SIMPLE REGRESSION AND CORRELATION

10.1 INTRODUCTION

The term *regression* was introduced by the English biometrician, Sir Francis Galton (1822–1911) to describe a phenomenon which he observed in analyzing the heights of children and their parents. He found that, though tall parents have tall children and short parents have short children, the average height of children tends to *step back* or to *regress* toward the average height of all men. This tendency toward the average height of all men was called a *regression* by Galton.

Today, the word *regression* is used in a quite different sense. It investigates the *dependence* of one variable, conventionally called the *dependent variable*, on one or more other variables, called *independent variables*, and provides an equation to be used for estimating or predicting the average value of the dependent variable from the known values of the independent variable. The dependent variable is assumed to be a random variable whereas the independent variables are assumed to have *fixed* values, *i.e.* they are chosen non-randomly. The relation between the expected value of the dependent variable and the independent variable is called a *regression relation*. When we study the dependence of a variable on a single independent variable, it is called a *simple* or *two-variable regression*. When the dependence of a variable on two or more than two independent variables is studied, it is called *multiple regression*. Furthermore, when the dependence is represented by a straight line equation, the regression is said to be *linear*, otherwise it is said to be *curvilinear*.

It is relevant to note that in regression study, a variable whose variation we try to explain is a *dependent variable* while an *independent variable* is a variable that is used to explain the variation in the dependent variable.

Some more terminology: The dependent variable is also called the *regressand*, the *predictand*, the *response* or the *explained variable* whereas the independent or the non-random variable is also referred to as the *regressor*, the *predictor*, the *regression variable* or the *explanatory variable*.

10.2 DETERMINISTIC AND PROBABILISTIC RELATIONS OR MODELS

The relationship among variables may or may not be governed by an exact physical law. For convenience, let us consider a set of n pairs of observation (X_i, Y_i). If the relation between the variables is *exactly linear*, then the mathematical equation describing the linear relation is generally written as

$$Y_i = a + bX_i,$$

where a is the value of Y when X equals zero and is called the *Y-intercept*, and b indicates the change in Y for a one-unit change in X and is called the *slope* of the line. Substituting a value for X in the equation, we can completely determine a *unique* value of Y. The linear relation in such a case is said to be a *deterministic model*. An important example of the deterministic model is the relationship between Celsius and Fahrenheit scales in the form of $F = 32 + \frac{9}{5}C$. Another example is the area of a circle expressed by the relation, area $= \pi r^2$. Such relations cannot be studied by regression.

In contrast to the above, the linear relationship in some situations is *not exact*. For example, we cannot precisely determine a person's weight from his height as the relationship between them is not expected to follow an exact linear form. The weights for given values of age are reasonably assumed to include measurement of random errors. The deterministic relation in such cases is then modified to allow

for the inexact relationship between the variables and we get what is called a *non-deterministic* or *probabilistic model* as

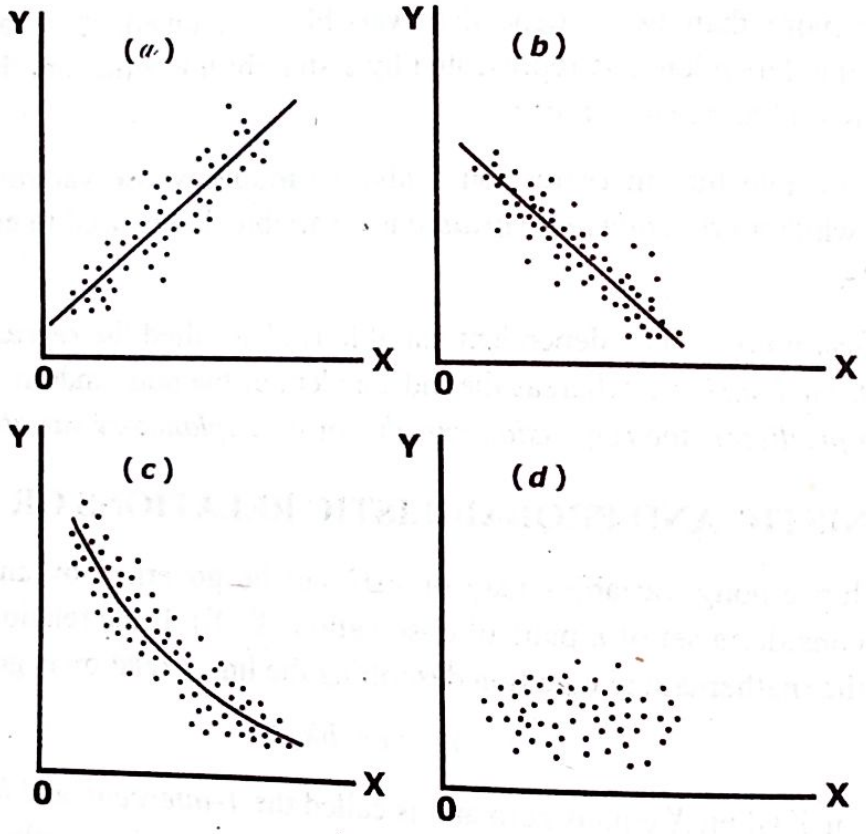
$$Y_i = a + bX_i + e_i, \qquad (i = 1, 2, \ldots, n)$$

where e_i's are the unknown random errors.

10.3 SCATTER DIAGRAM

A first step in finding whether or not a relationship between two variables exists, is to plot each pair of independent-dependent observations $\{(X_i, Y_i), i = 1, 2, \ldots, n\}$ as a point on graph paper, using the X-axis for the regression variable and the Y-axis for the dependent variable. Such a diagram is called a *scatter diagram* or a *scatter plot*. If a relationship between the variables exists, then the points in the scatter diagram will show a tendency to cluster around a straight line or some curve. Such a line or curve around which the points cluster, is called the *regression line* or *regression curve* which can be used to estimate the expected value of the random variable Y from the values of the nonrandom variable X.

The scatter diagrams shown below reveal that the relationship between two variable in (a) is positive and linear, in (b) is negative and linear, in (c) is curvilinear and in (d) there is no relationship.



10.4 SIMPLE LINEAR REGRESSION MODEL

We assume that the linear relationship between the dependent variable Y_i and the value X_i of the regressor X is

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where the X_i's are fixed or predetermined values,

the Y_i's are observations randomly drawn from a population,

the ε_i's are error components or random deviations,

$$= \frac{619.14}{646} = 0.958$$

A value of $r^2 = 0.958$ indicates that 95.8% of the variability in Y, the length of the spring, is demonstrated by its linear relationship with X, the weight on the spring.

10.5 CORRELATION

Correlation, like covariance, is a measure of the degree to which any two variables *vary* together. In other words, two variables are said to be *correlated* if they tend to simultaneously vary in some direction. If both the variables tend to increase (or decrease) together, the correlation is said to be *direct* or *positive*, *e.g.* the length of an iron bar will increase as the temperature increases. If one variable tends to increase as the other variable decreases, the correlation is said to be *negative* or *inverse*, *e.g.* the volume of gas will decrease as the pressure increases. It is worth remarking that in correlation, we assess the strength of the relationship (or interdependence) between two variables; both the variables are random variables, and they are treated symmetrically, *i.e.* there is no distinction between dependent and independent variable. In regression, by contrast, we are interested in determining the dependence of one variable that is random, upon the other variable that is non-random or fixed, and in predicting the average value of the dependent variable by using the known values of the other variable.

10.5.1 Pearson Product Moment Correlation Co-efficient. A numerical measure of strength in the *linear* relationship between any two variables is called the *Pearson's product moment correlation co-efficient* or sometimes, the *coefficient of simple correlation or total correlation.* The sample *linear* correlation coefficient for n pairs of observations (X_i, Y_i) usually denoted by the letter r, is defined by

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

The population correlation co-efficient for a bivariate distribution, denoted by ρ, has already been defined as

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var(X)Var(Y)}}}$$

For computational purposes, we have an alternative form of r as

$$r = \frac{\Sigma XY - (\Sigma X)(\Sigma Y)/n}{\sqrt{[\Sigma X^2 - (\Sigma X)^2/n][(\Sigma Y^2 - (\Sigma Y)^2/n)]}}$$

$$= \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

This is a more convenient and useful form, especially when $\bar{X}$ and $\bar{Y}$ are not integers. The coefficient of correlation r is a pure number (*i.e.* independent of the units in which the variables are measured) and it assumes values that can range from +1 for perfect positive linear relationship, to –1, for perfect negative linear relationship with the intermediate value of zero indicating no linear relationship between X and Y. The sign of r indicates the direction of the relationship or correlation.

It is important to note that $r = 0$ does not mean that there is no relationship at all. For example, if all the observed values lie exactly on a circle, there is a perfect *non-linear* relationship between the variables but r will have a value of zero as r only measures the linear correlation.

The linear correlation co-efficient, is also the square root of the linear co-efficient of determination, r^2.

We have $\hat{Y} = \bar{Y} + b(X - \bar{X})$

or $\hat{Y} - \bar{Y} = b(X - \bar{X})$

Squaring both sides, we get

$$(\hat{Y} - \bar{Y})^2 = b^2(X - \bar{X})^2$$

Substituting in the ratio, we find

$$\frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2} = \frac{b^2 \Sigma(X - \bar{X})^2}{\Sigma(Y - \bar{Y})^2}$$

$$= \frac{\Sigma(X - \bar{X})^2}{\Sigma(Y - \bar{Y})^2} \left[\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} \right]^2$$

$$= \left[\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(Y - \bar{Y})^2 \, \Sigma(X - \bar{X})^2}} \right]^2 = r^2$$

Example 10.5 Calculate the product moment co-efficient of correlation between X and Y from the following data:

X	1	2	3	4	5
Y	2	5	3	8	7

(P.U., B.A./B.Sc. 1973)

The calculations needed to compute r are given below:

X	Y	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	2	–2	4	–3	9	6
2	5	–1	1	0	0	0
3	3	0	0	–2	4	0
4	8	1	1	3	9	3
5	7	2	4	2	4	4
15	25	0	10	0	26	13