Regression Analysis

Md. Pear Hossain
Lecturer of Statistics
BSMRSTU, Bangladesh

Introduction

- In science, we frequently measure two or more variables on the same individual (case, object, etc).
- We do this to explore the nature of the relationship among these variables.
- There are two basic types of relationships.
  1. Cause-and-effect relationships.
  2. Functional relationships.

Function: a mathematical relationship enabling us to predict what values of one variable (Y) correspond to given values of another variable (X).

Y: is referred to as the **dependent variable**, the **response variable** or the **predicted variable**.

X: is referred to as the **independent variable**, the **explanatory variable** or the **predictor variable.**

Definition

Regression analysis is a technique to study the cause and effect relationship of two or more variables when variables are recorded from each sample point of a given sample.

- For example, the weight of a person is usually increased if his height is more, provided other conditions remain normal.
- The number of ever born children of a couple depends on their duration of marriage.
- The agricultural production depends on amount of fertilizer used.
- The performance of a student depends on his time spent for learning.

Regression model

The regression model is defined as

$$y = \alpha + \beta x + e$$

The parameter β is called regression coefficient of y on x which measures the rate of change of y for a unit change in the value of x.

α is a parameter measuring the value of y in absence of x and

e is called random component.

Assumptions for regression analysis

i. The dependent variable y is assumed to be normally distributed.
ii. The explanatory variable are non-random variables.
iii. The explanatory variables are uncorrelated.
iv. There is no measurement error in measuring the explanatory variables.
v. The error term e is normally and independently distributed.

Fitting of regression line

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n pairs of values from a random sample. Then the fitted regression equation is

$$\hat{y} = a + bx$$

where,

$$a = \bar{y} - b\bar{x}$$

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

Interpretation of $\boldsymbol{b}$: Let, $\boldsymbol{\hat{y} = 24.5 + 0.0509x}$.

Since $b = 0.0509$; it indicates as x increases by 1 unit, $\hat{y}$ will increase by 0.0509 unit.

Example

The following data represent the values of gestation age *(x, days)* and birth weight *(y, pound)* of some new born babies of 10 mothers:

x	265	250	270	255	260	258	255	265	248
y	6.2	5.8	7.0	5.6	6.0	5.6	6.4	6.8	5.2

1. Fit a regression line of *y* on *x*. **Hint:** In this regard, you have to first calculate the coefficients (both slop coefficient and regression coefficient) and then substitute it in the regression line.
2. predict y for a given value of x=210**.**
3. Test the significance of the regression.

Testing significance of regression coefficient β

Step 1. Let us we have to test the following hypothesis

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

Step 2: At α% level of significance we will test the above hypothesis.

Step 3: Under null hypothesis the test statistic is

$$\text{Type equation here.}$$