**Chapter- 7**

**Linear Regression**

**Objectives:**

* To learn how many business decisions depend on knowing the specific relationship between two or more variables;
* To use regression analysis to estimate the relationship between two variables; and
* To use the least- squares estimating equation to predict future values of the dependent variable.

**Contents:**

* Introduction;
* Definition and Types of Regression;
* Linear Regression Equation;
* Methods of Regression Analysis;
* Estimation Using the Regression Line.

# **Introduction:**

After knowing the relationship between two variables we may be interested in estimating (predicting) the value of one variable given the value of another. The variable predicted on the basis of other variables is called the “dependent” or the ‘explained’ variable and the other the *independent* or the *predicting variable*. The prediction is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise, is called the regression equation or the explaining equation.

For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

The relationship between two variables can be considered between, say, rainfall and agricultural production, price of an input and the overall cost of product, consumer expenditure and disposable income. Thus, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

# **Definition:**

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

# **Types of Regression:**

The regression analysis can be classified into:

1. Simple and Multiple
2. Linear and Non –Linear
3. Total and Partial

# **Simple and Multiple:**

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationships, more than two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones.

For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as y = f (x,z).

# **Linear and Non-linear:**

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predictive value, a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

# **Total and Partial:**

In the case of total relationships all the important variables are considered. Normally, they take the form of a multiple relationships because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

# **Linear Regression Equation:**



If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of **Y on X is Y = *a* + *b*X** ……. (1) and

**X on Y is**

**X = *a* + *b*Y**……. (2) where *a, b* are constants.

From (1) We can estimate Y for known value of X.

(2) We can estimate X for known value of Y.

# **Regression Lines:**

For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables.

For perfect correlation, positive or negative i.e., r = +1, the two lines coincide i.e., we will find only one straight line. If r = 0, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axes.

**Y** r = - 1

r = + 1

Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X- axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y- axis will touch the mean value of Y.

# **Principle of ‘Least Squares’:**

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of “least squares”. This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

* + - 1. The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.,

(X – Xc) = 0 or (Y- Yc ) = 0

Where Xc and Yc are the values obtained by regression analysis.

* + - 1. The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e.,

(Y – Yc)2 < (Y – Ai)2

Where Ai = corresponding values of any other straight line.

* + - 1. The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is

*x*, *y* .

# **Methods of Regression Analysis:**

There are two methods of regression analysis-

1. Graphic Method through Scatter Diagram; and
2. Algebraic Method through regression equations (normal equation and through regression coefficient).

# **Graphic Method:**

# **Scatter** **Diagram:**

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

# **Algebraic Methods:**

* + - 1. *Regression Equation.*

The two regression equations for X on Y; X = *a + b*Y

And for Y on X; Y = *a* + bX

Where X, Y are variables, and a,b are constants whose values are to be determined

For the equation, X = *a + b*Y, The normal equations are

X = n*a + b* Y and

XY = *a*Y + *b*Y2

 For the equation, Y= *a* + *b*X, the normal equations are

Y = n*a* + *b*X and

XY = *a*X + *b*X2

From these normal equations the values of *a* and *b* can be determined. After solving these two simultaneous equations, we may find the values of *a* and *b* as:

**Slope of the Best- Fitting Regression Line, b =** $\frac{∑XY-nX̅̅Y̅̅}{∑X^{2}-nX^{2}}$

**Y- Intercept of the Best- Fitting Regression Line, a = Y̅ - bX̅**

# **Example 1:**

***Find the two regression equations from the following data:***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***X:*** | ***6*** | ***2*** | ***10*** | ***4*** | ***8*** |
| ***Y:*** | ***9*** | ***11*** | ***5*** | ***8*** | ***7*** |

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | X2 | Y2 | XY |
| 6 | 9 | 36 | 81 | 54 |
| 2 | 11 | 4 | 121 | 22 |
| 10 | 5 | 100 | 25 | 50 |
| 4 | 8 | 16 | 64 | 32 |
| 8 | 7 | 64 | 49 | 56 |
| 30 | 40 | 220 | 340 | 214 |

Regression equation of Y on X is Y = *a* + *b*X and the normal equations are

Y = n*a* + *b*X

XY = *a*X + *b*X2

**Substituting the values, we get**

40 = 5*a* + 30*b* …… (1)

214 = 30*a* + 220*b* ……. (2)

**Multiplying (1) by 6**

240 = 30*a* + 180*b*……. (3)

 (2) – (3) - 26 = 40*b*

 b = $\frac{-26}{40}$ = - 0.65

Now, substituting the value of ‘b’ in equation (1) 40 = 5*a* – 19.5

5*a* = 59.5

 *a* = $\frac{59.5}{5}$ = 11.9

Hence, required regression line Y on X is Y = 11.9 – 0.65 X. Again, regression equation of X on Y is

X = *a* + *b*Y and

**The normal equations are**

X = n*a* + *b*Y and

XY = *a*Y + *b*Y2

Now, substituting the corresponding values from the above table, we get

 30 = 5*a* + 40*b* …. (3)

214 = 40*a* + 340*b* …. (4)

**Multiplying (3) by 8, we get**

240 = 40*a* + 320 *b* …. (5)

(4) – (5) gives

-26 = 20*b*

 b = $\frac{-26}{20}$ = - 1.3

Substituting *b* = - 1.3 in equation (3) gives 30 = 5*a* – 52

 5*a* = 82

 *a =* $\frac{82}{5}$*=* 16.4

**Hence, required regression line of X on Y is**

X = 16.4 – 1.3Y

* + - 1. *Regression Co-efficient:*

The regression equation of Y on X is y = *ȳ+r*$\frac{σy}{σx}$*(x- x)*

Here, the regression Co-efficient of Y on X is

b1 = byx = *r*$\frac{σy}{σx}$

The regression equation of X on Y is x = *x̅ +r*$\frac{σx}{σy}$*(x- x)*

 Here, the regression Co-efficient of X on Y is

b2 = bxy = *r*$\frac{σx}{σy}$

If the deviation are taken from respective means of x and y

b1 = byx = $\frac{∑xy}{∑x^{2}}$

b2 = bxy = $\frac{∑xy}{∑y^{2}}$

# **Properties of Regression Co-efficient:**

1. Both regression coefficients must have the same sign, ie either they will be positive or negative.
2. Correlation coefficient is the geometric mean of the regression coefficients i.e., r = $\pm \sqrt{b\_{1}b\_{2}}$
3. The correlation coefficient will have the same sign as that of the regression coefficients.
4. If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
5. Regression coefficients are independent of origin but not of scale.
6. Arithmetic mean of b1 and b2 is equal to or greater than the coefficient of correlation i.e., $\frac{b\_{1}+b\_{2}}{2}$ $\geq r$
7. If r=0, the variables are uncorrelated, the lines of regression become perpendicular to each other.
8. If r= +1, the two lines of regression either coincide or parallel to each other.
9. Angle between the two regression lines is $θ= tan^{-1}\frac{⃒m1-m2⃒}{⃒1+m1.m2⃒}$

Where m1 and m2 are the slopes of the regression lines X on Y and Y on X respectively

1. The angle between the regression lines indicates the degree of dependence between the variables

# **Difference between Correlation and Regression:**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Correlation** | **Regression** |
| 1. | Correlation is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction. | Regression means going back and it is a mathematical measure showing the average relationship between two variables |
| 2. | Both the variables X and Y are random variables | Here X is a random variable and Y is a fixed variable. Sometimes both the variables may be random variables. |
| 3. | It finds out the degree of relationship between two variables and not the cause and effect of the variables. | It indicates the causes and effect relationship between the variables and establishes functional relationship. |
| 4. | It is used for testing and verifying the relation between two variables and gives limited information. | Besides verification it is used for the prediction of one value, in relationship to the other given value. |
| 5. | The coefficient of correlation is a relative measure. The range of relationship lies between –1 and +1 | Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable. |
| 6. | There may be spurious correlation between two variables. | In regression there is no such spurious regression. |
| 7. | It has limited application, because it is confined only to linear relationship between the variables. | It has wider application, as it studies linear and non- linear relationship between the variables. |
| 8. | It is not very useful for further mathematical treatment. | It is widely used for further mathematical treatment. |
| 9. | If the coefficient of correlation is positive, then the two variables are positively correlated and vice-versa. | The regression coefficient explainsthat the decrease in one variable is associated with the increase in the other variable. |

**Example 2: *If 2 regression coefficients are b1=*** $\frac{4}{5}$ ***and b2 =*** $\frac{9}{20} $***What would be the value of r?***

**Solution:** The correlation coefficient, r = $\pm \sqrt{b1.b2}$ = $\sqrt{\frac{4}{5} .\frac{9}{20}}$ = $\sqrt{\frac{36}{100}}$ = $\frac{6}{10}$ = 0.6

# **Example 3: *Given b1=*** $\frac{15}{8}$ ***and b2 =*** $\frac{3}{5}$ ***Find the value of r.***

**Solution:** The correlation coefficient, r = $\pm \sqrt{b1.b2}$ = $\sqrt{\frac{15}{8} .\frac{3}{5}}$ = $\sqrt{\frac{9}{8}}$ = 1.06

It is not possible since *r*, cannot be greater than one. So the given values are wrong.

# **Example 4:**

***Compute the two regression equations from the following data.***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***X*** | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** |
| ***Y*** | ***2*** | ***3*** | ***5*** | ***4*** | ***6*** |

***If x =2.5, what will be the value of y?***

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y | *x* *X* *X̅* | *y* *Y* *Y* | x2 | y2 | xy |
| 1 | 2 | -2 | -2 | 4 | 4 | 4 |
| 2 | 3 | -1 | -1 | 1 | 1 | -1 |
| 3 | 5 | 0 | 1 | 0 | 1 | 0 |
| 4 | 4 | 1 | 0 | 1 | 0 | 0 |
| 5 | 6 | 2 | 2 | 4 | 4 | 4 |
| 15 | 20 | 20 |  | 10 | 10 | 9 |

Mean of X (X̅) = $\frac{\sum\_{}^{}X}{n}$ = $\frac{15}{5}$ = 3

Mean of Y (Y̅) = $\frac{\sum\_{}^{}Y}{n}$ = $\frac{20}{5}$ = 4

Regression Co efficient of Y on X

byx = $\frac{\sum\_{}^{}xy}{∑x2}$ = $\frac{9}{10}$ = 0.9

Hence regression equation of Y on X is

*Y* *Y* *byx* ( *X* *X* )

= 4 + 0.9 ( *X* – 3 )

= 4 + 0.9X – 2.7

=1.3 + 0.9X

when X = 2.5

Y = 1.3 + 0.9 × 2.5

= 3.55

Regression co efficient of X on Y

bxy = $\frac{\sum\_{}^{}xy}{∑y2}$ = $\frac{9}{10}$ = 0.9

So, regression equation of X on Y is

*X* *X* *bxy* (*Y* *Y* )

= 3 + 0.9 ( Y – 4 )

= 3 + 0.9Y – 3.6

= 0.9Y - 0.6

# **Example 5:**

***Obtain the equations of the two lines of regression for the data given below:***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***X*** | ***45*** | ***42*** | ***44*** | ***43*** | ***41*** | ***45*** | ***43*** | ***40*** |
| ***Y*** | ***40*** | ***38*** | ***36*** | ***35*** | ***38*** | ***39*** | ***37*** | ***41*** |

# **Example 6: *In a correlation study, the following values are obtained***

|  |  |  |
| --- | --- | --- |
|  | ***X*** | ***Y*** |
| ***Mean*** | ***65*** | ***67*** |
| ***S.D*** | ***2.5*** | ***3.5*** |

***Co-efficient of correlation = 0.8***

***Find the two regression equations that are associated with the above values.***

 **Solution:**

Given that X̅ = 65; Y̅ = 67; σX = 2.5; σY = 3.5 and r = 0.8

**Regression Equation of Y on X is**

Y = Y̅ + byx(X- X̅) = 67 + r. $\frac{σY}{σX}$ (X - 65) = 67 + 0.8× $\frac{3.5}{2.5}$ (X - 65) = 67+1.12(X - 65)

Y = 67 + 1.12X – 72.8

Y= - 5.8 + 1.12X

**Regression Equation of X on Y is**

 X = X̅ + bxy (Y- Y̅) = 65 + r. $\frac{σx}{σy}$ (Y - 67) = 65 + 0.8× $\frac{2.5}{3.5}$ (Y - 67) = 65 + 0.57(Y - 67)

 X = 65 + 0.57Y – 38.28

 X = 26.72 + 0.57Y