Regression Analysis Concept

The English scientist Francis Galton (1822-1911) is considered the first to use the concept of regression in biological applications with the aim of discovering some relationships between some biological variables.

The subject of regression analysis aims to estimate the mathematical relationship that links two or more variables together.

In regression analysis, there are two types of variables, the first is the dependent variable and the second is the independent variable (or a set of independent variables).

The dependent variable is the variable whose values are affected in the event of a change in the values of the independent variable and is not affected by it.

For example, the monthly family spending is affected by its monthly income, while the opposite is not true.

The independent variable is often called the predicted variable or the explanatory variable.

Definition of Regression Analysis

Regression analysis is a statistical method used to analyze the relationship between one or more independent variables and a dependent variable.

Or it can be defined in general as a mathematical measure of the average relationship between two or more variables in terms of the units of measurement of the variables in the relationship. Relations of this type are often called regression models.

Uses of Regression Analysis

1- Data Description: A set of data can be summarized and described by the researcher to find the regression equation that describes that data.

2- Parameter Estimation: The unknown parameters of the model can be estimated, through which the importance, strength and direction of the relationship between the variables can be inferred.

3- Prediction: The response can be estimated and predicted, which is very useful in planning and decision-making.

4- Control: After finding the models that describe the relationship between the independent variables and the dependent variable, the values of the dependent variable can be controlled by changing the values of the independent variables.

Regression and causality

What is meant by causality is that if regression analysis proves the existence of a relationship between X and y, this does not mean that X is the cause of y, unless this is proven by other experiments and other variables are taken into consideration.

Types of Regression

1- Simple Linear Regression (in the absence of repetition in the Xi values)
Simple linear regression includes two variables, the first is the dependent variable and the second is the independent variable. An example of this is the relationship between age X and blood pressure y for fifty people in a particular city.
Simple linear regression can be defined as the process of estimating the relationship between two variables, one of which is independent and the other dependent.
Simple linear regression that explains the relationship between two variables generates a pair of values. For example, the variable X (the independent variable) and the variable y (the dependent variable) are the variables that represent the pair of values (X and y), meaning that Xi represents the values i of the variable X, which corresponds to the value yi, which represents the value i of the variable y.
If we represent the pair of values (Xi, yi), we will get a graph called the scatter plot Figure (1)). Note the figure below Figure (1):

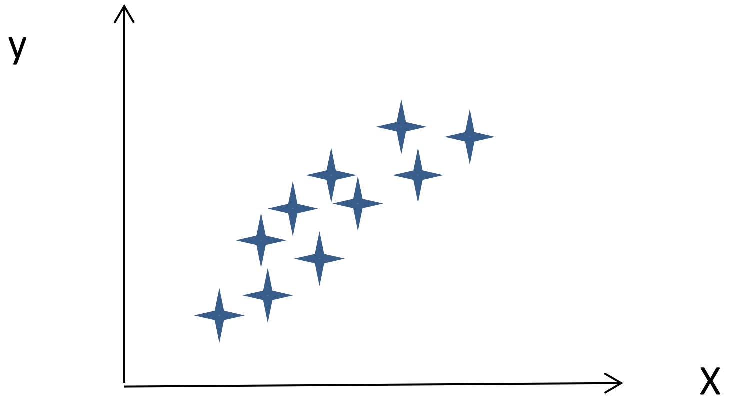

Figure (1) plot of the pair of values (Xi,yi)

Simple linear regression can be represented by a linear equation called the simple regression line equation, which takes the following form:

General linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i \quad , \; i = 1, 2, \ldots, n$$

Whereas:
y_i: It is the value of the dependent variable or the response variable.

x_i: It is the value of the independent variable.

β_0, β_1: They are constants called regression equation parameters.

Whereas:
β_0: It is the point of intersection of the regression line with the y-axis.

β_1: The regression coefficient of y on X_{i1} represents the amount of change in y when X_{i1} increases by one unit.

e_i: It is the value of the random or residual error.

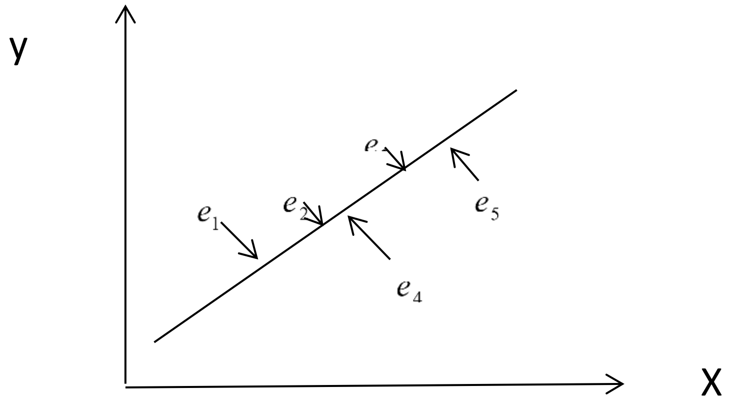

Figure (2) plot of random errors

Through the simple regression line equation above (Equation (1)):
If the value of β_1 is equal to zero, then β_0 represents the amount of increase or decrease in the value of y.
We notice from the figure above that by drawing the pair of values (Xi, yi), we can find the equation of the regression line that passes through these points so that the sum of the squares of these errors is as small as possible.

The regression line equation can be explained more clearly by drawing this equation and determining the values of the parameters β_1 and β_0. Note the figure below (Figure (3)):

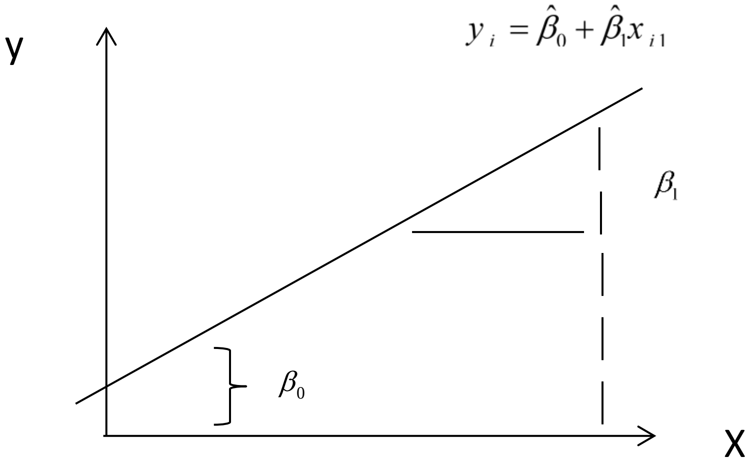

Figure (3) Regression line equation and parameters β_1 and β_0

We note that β_0 is the point of intersection of the regression line equation with the vertical axis.

As for β_1, it represents the slope of the regression line (Slop) $\frac{\Delta y}{\Delta X}$, which is the amount of change in y resulting from a one-unit increase in X.

The sign of the parameter β_1 represents the relationship between X and Y. If it is positive, the relationship is direct, and if it is negative, the relationship is inverse.

For the error e_i, it represents the difference between the observed value y_i and the estimated value $\hat{y}_i$, i.e.:

$$e_i = y_i - \hat{y}_i$$

were

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

were

$\hat{\beta}_0$: It is the estimated parameter for parameter β_0.
$\hat{\beta}_1$: It is the estimated parameter for parameter β_1.
$\hat{y}_i$: is the estimated value of the value y_i.

Analysis Assumptions

1- y is a random variable, and its values are statistically independent of each other and are normally distributed with an arithmetic mean $\mu_{y/x}$ and variance $\sigma^2 = \sigma^2_{y/x}$. y is a random variable because its values change from one value to another of the values of X. The random variable y that depends on a fixed value of X is usually symbolized by y(xi) and its probability distribution is symbolized by F(y/x). Note the figure below:

Assumptions of analysis

The mean of the y values (i.e. $\mu_{y/x}$) is a straight-line function, i.e. the value lies on a straight line.
1-The variance of y ($\sigma^2_{y/X}$) has one value from any value of X, i.e. $\sigma^2_{y/X} = \sigma^2$ this property is called homosccdasticity, i.e. the stability or homogeneity of the error variance.

2- e_i is a random error and is normally distributed with a mean of zero and a variance of X, i.e. $e_i \sim N(0, \sigma_e^2)$.

3-The e_{is} values are not correlated, meaning that the common variance between them is zero, meaning that $E(ee) = 0$, also e_i, is a value in one period that is not correlated with its value in another period.

4-The values of the variable X_{i1} are constant and measured without error, and its values are not related to the error values, i.e. $E(x_i' e_i) = 0$.

Estimating regression parameters

1- Least squares method

The simple linear regression model is represented by the following equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i, i = 1,2, \dots, n$$