

Chapter Ten

Linear Correlation and Regression Coefficient

Introduction:

In this chapter we present analysis to determine **the strength of relationship between two variables.**

In the case of linear regression, we will examine the amount of variability in one variable (**Y, the dependent variable**) that is explained by changes in another variable (**X, the independent variable**).

Item	correlation coefficient	regression coefficient
Definition	Measure the strength and direction of relationship between two variables	used to describe the functional relationship between two variables.. Predict the value of a dependent variable (Y) based on the value of independent variable (X).
	r → sample ρ (rho) → population	b → sample β (beta) → population
Nature of variables (Dependent and independent)	Both variable x and y are random and mutually dependent	X is random variable (independent) fixed – y is variable(dependent)
Range	Ranged from -1 to +1	take any value
Coefficient value	Value is relative	Value is absolute
Unit of measurement	Not take unit	Take unit of measurement

I. Simple linear correlation analysis

Correlation analysis is used to measure the intensity of association between one pair of variables.

Correlation show the extent to which two quantitative (continuous) variables, X and Y, "go together." When high values of X are associated with high values of Y, a positive correlation is said to exist. When high values of X are associated with low values of Y, a negative correlation is said to exist.

A widely used index of the association of two quantitative variables is the **Pearson product-moment correlation coefficient**, usually just called **correlation coefficient**.

The **correlation coefficient**, denoted symbolically as **r**, defines **both the strength and direction of the linear relationship** between two variables.

Characteristics of the correlation coefficient:

A. Correlation coefficient is an index number between -1 and +1.



- When $r = -1$, the variables **have a perfect negative linear relationship**. In this case, all points in the scatter diagram fall exactly on a straight line that slopes downward from left to right.

- When $r = +1$, the study variables **have a perfect positive linear relationship**. In this case, all points in the scatter diagram fall exactly on a straight line that slopes upward from left to right.

- When $r = 0$, **there is not a linear relationship between the study variables**. The relationship may be nonlinear; alternatively, the study variables may be unrelated, that is, a change in one variable has no effect on the other.

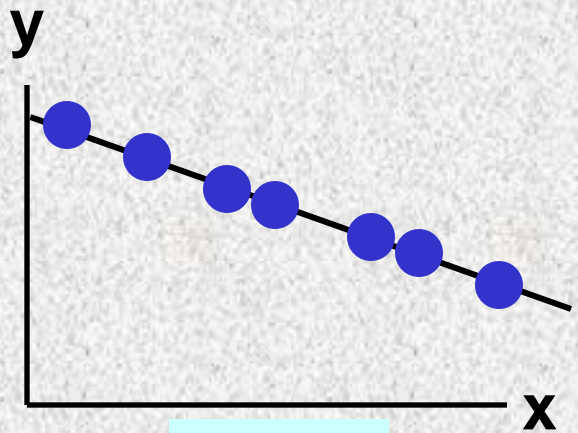


Features of ρ and r

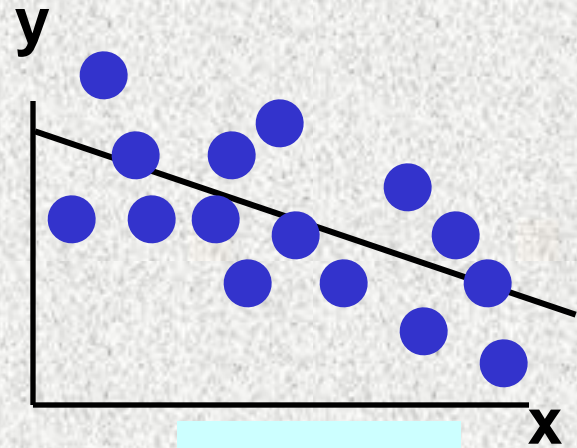
- **Unit free**
 - **Range between -1 and 1**
 - **The closer to -1, the stronger the negative linear relationship.**
 - **The closer to 1, the stronger the positive linear relationship.**
 - **The closer to 0, the weaker the linear relationship.**
- 
- 



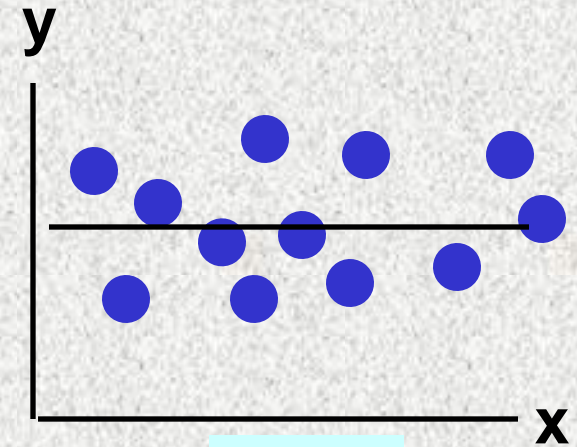
Examples of Approximate r Values



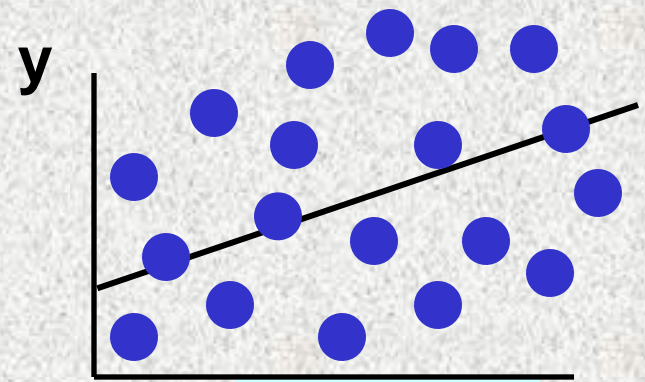
$r = -1$



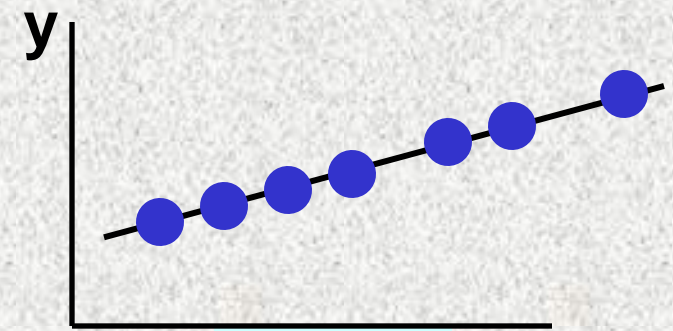
$r = -0.6$



$r = 0$



$r = +0.3$



$r = +1$

B. The better the points on the scatter diagram approximate a straight line, the greater the magnitude of r .

C. The correlation coefficient calculated for a sample drawn from a population of interest (r) is an estimate of the population correlation coefficient, denoted as (ρ). The population correlation coefficient is a measure of the linear association between the study variables for all members of the population. In other words, r is the statistic that estimates the population parameter.

Calculating the correlation coefficient:

We draw a random sample from the population of interest, compute r , the estimator of ρ , and test the hypothesis:

$$H_0 : \rho = 0$$

$$H_0 : \rho \neq 0$$

An alternative formula for calculating r is given by:

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Example:

The following data represent the relationship between two traits X1 and X2. The question is to calculate the correlation coefficient between them.

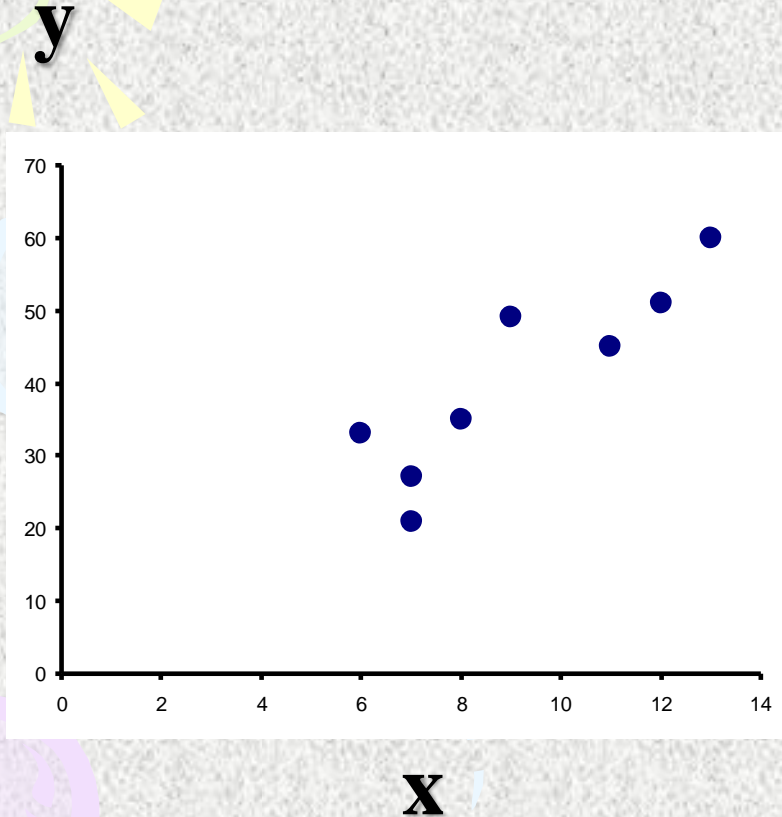
X1	8	9	7	6	13	7	11	12
X2	35	49	27	33	60	21	45	51

Calculation Example

y	x	xy	y²	x²
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ = 321	Σ = 73	Σ = 3142	Σ = 14111	Σ = 713

Calculation Example

(continued)



$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$
$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$
$$= 0.886$$

$r = 0.886 \rightarrow$ relatively strong positive linear association between x and y

Testing the correlation coefficient:

The suitable statistic for testing r is t , and is calculated as follows:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

$H_0: \rho = 0$ (No correlation)

$H_1: \rho \neq 0$ (correlation exists)

This value is to be compared to the critical value of t from the table at 6 d.f. (n-2), at 0.01 level of significance.

As the calculated t exceeds this value, then the null hypothesis of no correlation is rejected,

it is concluded that the two variables are correlated.

II. Simple Linear Regression:

regression is used to describe the **functional relationship between two variables.**

Predict the value of a dependent variable (Y) based on the value of independent variable (X).



Population Linear Regression

The population regression model:

Dependent Variable

Population y intercept

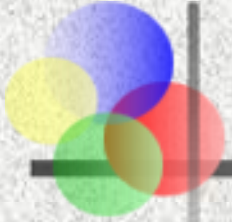
Population Slope Coefficient

Independent Variable

$$y = a + bX$$

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

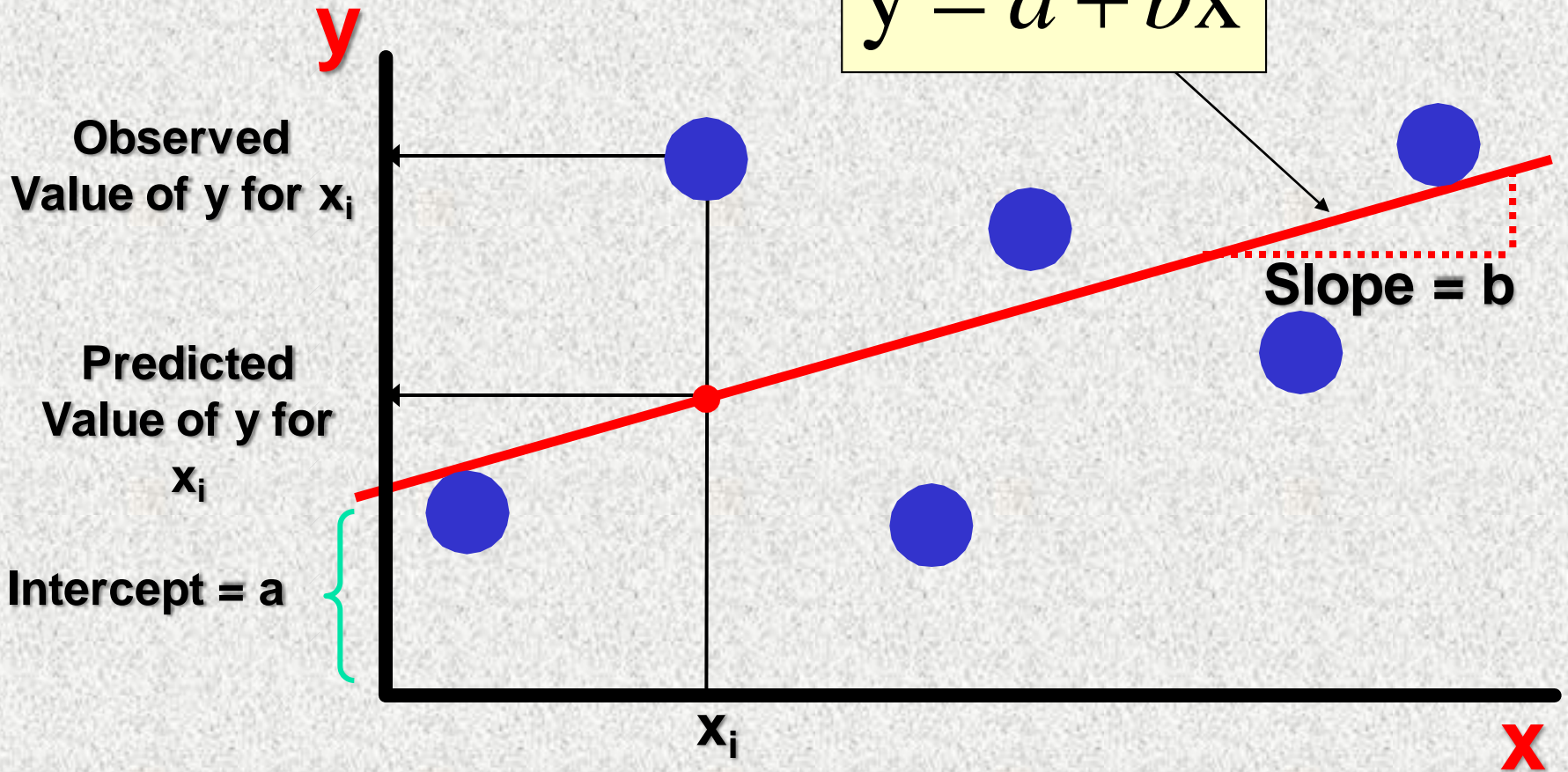
$$a = \sum y - b \sum x$$




Population Linear Regression

(continued)

$$y = a + bx$$





Interpretation of the Slope and the intercept

- **a** is the estimated average value of y when the value of x is zero
- **b** is the estimated change in the average value of y as a result of a one-unit change in x

1. The least squares line: ▽

It would seem simple to draw a free hand line through the points in the scatter diagram that tends to describe the relationship between the variable X and Y.

However, this method is not accurate, and is subject to personal judgment, which may differ between different investigators.

The most accurate regression line is the least squares line. This shows the minimum sum of squares of the distance between the data points and the fitted regression line. It follows the general equation for the straight line:

$$y = a + bx \quad \text{where:}$$

y is the value in the vertical axis,

x is the corresponding value in the horizontal axis,

a is the point at which **the regression line crosses the vertical axis (y intercept)**, and

b is the amount by which **y increases for each unit increase in x**, **b** is also called **the slope of the line**, or **the regression coefficient**.