Correlation and Regression Analysis

Dr. Salman Ahmad

Correlation

- Correlation is a statistical method that determines the degree of relationship between two different variables. It is also known as a “bivariate” statistic, with bi- meaning two and variate indicating variable or variance. The two variables are usually a pair of scores for a person or object. The relationship between any two variables can vary from strong to weak or none. When a relationship is strong, this means that knowing a person's or object’s score on one variable helps to predict their score on the second variable

Strong positive relationship

- Strong or high degree of relationship between the two variables. This also means that the higher the score of a participant on one variable, the higher the score will be on the other variable. Also, if a participant scores very low on one variable then their score will also be low on the other variable. For example, there is a positive correlation between years of education and wealth. Overall, the greater the number of years of education a person has, the greater their wealth. A strong correlation between these two variables also means the lower the number of years of education, the lower the wealth of that person. If the correlation was perfect one ($r = +1.00$), then there would be not a single exception in the entire sample to increasing years of education and increasing wealth. It would mean that there would be a perfect linear relationship between the two variables. However, perfect relationships do not exist between two variables in the real world of statistical sampling.

Strong negative relationship

- Strong negative relationship. This means that the higher the score of a person on one variable, the lower the score will be on the other variable. For example, there might be a strong negative relationship between the value of gold and the Dow Jones Industrial Average. In other words, when the value of gold is high, the stock market will be lower and when the stock market is doing well, the value of gold will be lower. A correlation coefficient that is close to r = 0.00 (note that the typical correlation coefficient is reported to two decimal places) means knowing a person's score on one variable tells you nothing about their score on the other variable.

Correlation coefficient = r

- A measure of the strength of a relationship between two continuous variables.

Its value is in between +1 and -1.

The value of “r” can not be +1 or -1.

The value of “r” may be +0.99 or -0.99.

How correlation analysis is performed in Excel

- See the video sent in our group

What is regression

- Regression analysis is a conceptually simple method for investigating functional relationships among variables. In simple word, it determines the dependence of dependent variable (in our case dependent variable is disease severity) over the independent variables (maximum and minimum temperatures, relative humidity, rainfall, wind speed and sunshine radiations, soil moisture etc.)
- Regression relationship is expressed in the form of an "equation" or a "model" connecting the response or dependent variable and one or more explanatory or predictor variables (these are independent variables).

- We denote the response variable (dependent variable) by Y and the set of predictor variables (independent variables) by Xl, X 2, . ", Xp , where p denotes the number of predictor variables.
- True relationship between Y and Xl, X2, ... ,Xp can be approximated by the regression model.
- Regression is of two types mainly.

A. Simple linear regression equation is:

$Y = a + bX$

Where “a” is called “intercept” or “constant” (this value comes through regression analysis)

“b” is called slope (this value comes through regression analysis)

“X” is called independent/predictor/regressor variables. We know the values of these like maximum temperature, minimum temperature etc.

Simple linear regression

- It is the relationship between dependent variable and single independent variable as clear from the equation mentioned in the previous slide.

Multiple regression

- It is the relationship between dependent variable and number of independent variables. Its equation is:

y = b 

It is clear from the equation that independent variables i.e. X1, x2 and Xn..are many so this know as multiple regression.

Coefficient of Determination (R^2)

- R^2 tells about the degree of dependence of dependent variable over independent variables. Its value is in between +1 and -1. Its value can not be +1 or -1 exact. Its maximum value may 0.99 and -0.99.

How to perform regression analysis using Excel

- See video attached.