# Correlation and Regression

Introduction

While dealing with data researchers very often have to come across situation where they need to handle multiple variables simultaneously. Under the circumstances researchers required to identify different features of the interrelationship between the variables under study such as existence of relationships, directions and extent of relationships, potential changes in the value of variables in response to the change of other variables and so on. To envelop such issues statisticians develop numerous numeric techniques namely correlation analysis and regression analysis.

Although these two methods are intimately related but conceptually they are quite different. In correlation analysis numerical measures is developed in order to identify the existence and quantify the extent of relationship between the variables understudy as well as to indicate the direction of the relationship. While in regression analysis emphasis is put on developing a mathematical equation of the line that best fits the variables understudy. Such equation is also used to estimate or predict the value of one variable given the value of another.

**Correlation Analysis**

The statistical tool with the help of which the relationship between two or more variables is studied is called correlation.

Whenever two variables are so related that a change in the value of one is accompanied by a change in the value of the other in such a way that

1. An increase in the one is accompanied by an increase or decrease in the other

Or

1. A decrease in the one is accompanied by a decrease or increase in the other

- Then the variables are said to be correlated.

A M Tuttle gives a very simple definition of correlation analysis as “An analysis of relation of two or more variables is usually called correlation”.

Thus correlation analysis can be defined as a group of techniques to measure the association between variables (two or more). The basic idea of correlation analysis is to measure the association between variables. In simplest case (when dealing with only two variables) the usual first step is to plot the data in a scatter diagram.

The variable that is being predicted or estimated is known as **dependent variable**. It is usually scaled on the Y – axis.

The variable used to estimate the value of the dependent variable is known as the **independent variable**. It is usually scaled on the X axis.

**Types of Scatter diagram:**

In analyzing the relationship between variables, it is often desirable to present the sample data in a diagram variously called scatter diagram, scatter gram or scatter plot. It gives us a visual impression of the relationship involved and suggests the type of model that may best fit the data.

Figure 1: Different forms of Scatter Diagram

The conventional procedure in constructing such a diagram is to have the independent variable X scaled on the horizontal axis and the dependent variable Y on the vertical axis. A point representing a pair of observations of X and Y is plotted, the resulting graph of all the points thus plotted for all the pairs of X and Y values in the sample is the scatter diagram. Such a typical diagram appears in figure 1.

**Example 1:**

In Bangladesh, Global Brand Pvt. Ltd., in one of the exclusive distributor of several internationally reputed electronic appliance such as ASUS, DELL, A4 Tech and so on. Ms. Moushumi Zahur was recently promoted to the position of national sales manager of the Global Brand Pvt. Ltd. at the upcoming sells meeting, the sales representatives from all over the country will be in attendance. She would like to impress upon them the importance of making that extra sells call each day. She decides to gather some information on the relationship between the numbers of ASUS Laptop sold. She selects a random number of 10 sales representatives and determines the number of sales calls and number of laptop sold as given in table 1. What observations can you make about the relationship between the number of sales calls and the number of laptop sold? Develop a scatter diagram to display the information.

**Solution:**

**Table 1: Number of Sales calls and Laptops sold for 10 sales people**

|  |  |  |
| --- | --- | --- |
| Sales representative | Number of sales calls | Number of laptops sold |
| Mr . Iftekhar kalam | 20 | 30 |
| Mr. Mahmudul hasan | 40 | 60 |
| Mr. Zakir Hossen | 20 | 40 |
| Mr. Maruf Ahmed | 30 | 60 |
| Ms. Lopamudra | 10 | 30 |
| Ms. Sharmina Hossain | 10 | 40 |
| Mr. Al Amin Kabir | 20 | 40 |
| Ms. Sharmina Hossain | 20 | 50 |
| Mr. Md. Lutfor Rahman | 20 | 30 |
| Ms. Farhana Islam | 30 | 70 |



***Made Easy – Drawing a scatter diagram***

In order to draw a scatter diagram we plot the independent variable – “number of sales call (X)” along horizontal axis and the dependent variable – “number of laptop sold (Y)” along the vertical axis. To plot the first pair of observation (X=20 and Y=30) move along the horizontal axis X=20, then go vertically to Y=30 and place a dot at the intersection. This process is continued until all the paired data are plotted as shown in the above figure.

**Uses of the study of correlation:**

The utility of the study of correlation is obvious from the following

1. Correlation analysis is used to determine the existence & extent of the relationship between two phenomena.
2. Sometimes correlation helps us to predict what will happen in future.
3. Once we know that two variables are closely related we can estimate the value of one variable given the value of another with the help of regression equations.

**The Coefficient of Correlation:**

Although a scatter plot is an essential and important first step in studying the association between two variables, it is often useful to quantify the strength of the association by calculation of a summary index. One such commonly used measure is the Pearson’s correlation coefficient. Originated by Karl Pearson about 1900, the coefficient of correlation describes the strength of the relationship between two sets of interval scaled or ratio scaled variables. Usually denoted by  following is the simple definition of correlation coefficient:

***Coefficient of correlation or correlation coefficient is a quantitative measure of the direction and strength of linear relationship between two numerically measured variables.***

If for two variables  be the n pairs of observations then the correlation coefficient between X and Y denoted by  is defined as



Where

,





An alternative form of the given formula is 

For computational convenience we will use the following form of correlation coefficient as given below:



**Assumptions of**

* The use of the correlation coefficient specifically assumes:
* Both variables are measured numerically.
* The two variables follow bivariate normal distribution.
* The relationship between the variables is linear.

**Some properties of**

The coefficient of correlation has some appealing properties. These appear as follows:

* The correlation coefficient is a symmetric measure. The symmetric property states that.
* The correlation coefficient will be positive or negative depending on whether the sign of the numerator of the formula is positive or negative.
* The correlation coefficient lies between –1 and +1. Symbolically.
* The correlation coefficient is a dimensionless quantity, implying that it is not expressed in any unit of measurement.
* The coefficient of correlation is independent of origin and scale of measurement.

**Interpretation of**



Figure 2: Scatter plots with different correlation patterns (adapted from Gujarati, 1988, p-69)

The following general rules would help in interpreting the value of:

1. A value of +1 indicates that X and Y are perfectly related in a positive linear sense. In this case, all the points in a scatter diagram lie on a straight line that has a positive slope (Figure 2a).
2. A value of –1 for indicates that X and Y are perfectly related in a negative linear sense. That is all the points lay on a straight line that has a negative slope (Figure 2b).
3. Values of lying between –1 and +1 indicate varying degrees of linear association. Values of close to 1 indicate a strong linear relationship with positive slope (Figure 2c). Positive values of close to 0 indicate a weak linear association with positive slope (Figure 2e). Values of close to –1 indicate a strong linear relationship with negative slope (Figure 2d) and negative values of close to 0 indicate a weak linear association with negative slope (Figure 2f).
4. =0 indicates that there is **no linear relationship** between the two variables.

The following drawing summarizes the strength and direction of the coefficient of correlation.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Perfectly negative linear relationship | | |  | | | | No linear relationship | |  | | | | Perfectly positive linear relationship | |
|  |  | | | | | | |  | | | | | |  |
|  | Strong negative correlation | | | Moderate negative correlation | | Weak negative correlation | | Weak positive correlation | | Moderate positive correlation | | Strong positive correlation | |  |
|  | | | |  | | |  | | |  | | |
| = -1.00 | | = -0.5 | | | | | = 0 | | = 0.5 | | | | = 1.00 | |
|  | Negative linear relationship | | | | | | | Positive linear relationship | | | | | |  |

**Example:**

The CEO of a Micro Finance Institution (MFI) believes there is a relationship between the number of client contacts by the field officers and reimbursement of loan by the clients. To document this assertion, he gathered the information indicating the number of client contacts and amount of loan reimbursement as given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Number of client contacts by the field officers | Amount of loan reimbursement by the clients (Tk. thousands) | Number of client contacts by the field officers | Amount of loan reimbursement by the clients (Tk. thousands) |
| 14 | 24 | 23 | 30 |
| 12 | 14 | 48 | 90 |
| 20 | 28 | 50 | 85 |
| 16 | 30 | 55 | 120 |
| 46 | 80 | 50 | 110 |

1. Draw a scatter diagram for the given data and comment on the approximate relationship between the variables under study.
2. Determine using Pearson’s coefficient of correlation, whether there exist any relationship between the number of client contacts by the field officers and the amount of loan reimbursement by the clients.

**Answer:**

1. 

* 1. For the calculation of correlation coefficient follow the steps below:

***Step 1*** - Complete the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y |  |  |  |
| 14 | 24 | 196 | 576 | 336 |
| 12 | 14 | 144 | 196 | 168 |
| 20 | 28 |  |  |  |
| 16 | 30 |  |  |  |
| 46 | 80 |  |  |  |
| 23 | 30 |  |  |  |
| 48 | 90 |  |  |  |
| 50 | 85 |  |  |  |
| 55 | 120 | 3025 | 14400 | 6600 |
| 50 | 110 | 2500 | 12100 | 5500 |
|  |  |  |  |  |

***Step 2*** – Compute the Pearson’s correlation coefficient using the following equation

For this particular problem (Apprx.)

***Step 3 –***

Comment on strength and direction of the relationship between the variable concerned based on the value of.

For this particular problem the comment would be –

*Since hence there exists strong positive linear relationship between number of client contacts by the field officers (X) and the amount of loan reimbursement by the clients (Y).*

**Exercise:**

The Following table gives indices of industrial production and number of registered unemployed people (in Lakh). Calculate the value of the correlation coefficient and comment.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 |
| Index of Production | 100 | 102 | 104 | 107 | 105 | 112 | 103 | 99 |
| Number unemployed | 154 | 12 | 13 | 11 | 12 | 12 | 19 | 26 |

***Answer:***

***J K Sharma, 457***

**Exercise:**

With the following data in 6 cities, determine and comment on the relation between the density of population and death rate:

|  |  |  |  |
| --- | --- | --- | --- |
| City | Area (in sq Km) | Population (in ‘000) | Number of deaths |
| A | 150 | 30 | 300 |
| B | 180 | 90 | 1440 |
| C | 100 | 40 | 560 |
| D | 60 | 42 | 840 |
| E | 120 | 72 | 1224 |
| F | 80 | 24 | 312 |
|  |  |  |  |

***Answer:***

***J K Sharma, 462***

**Exercise:**

The data given below shows the quantity of fertilizer (X) and yield (Y). For the given data plot the scatter diagram and calculate the sample coefficient of correlation.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 33 | 37 | 40 | 38 | 35 | 33 | 40 | 32 | 34 | 28 |
| Y | 38 | 32 | 33 | 34 | 30 | 26 | 29 | 31 | 24 | 24 |

**Solution:**

**Scatter plot**: Do yourself.

For the calculation of correlation coefficient follow the steps below:

We will use the equation 

Complete the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y |  |  |  |
| 33 | 38 |  |  |  |
| 37 | 32 |  |  |  |
| … | … | … | … | … |
| 28 | 24 |  |  |  |
|  |  |  |  |  |

Then compute: 

**Exercises on Correlation:**

**Problem 1:**

The following sample observations were randomly selected

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X: | 4 | 5 | 3 | 6 | 10 |
| Y: | 4 | 6 | 5 | 7 | 7 |

Determine the coefficient of correlation and the coefficient of determination. Interpret.

**Problem 2:**

The production department of ‘***CATS EYE***’ wants to explore the relationship between the number of employees who assemble polo shirt and the number produced. As an experiment, two employees were assigned to assemble polo shirt. They produced 15 during a one hour period. Then four employees assembled them and produced 25 during a one – hour period. The complete set of paired observations follows.

|  |  |
| --- | --- |
| Number of assemblers | One-Hour Production  (Units) |
| 2 | 15 |
| 4 | 25 |
| 1 | 10 |
| 5 | 40 |
| 3 | 30 |

The dependent variable is production; that is assumed that the level of production depends upon the number of employees.

1. Draw a scatter diagram.
2. Based on scatter diagram, does there appear to be any relationship between the number of assemblers and production? Explain.
3. Compute the coefficient of correlation and hence comment on the relationship of the concerned variables.
4. Evaluate the strength of the relationship by computing the determination.

**Problem 3:**

The following sample observations were randomly selected

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X: | 5 | 3 | 6 | 3 | 4 | 4 | 6 | 8 |
| Y: | -13 | 15 | 7 | 12 | 13 | 11 | 9 | 5 |

Determine the coefficient of correlation and the coefficient of determination. Interpret the association between X and Y.

**Problem 4:**

Dhaka Metropolitan Police (DMP) authority is considering increasing the number of police in an effort to reduce crime. Before making a final decision the DMP authority asks the officer in charge (OC) of different police stations to submit report on the number of police force in each of the police stations and the number of reported crimes in each of the selected area covered by different police station. The DMP collected the following sample data:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name of Police station | Number of police force | Number of crimes reported | Name of Police station | Number of police force | Number of crimes reported |
| Badda | 150 | 172 | Mohammadpur | 175 | 214 |
| Mugdapara | 177 | 135 | Dhanmondi | 125 | 65 |
| Khilgaon | 255 | 56 | Mirpur | 112 | 76 |
| Uttara | 278 | 78 | Hazaribahg | 224 | 85 |
| Shahbagh | 176 | 74 | Turag | 165 | 112 |

1. If we want to estimate crimes on the basis of the number of police, which variable is the dependent variable and which is the independent variable?
2. Draw a scatter diagram.
3. Determine the coefficient of correlation.
4. Determine the coefficient of determination.
5. Interpret these statistical measures.

**Problem 5:**

It is assumed that sales of major home appliances vary with the standing housing market: when new home sales are good, so are sales of dishwashers, washing machines, driers and refrigerators. A trade association compiled the following data (in thousand of units) on housing sales and major home appliances sales:

|  |  |  |  |
| --- | --- | --- | --- |
| Apartment Sales (thousand) | Appliance Sales (thousand) | Apartment Sales (thousand) | Appliance Sales (thousand) |
| 2.0 | 5.0 | 4.0 | 7.7 |
| 2.5 | 5.5 | 4.2 | 8.4 |
| 3.2 | 6.0 | 4.6 | 9.0 |
| 3.6 | 7.0 | 4.8 | 9.7 |
| 3.3 | 7.2 | 5.0 | 10.0 |

1. Draw a scatter diagram for the given data and comment on the approximate relationship between the variables under study.
2. Using Pearson’s correlation coefficient, determine whether there exists any relationship between the apartment sales and the amount of home appliance sales and also comment on the strength of the relationship.
3. Develop an equation for the relationship between appliances sales (in thousand) and apartment sales (in thousands) and comment.
4. Using coefficient of determination comment on how much variation in the home appliances sales is expressed by apartment sales.

**Regression Analysis:**

The regression analysis is a technique of studying the dependence of one variable (called dependent variable), on one or more variables (called explanatory variables), and with a view to estimating or predicting the average value of the dependent variable in terms of the known or fixed values of the independent variables.

The regression technique is primarily used to:

1. Estimate the relationship that exists, on the average, between the dependent variable and the explanatory variables.
2. Determine the effect of each of the explanatory variables on the dependent variable, controlling t he effects of all other explanatory variables.
3. Predict the value of the dependent variable for a given value of the explanatory variable.

**Regression model:**

In dealing with the concept of regression analysis we frequently use the term ‘model’. A model is simply a set of mathematical equations that describes the relationship between variable and a set of independent variables.

A model in its simplest from involving two variables may be of the type



Where

Dependent variable

Independent variable

Intercept of the regression line  measures the value of Y when the value of X assumes zero.

Regression coefficient of Y on X  measures Average change in Y as a result of per unit change in X

Error term

 &  are the parameters of the model.

**Estimating parameters regression line:**

One important objectives of regression analysis is to find estimates for  &  for a given model. There are several methods of estimating the parameters of a regression model. Such as:

1. Graphical method
2. Least square method

We will discuss here about the most commonly used method that is least square method for estimating the parameters of a regression model.

**Estimating the regression coefficient:**

We will designate the estimate of  by  and the required equation for estimating  is:



**Estimating the intercept **

We will designate the estimate of  by and the required equation for estimating  is:



**The fitted regression line thus:**



**Example:**

A departmental store has the following statistics of sales for a period of last one year of 10 salesmen, who have varying years of experience. Find the regression line of annual sales on year of experience. And hence predict the annual sales of persons who have 12 and 15 years of experience.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sales person | Years of experience | Annual sales (TK. 000s) | Sales persons | Years of experience | Annual sales (TK. 000s) |
| 1 | 1 | 80 | 6 | 8 | 111 |
| 2 | 3 | 97 | 7 | 10 | 119 |
| 3 | 4 | 92 | 8 | 10 | 123 |
| 4 | 4 | 102 | 9 | 11 | 117 |
| 5 | 6 | 103 | 10 | 13 | 136 |

Answer: For required computation complete the following table:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y |  |  |
| 1 | 80 |  |  |
| 3 | 97 |  |  |
| … | … | … | … |
| 11 | 117 |  |  |
| 13 | 136 |  |  |
|  |  | = | = |

Then compute

=



Hence the estimating equation is =

Then evaluate the value of  using the value of X =15 and X=12.

**Some other relevant calculation:**

* Calculation of standard error of estimate:



* Calculate standard error of estimate for the given example.
* Determine the standard error of b:



* Calculate standard error of b for the given example.

#### Assignment on Correlation and Regression

**Problem 1:**

For the following data set

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 13 | 16 | 14 | 11 | 17 | 9 | 13 | 17 | 18 | 12 |
| Y | 6.2 | 8.6 | 7.2 | 4.5 | 9.0 | 3.5 | 6.5 | 9.3 | 9.5. | 5.7 |

Answer the following

1. Plot the scatter diagram
2. Develop the estimating equation that best describe the data/ fit regression line of y on x / Fit regression line of y on x by LSM.
3. Predict Y for X= 10, 15, 20.
4. Calculate the standard error of estimate or estimate mean square error (MSE).
5. Determine the standard error of b.

**Problem 2:**

For the following data

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 33 | 38 | 24 | 61 | 52 | 45 | 65 | 82 | 29 | 79 | 50 | 63 |
| Y | 3 | 7 | 6 | 6 | 10 | 12 | 12 | 13 | 12 | 15 | 14 | 13 |

1. Plot the scatter diagram
2. Develop the estimating equation that best describe the data.
3. Calculate the sample coefficient of determination and sample coefficient of correlation.

**Problem 3:**

The CEO of a Micro Finance Institution (MFI) believes there is a relationship between the number of client contacts by the field officers and reimbursement of loan by the clients. To document this assertion, he gathered the information indicating the number of client contacts and amount of loan reimbursement as given below:

|  |  |  |  |
| --- | --- | --- | --- |
| Number of client contacts by the field officers | Amount of loan reimbursement by the clients (Tk. thousands) | Number of client contacts by the field officers | Amount of loan reimbursement by the clients (Tk. thousands) |
| 14 | 24 | 23 | 30 |
| 12 | 14 | 48 | 90 |
| 20 | 28 | 50 | 85 |
| 16 | 30 | 55 | 120 |
| 46 | 80 | 50 | 110 |

1. Draw a scatter diagram for the given data and comment on the approximate relationship between the variables under study.
2. Determine using Pearson’s coefficient of correlation, whether there exists any relationship between the number of client contacts by the field officers and the amount of loan reimbursement by the clients.
3. Determine the regression equation of loan reimbursement by the clients on the number of client contacts by the field officers.
4. Determine the estimated loan reimbursement if 40 and 75 client contacts have been made by the field officers.

**Problem 4:**

The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales (in tk. ‘000s)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Salesmen: | A | B | C | D | E | F | G | H | I |
| Test scores | 50 | 60 | 50 | 60 | 80 | 50 | 80 | 40 | 70 |
| Weekly sales | 30 | 60 | 40 | 50 | 60 | 30 | 70 | 50 | 60 |

1. Determine whether there is any relation between the test scores and weekly sales.
2. Obtain regression equation of sales on intelligence test scores of the salesmen
3. If the intelligence test scores of a salesman in 65, what would be his expected weekly sales?

***Answer:***

1. **and**

***J K Sharma, 490***

**Problem 5:**

Cost accounts often estimate overhead based on the level of production. At the standard limiting co. they have collected information on overhead expenses and units produced at different plants.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Overhead (Y) | 191 | 170 | 272 | 155 | 280 | 173 | 234 | 116 | 153 |
| Units (X) | 40 | 42 | 53 | 35 | 56 | 39 | 48 | 30 | 37 |

1. Compute the value of correlation co-efficient and hence comment.
2. Develop the regression equation for the cost accounts.
3. Predict the overhead when 50 units are produced.

**Problem 6:**

A study by the Atlanta, Georgia, and Department of Transportation on the effect of bus-ticket prices on the number of passengers produced the following results:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ticket prices (Cuts) (X) | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 |
| Passenger per 100 miles (Y) | 800 | 780 | 780 | 660 | 640 | 600 | 620 | 620 |

1. Compute the value of correlation co-efficient and hence comment.
2. Develop the regression equation that best describe the data.
3. Predict the number of passengers if the ticket price is 50 cents.

**Problem 7:**

A company is introducing a job evaluation scheme in which all jobs are graded by points for skill, responsibility and so on. Monthly pay scales (tk. in ‘000) are then drawn up according to the number of points allocated and other factors such as experience and local conditions. To date the company has applied this scheme to jobs:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Job: | A | B | C | D | E | F | G | H | I |
| Points | 5 | 25 | 7 | 19 | 10 | 12 | 15 | 28 | 16 |
| Pay (in ‘000 tk.) | 3.0 | 5.0 | 3.25 | 6.5 | 5.5 | 5.6 | 6.0 | 7.2 | 6.1 |

1. Find the least squares regression line for linking pay scales to points
2. Estimate the monthly pay for a job graded by 20 points.

***Answer:* and**

***J K Sharma, 492***