Correlation and Regression Analysis

Pembe Begul GUNER

Contents

Introduction

- **Correlation analysis: Examines between two or more variables the relationship.**
- **Regression analysis: Change one variable when a specific volume, examines how other variables that show a change.**

Correlation Analysis

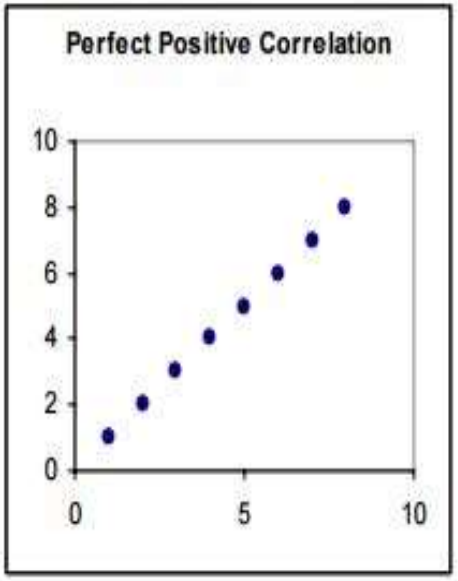
There are two important types of correlation.

(1) Positive and Negative Correlation

(2) Linear and Non – Linear Correlation

Positive and Negative Correlation

If the values of the two variables deviate in the same direction i.e. if an increase (or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.



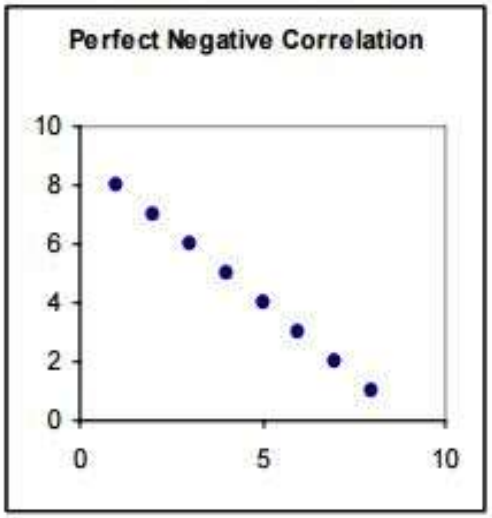
This graph illustrates perfect positive correlation. The two variables of interest are on the x and y axis, respectively. When graphed this way, it is apparent that a (positive) linear relationship exists between the two variables.

Some examples of series of positive correlation are:

- Heights and weights;
- Household income and expenditure;
- Price and supply of commodities;
- Amount of rainfall and yield of crops.

Negative Correlation

Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if the increase in the variables deviate in opposite direction. That is, if increase (or decrease) in the values of one variable results on an average, in corresponding decrease (or increase) in the values of other variable.



This graph illustrates perfect negative correlation. The two variables of interest are on the x and y axis, respectively. When graphed this way, it is apparent that a (negative) linear relationship exists between the two variables, i.e. the variables "move together".

Some examples of series of negative correlation are:

- Volume and pressure of perfect gas;
- Current and resistance [keeping the voltage constant] (R =V / I) ;
- Price and demand of goods.

Linear and Non – Linear Correlation

The correlation between two variables is said to be **linear** if the change of one unit in one variable result in the corresponding change in the other variable over the entire range of values.

For Example;

X	2	4	6	8	10
Y	7	13	19	25	31

Thus, for a unit change in the value of x, there is a constant change in the corresponding values of y and the above data can be expressed by the relation ;

- $y = 3x + 1$

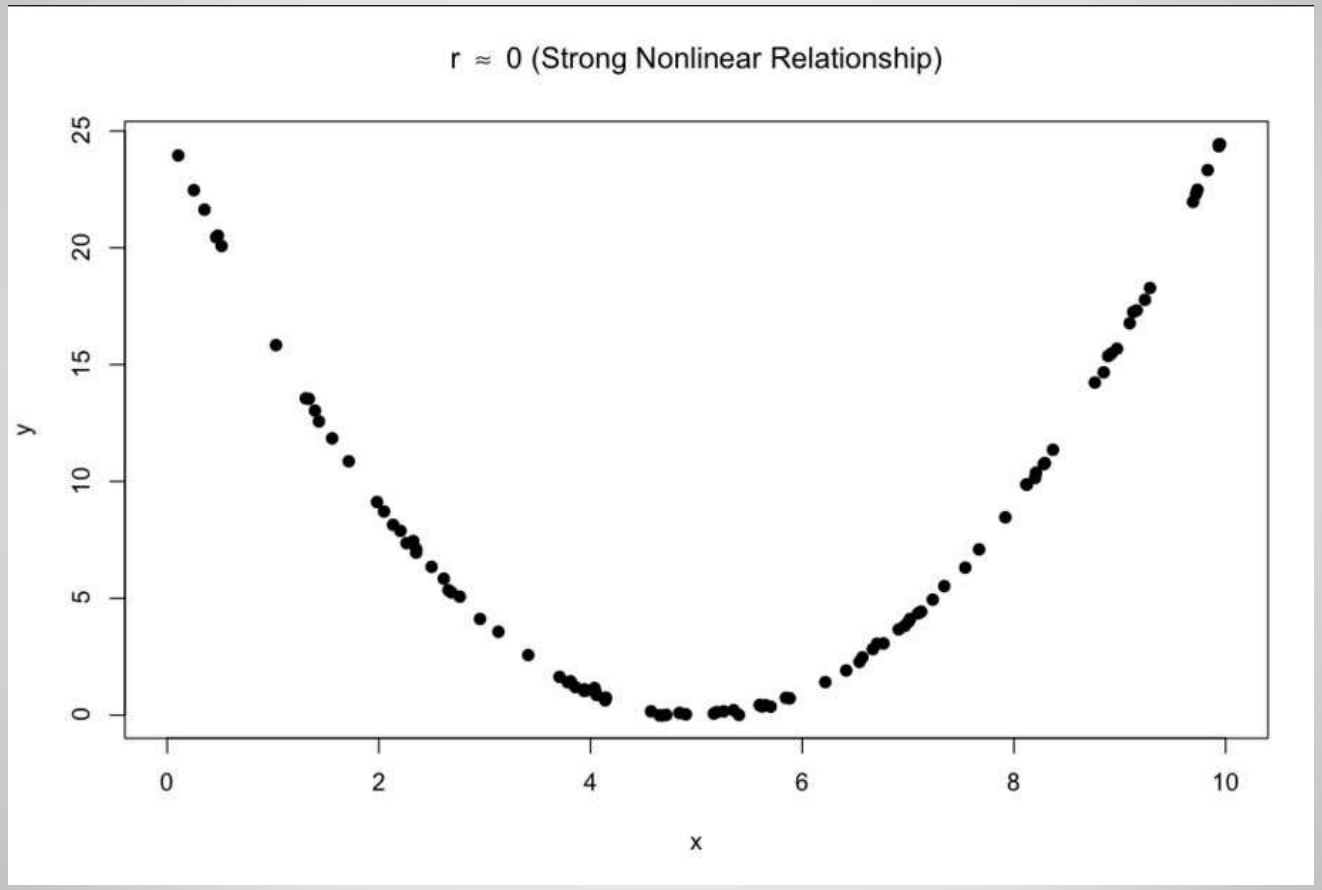
In general ;

- $y = a + bx$

The relationship between two variables is said to be **non – linear** if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate. In such cases, if the data is plotted on a graph sheet we will not get a straight line curve. For example, one may have a relation of the form

- $y = a + bx + cx2$

or more general polynomial.

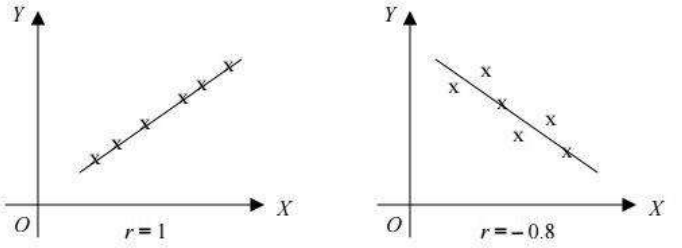

r ≈ 0 (Strong Nonlinear Relationship)
y
x
0
5
10
15
20
25
0
2
4
6
8
10

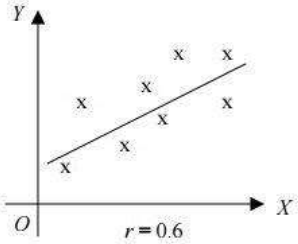
The Coefficient of Correlation

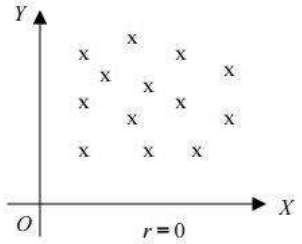
$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

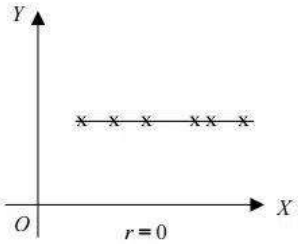
One of the most widely used statistics is the **coefficient of correlation** 'r' which measures the degree of association between the two values of related variables given in the data set.

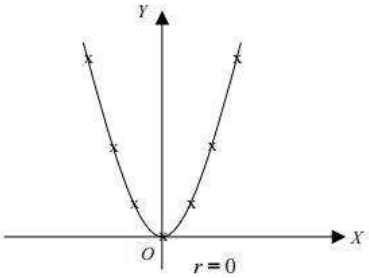
- It takes values from + 1 to – 1.
- If two sets or data have r = +1, they are said to be perfectly **correlated positively .**
- If r = -1 they are said to be perfectly **correlated negatively**; and if r = 0 they are uncorrelated**.**











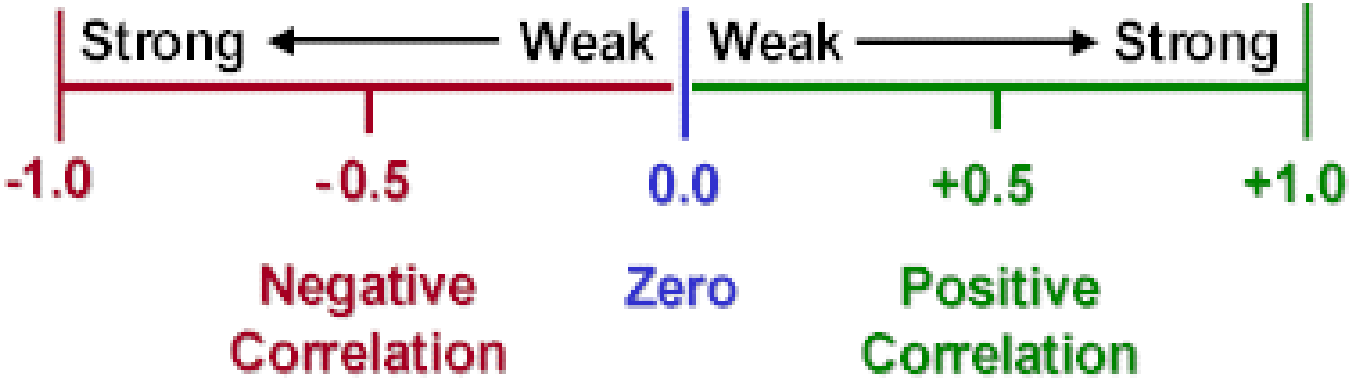
Y is independent of X, that is, Y assumes the same value irrespective of X.

X and Y have a non-linear relationship.

Fig. 2.3 Using scattering diagrams to determine *r* approximately

Correlation Coefficient

Shows Strength & Direction of Correlation



For Example: A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The following set of data was arrived at from a clinical study. Let us determine the coefficient of correlation for this set of data. The first column represents the serial number and the second and third columns represent the weight and blood pressure of each patient.

S. No.	Weight	Blood Pressure
1.	78	140
2.	86	160
3.	72	134
4.	82	144
5.	80	180
6.	86	176
7.	84	174
8.	89	178
9.	68	128
10.	71	132

Solution:

x	y	x^2	y^2	xy
78	140	6084	19600	10920
86	160	7396	25600	13760
72	134	5184	17956	9648
82	144	6724	20736	11808
80	180	6400	32400	14400
86	176	7396	30976	15136
84	174	7056	30276	14616
89	178	7921	31684	15842
68	128	4624	16384	8704
71	132	5041	17424	9372
796	1546	63,776	243036	1242069

Then

$$r=\frac{10(124206)-(796)(1546)}{\sqrt{\left[(10)63776-(796)^2\right]\left[(10)(243036)-(1546)^2\right]}}$$

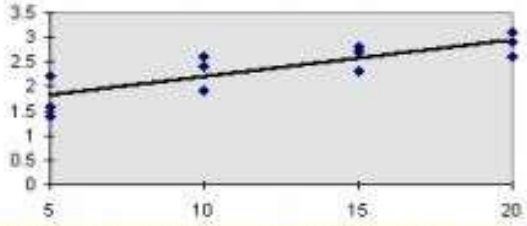
r = 0,5966

Regression Analysis

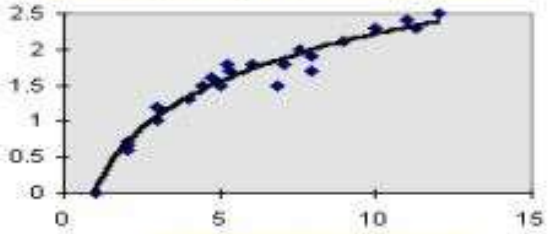
Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of the other variable. It is one of the most important statistical tools which is extensively used in almost all sciences – Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation of demand and supply graphs, cost functions, production and consumption functions and so on.

Types of Regression Models

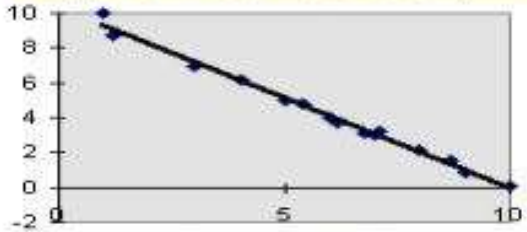
Positive Linear Relationship



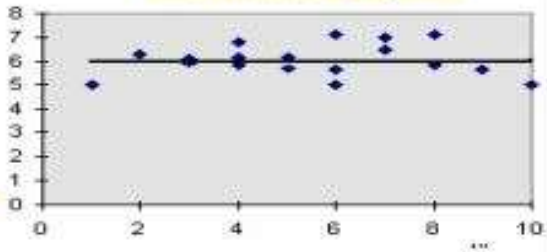
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Regression Equation

Suppose we have a sample of size ‘n’ and it has two sets of measures, denoted by x and y. We can predict the values of ‘y’ given the values of ‘x’ by using the equation, called the **regression equation**.

$$y^* = a + bx$$

where the coefficients a and b are given by

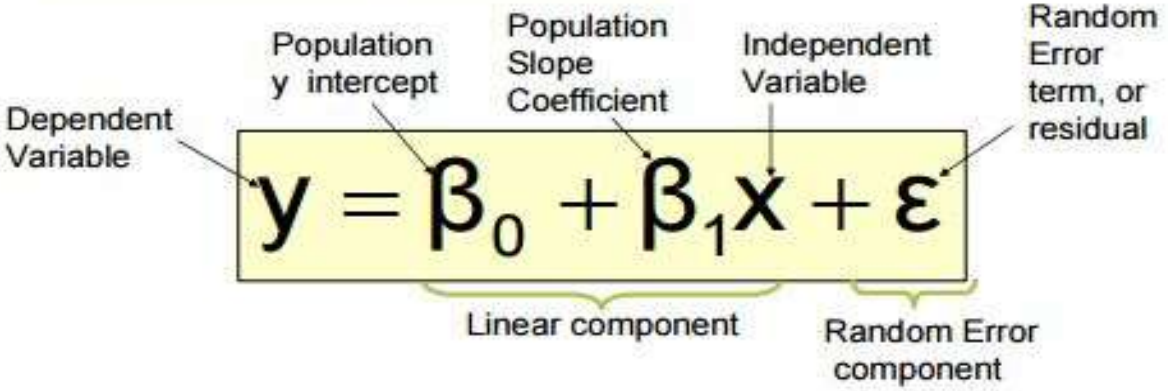
$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{\sum y - b\sum x}{n}$$

The symbol y* refers to the predicted value of y from a given value of x from the regression equation.

Population Linear Regression

The population regression model:

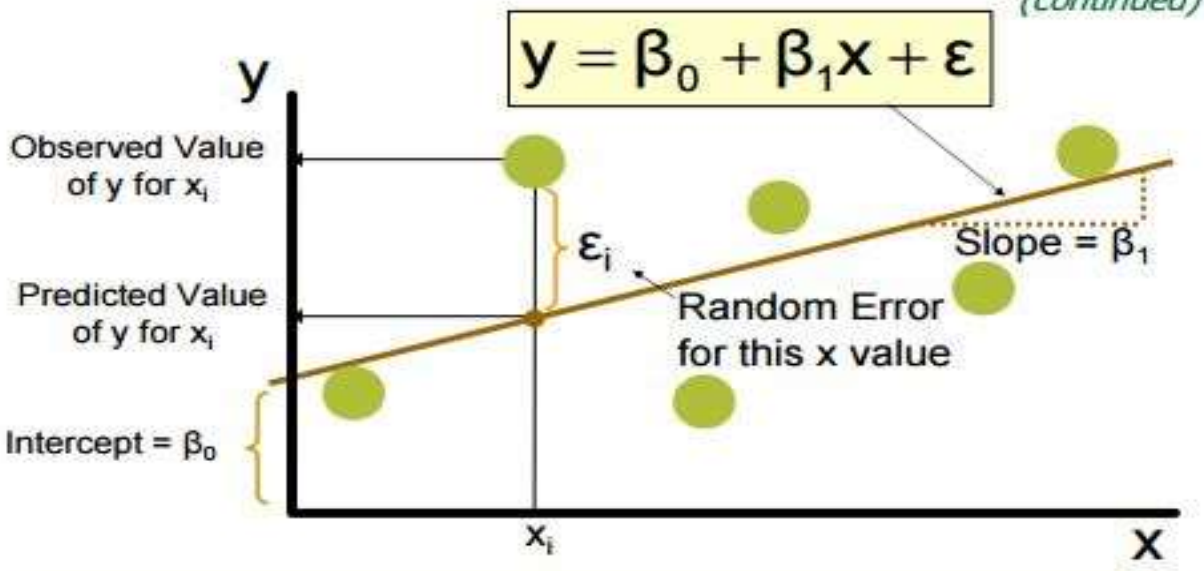


Linear Regression Assumptions

- Error values (ε) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the x variable and the y variable is linear

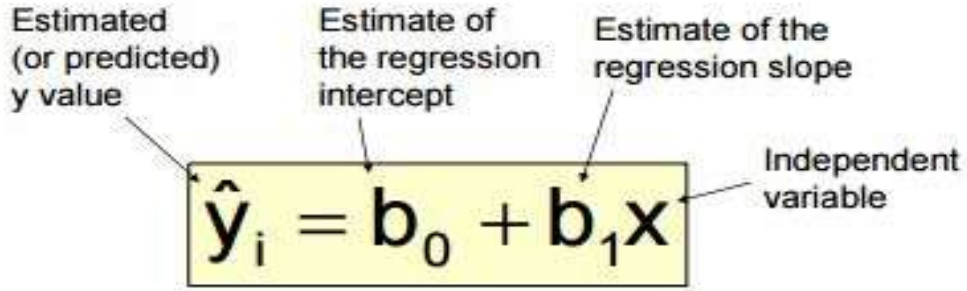
Population Linear Regression

(continued)



Estimated Regression Model

The sample regression line provides an estimate of the population regression line



The individual random error terms e_i have a mean of zero

Example:

Scores made by students in a statistics class in the mid - term and final examination are given here. Develop a regression equation which may be used to predict final examination scores from the mid – term score.

STUDENT	MID – TERM	FINAL
1.	98	90
2.	66	74
3.	100	98
4.	96	88
5.	88	80
6.	45	62
7.	76	78
8.	60	74
9.	74	86
10.	82	80

Solution:

We want to predict the final exam scores from the mid term scores. So let us designate ‘y’ for the final exam scores and ‘x’ for the mid – term exam scores. We open the following table for the calculations.

Stud	x	y	X^2	xy
1	98	90	9604	8820
2	66	74	4356	4884
3	100	98	10,000	9800
4	96	88	9216	8448
5	88	80	7744	7040
6	45	62	2025	2790
7	76	78	5776	5928
8	60	74	3600	4440
9	74	86	5476	6364
10	82	80	6724	6560
Total	785	810	64,521	65,071

Numerator of b = 10 * 65,071 – 785 * 810 = 6,50,710 – 6,35,850 = 14,860
Denominator of b = 10 * 64, 521 – $(785)^2$ = 6,45,210 – 6,16,225 = 28,985

Therefore, b = 14,860 / 28,985 = 0.5127

Numerator of a = 810 – 785 * 0.5127 = 810 – 402.4695 = 407.5305
Denominator of a = 10

Therefore a = 40.7531

Thus , the regression equation is given by

$$y^* = 40.7531 + (0.5127)\,x$$

We can use this to find the projected or estimated final scores of the students.
For example, for the midterm score of 50 the projected final score is
$y^* = 40.7531 + (0.5127)\,50 = 40.7531 + 25.635 = 66.3881$
which is a quite a good estimation.
To give another example, consider the midterm score of 70. Then the projected final score is
$y^* = 40.7531 + (0.5127)\,70 = 40.7531 + 35.889 = 76.6421,$
which is again a very good estimation.

Specify the Source

- https://medicine.tcd.ie/neuropsychiatric-genetics/assets/pdf/2009_4_Regression.pdf
- http://www2.sas.com/proceedings/forum2008/364-2008.pdf
- http://stud.pam.szczecin.pl/edu/eng/Chapter-5.pdf
- http://www.surgicalcriticalcare.net/Statistics/correlation.pdf
- http://www.personal.kent.edu/~mshanker/personal/Classes/f06/ch13_F06.pdf
- http://pages.intnet.mu/cueboy/education/notes/statistics/pearsoncorrel.pdf