Regression analysis

Regression analysis is concerned with the study of the dependence of one variable (the *dependent variable)* on one or more other variables (the *independent variables*). In regression analysis we use independent variables to estimating or predicting the average changes of the dependent variables.

Examples:

- The marketing director of a company may want to know how the demand for the company's product is related, to say advertising expenditure.
- Suppose the sales manager of a company say X, wants to determine how the number of credit cards sell is related to the number of call.

Some terms related to regression analysis:

A Scatter plot: is a chart that portrays the relationship between two variables.

Dependent variable: The dependent variable is the variable being predicted or estimated. It is denoted by *y*.

Independent variable: The Independent Variable provides the basis for estimation. It is the predictor variable. It is denoted by *x*.

Example: If we want to know the expected weekly production of a company then production will be the dependent variable and the predictor/independent variables could be the capital, number of labours engaged, supply of raw materials etc.

Linear Regression Model: The equation that describes how *y* is related to *x* and an error term is called the regression model.

The linear regression equation is:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

The estimated linear regression equation is give as:

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

Where,

$\hat{y}_i$ is the average predicted value of y_i for any x_i.

β_0 and β_1 are called parameters of the model.

β_0 is called intercept or it is the estimated $\hat{y}_i$ value when *X=0*

β_1 is called slope or the average change in estimated value of $\hat{y}_i$ for each change of one unit in x.

e_i is a random variable called the error term.

Slope: A slope of 2- means that every 1-unit change in X yields a 2-unit change in Y.

Error term: The deviation of a particular value of y from $\hat{y}_i$ is called error term or the stochastic disturbance term. It is written as $e_i = (y_i - \hat{y}_i)$.

Assumption of Linear Regression Model:

1. The y 's are linear functions of x plus a random error term.

   $y_i = \beta_0 + \beta_1 x_i + e_i$

2. The dependent values (y_i) are independent of each other, i.e. if we obtain a large value of y_i on the first observation, the result of the second and subsequent observations will not necessarily provide a large value.
3. The error terms are random variables with mean, $\mu = 0$ and variance, σ^2 .i.e. $E(e_i) = 0$ and $E(e_i^2) = \sigma^2$ for i=1,2,….n.
4. The random error terms, e_i are not correlated with one another,

   So that $E(e_i e_j) = 0$ for all $i \neq j$.

Least Squares Method: To estimate the regression parameters we used least square method. This regression technique that calculates the β so as to minimize the sum of the squared errors. That is

$$\min \sum (y_i - \hat{y}_i)^2$$

Where:

y_i = observed value of the dependent variable for the ith observation

$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

Slope for the estimated regression equation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Intercept for the estimated regression equation:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

x_i = value of independent variable for *i*th observation

y_i = value of dependent variable for *i*th observation

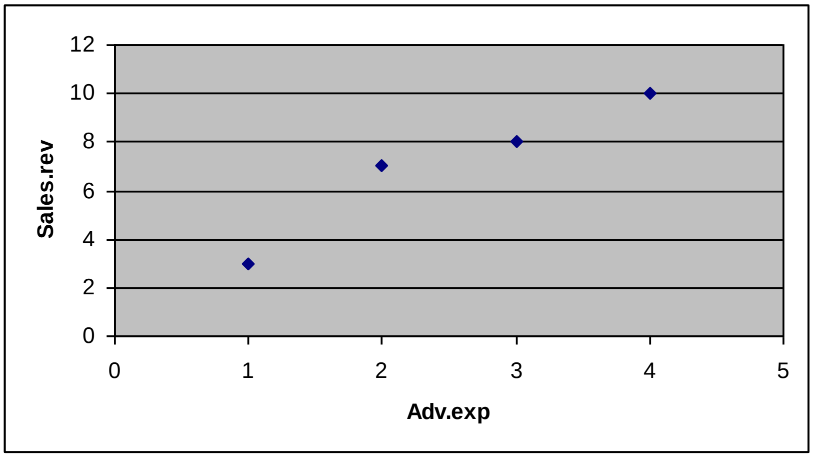
$\bar{x}$ = mean value for independent variable
$\bar{y}$ = mean value for dependent variable
n = total number of observations

Problem: Suppose a company's owner wants to forecast sales on the basis of advertising expenses. The owner would like to review the relationship between sales and the amount spent on advertising. Below is the information on sales and advertising expense for the last four months:

Month	Advertising expense(x) ($ million)	Sales revenue(y) ($ million)
July	2	7
August	1	3
September	3	8
October	4	10

a) Draw scatter plot.
b) Determine the estimated regression equation.
c) Interpret the value β_0 and β_1.
d) Estimate sales when $3 million is spent on advertising.

Solution:
a) If we plot the data and draw scatter plot we get as below plot



b) We know, $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

$$\overline{x} = \frac{\sum_{i=1}^{4} x_i}{n} = \frac{10}{4} = 2.5 \quad \text{and}$$

$$\overline{\text{y}} = \frac{\sum_{i=1}^{4} y_i}{n} = \frac{28}{4} = 7$$

Advertising expense(x)	Sales revenue(y)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
2	7	-0.5	0	0.25	0
1	3	-1.5	-4	2.25	6
3	8	0.5	1	0.25	0.5
4	10	1.5	3	2.25	4.5
				$\sum_{i=1}^{4}(x_i - \overline{x})^2$ =5	$\sum_{i=1}^{4}(x_i - \overline{x})(y_i - \overline{y})$ =11

Now,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{11}{5}$$

$\hat{\beta}_1$ =2.2

$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ =7-2.2*2.5

$\hat{\beta}_0$ =1.5

Now if we put the values of **estimated regression equation is**

$\hat{y}_i = 1.5 + 2.2x_i$

c) Interpretation:

Here slope is, $\hat{\beta}_1$ =2.2.This means that an increase of $1million in advertising cost, the sales revenue will increase $2.2 million.

And $\hat{\beta}_0$ =1.5 means that, that is if there is no advertisement cost, then sales revenue would be $1.5 million.

d) Now if x= 3, then

$\hat{y} = 1.5 + 2.2 * 3 = 8.1$

So, when advertisement cost is \$3 million, the sales revenue would be 8.1 million.

Standard Error: The Standard Error of Estimate measures the scatter, or dispersion, of the observed values around the line of regression .The formula that is used to compute the standard error:

$$s_{y.x} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n-2}}$$

Example: From the previous example we calculate the value of $\hat{y}_i$ .As we know

Standard Error, $s_{y.x} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n-2}}$

y_i	$\hat{y}_i$	$\left(y_i - \hat{y}_i \right)$	$\left(y_i - \hat{y}_i \right)^2$
7	5.9	1.1	1.21
3	3.7	-0.7	0.49
8	8.1	-0.1	0.01
10	10.3	-0.3	0.09
			$\sum_{i=1}^{4} \left(y_i - \hat{y}_i \right)^2$ =1.8

Here, n=4.

Now, by putting values we get,

$$S_{y,x} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{4-2}}$$

$$= \sqrt{\frac{1.8}{4-2}}$$

$$= \sqrt{0.9}$$

$$= 0.95$$

So, the standard error is 0.95.

Coefficient of Determination (r^2): The coefficient of determination tells the percent of the variation in the dependent variable that is explained (determined) by the model and the explanatory variable.

Interpretation of r^2 **:** Suppose r^2 =92.7%.
Interpretation: Almost 93% of the variability of the dependent variables explained by the independent variables.

Relationship between regression and correlation:

$$\hat{r} = \hat{\beta}_1 \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}}$$

Regression versus correlation:

- **In Correlation** the primary objective is to measure the *strength* or *degree* of *linear association* between two variables.

  In regression analysis, we are not primarily interested in such a measure. Instead, we try to estimate or predict the average value of one variable on the basis of the fixed values of other variables

- In correlation analysis, there is no distinction between the dependent and explanatory variables.

  In regression analysis there is an asymmetry in the dependent and explanatory variables are treated. The dependent variable is assumed to be statistical, random, or stochastic (that is, to have a probability distribution). The explanatory variables, on the other hand, are assumed to have fixed values (in repeated sampling),