

CHAPTER TEN

SIMPLE LINEAR REGRESSION AND CORRELATION

Objectives:

Having studied this unit, you should be able to:

- ✓ formulate a simple linear regression model.
- ✓ express quantitatively the magnitude and direction of the association between two variables

Introduction

The statistical methods discussed so far are used to analyze the data involving only one variable. Often an analysis of data concerning two or more variables is needed to look for any statistical relationship or association between them. Thus, regression and correlation analysis are helpful in ascertaining the probable form of the relationship between variables and the strength of the relationship.

8.1 Simple linear regression analysis

Regression analysis is the statistical method that helps to formulate a functional relationship between two or more variables. It can be used for assessment of association, estimation and prediction. For instance one might be interested to formulate a statistical model to relate the height of fathers and their sons, blood pressure and age, fertilizer amount and yield, etc.

A simple model to relate dependent (response) variable Y and with only one predictor variable X is to consider a linear relationship.

The first step in regression analysis involving two variables is to construct a scatter plot (diagram) of the observed data. Scatter diagram is a plot of all ordered pairs (X_i, Y_i) on the coordinate plane which is helpful for determining an apparent relationship between two variables.

The simple linear regression of Y on X can be expressed with respect to the population parameters α and β as

$$Y = \alpha + \beta X + \varepsilon$$

where α = y-intercept that represents the mean value of the dependent variable Y when the independent variable X is zero; β = slope of the regression line that represents the change in the mean of Y for a unit change in the value of X ; ε = error term

The population parameters α and β can be estimated from sample data using the least square technique. The estimators of α and β are usually denoted by a and b , respectively. The resulting regression line is

$$\hat{Y} = a + b X$$

and the equation is known as the fitted regression line. The estimated values of Y are denoted by $\hat{Y}$. The observed values of Y are denoted by y . The difference between the observed and the

estimated values, $Y - \hat{Y}$, is known as error or residual, and is denoted by $\{\hat{\epsilon}\}$. The residual can be positive, negative or zero.

A best fitting line is the one for which the sum of squares of the residuals, $\sum \hat{\epsilon}^2$ has the minimum value. This is called the method of least squares. According to this method, one would select a and b such that $\sum \hat{\epsilon}^2 = \sum (Y - \hat{Y})^2$ is minimum. The solution of this minimization problem using partial differentiation is as follows:

$$b = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} \quad \text{and} \quad a = \bar{Y} - b\bar{X}$$

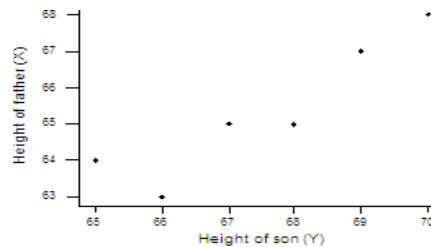
Example 8.1: A researcher wants to find out if there is any relationship between height of the son and his father. He took random sample of 6 fathers and their sons. The height in inch is given in the table below:

Height of father (X)	63	65	64	65	67	68
Height of the son (Y)	66	68	65	67	69	70

- Draw the scatter diagram and comment on the type of relationship.
- Fit the regression line of Y on X.
- Predict the height of the son if his father's height is 66 inch.

Solution:

i)



From the scatter plot one can see that the points are roughly on straight line.

ii)

$$n = 6, \sum X = 392, \sum Y = 405, \sum X^2 = 25628, \sum XY = 26476, \sum Y^2 = 27355$$

$$b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} = \frac{6(26476) - (392)(405)}{6(25628) - (392)^2} = 0.923 \quad a = \bar{Y} - b\bar{X} = \frac{405}{6} - 0.923 \frac{392}{6} = 7.2$$

Then the fitted (regression) line of Y on X is given by:

$$\hat{Y} = a + bX = 7.2 + 0.923X$$

- ✓ The slope of the line, i.e. $b=0.923$, tells us that a unit (one inch) increase in the height of the father results in 0.923 inch increase in the height of the son.
- ✓ The y-intercept of the line, i.e. $a=7.2$, is the value of Y when the value of X is zero (do you think that the intercept is meaningful?)

iii) $Y = 7.2 + 0.923(66) = 68.118$, thus the height of the son is 68.118 inch.

8.2 The covariance and the correlation coefficient

Correlation coefficient measures the degree of linear relationship between two variables. The population correlation coefficient is represented by ρ and its estimator is r . For a set of n pairs of sample values X and Y , Pearson's correlation coefficient is calculated as the ratio of the covariance of the variables X and Y to the product of the standard deviations of X and Y . symbolically,

$$r = \frac{Cor(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\sum \frac{(X - \bar{X})(Y - \bar{Y})}{n-1}}{\sqrt{\sum \frac{(X - \bar{X})^2}{n-1}} \sqrt{\sum \frac{(Y - \bar{Y})^2}{n-1}}} \\ = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

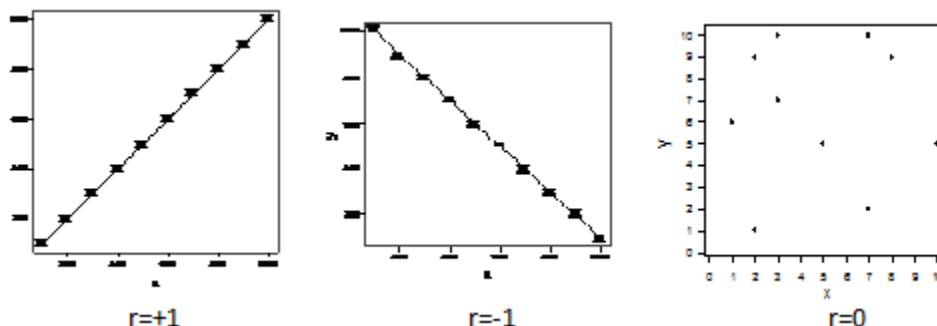
Alternatively, the Pearson's correlation coefficient r can be obtained as:

$$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}}$$

Properties of Pearson's correlation coefficient r ,

- It is appropriate to calculate when both variables X and Y are measured on an interval or ratio scale.
- The value of r is independent of the unit in which X and Y are measured. i.e., it is a pure number.
- The value of r ranges from +1 to -1.
- $r = +1$ indicates a perfect linear relationship between X and Y with positive slope.
- $r = -1$ indicates a perfect linear relationship between X and Y with negative slope.
- $r = 0$ indicates no linear relationship between the two variables X and Y .
- as r approaches +1 indicates strong and positive linear relationship between the two variables
- as r approaches -1 indicates strong and negative linear relationship between the two variables
- as r approaches 0 indicates weak linear relationship between the two variables

Examples of correlation coefficients:



Example 8.2: In some locations, there is strong association between concentrations of two different pollutants. An article reports the accompanying data on ozone concentration x (ppm) and secondary carbon concentration y ($\mu g / m^3$):

X	0.066	0.088	0.120	0.050	0.162	0.186	0.057	0.100
Y	4.6	11.6	9.5	6.3	13.8	15.4	2.5	11.8

0.112	0.055	0.154	0.074	0.111	0.140	0.071	0.110
8.0	7.0	20.6	16.6	9.2	17.9	2.8	13

- a. Calculate the correlation coefficient and comment on the strength and direction of the relationship between the two variables.

Solution: The summary quantities are

$$n = 16, \sum x_i = 1.656, \sum y_i = 170.6, \sum x_i y_i = 20.0397, \sum x_i^2 = 0.196912, \sum y_i^2 = 2253.56$$

The Person's correlation coefficient is

$$\begin{aligned}
 r &= \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}} \\
 &= \frac{16(20.0397) - (1.656)(170.6)}{\sqrt{16(0.196912) - (1.656)^2} \sqrt{16(2253.56) - (170.6)^2}} \\
 &= \frac{320.6352 - 282.5136}{\sqrt{0.408256} \sqrt{6952.6}} = \frac{38.1216}{(.639)(83.38)} \\
 &= 0.716
 \end{aligned}$$

The value of 0.716 indicates that there is somehow strong and positive relationship between ozone concentration and secondary carbon concentration.