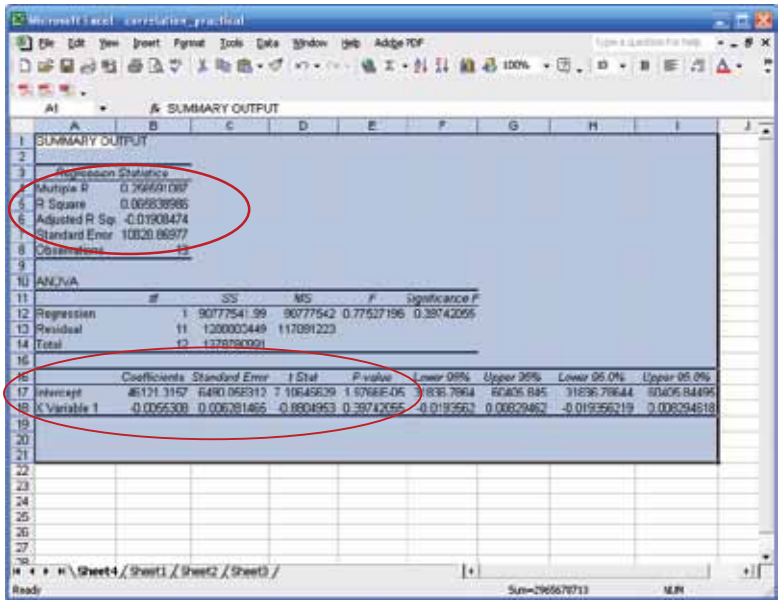
To do Regression in Microsoft Excel, choose **Data>Data Analysis>Regression**



- Enter the variable data, y as the dependent and x as the independent



- Check labels, if including column labels
- Check Residuals, Confidence levels to displayed them in the output
- The **SUMMARY OUTPUT** is displayed below



The coefficients of the regression line can be obtained from the summary output. The slope of the line (b) is the coefficient corresponding to *X variable1 and* a value is the coefficient corresponds to intercept. To check whether your results are statistically significant, look at Significance F. If this value is less than 0.05, the model is OK. If Significance F is greater than 0.05, it's probably better to stop using this set of independent variables. For testing the significance of the regression coefficients you can check the corresponding p-value.