CHAPTER-VI

REGRESSION ANALYSIS

The term ñregression ñliterally means ñstepping back towards the averageò. It was first used by a British biometrician, Sir Francis Galton (1822-1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to ñregressò or ñstepbackò to the average population height.

By regression we mean average relationship between two or more variables. One of these variables is called the dependent variable or response variable, and the other variable is the independent or the explanatory variable. If the explanatory variables are two or more then it is called the multiple regression analysis. Regression analysis can be further divided into linear and non-linear. In the linear regression, the dependent variable varies at a constant rate with a given change in the independent variable, the constant rate of change can be in absolute terms or in terms of percentage. In the non linear regression, the dependent variable changes at varying rates with a given change in the explanatory variable.

6.1 Linear Regression:

Linear regression line is one which gives the best estimate of one variable (Y) for any given value of the other variable (X). It should be noted that the two regression lines i.e. one of Y on X and another of X on Y cut each other at the point of average of X and Y. The regression equation of Y on X can be expressed as $Y = a + b X + e$, where Y is dependent variable and X is an independent variable, óaô is the intercept which the regression line makes with the y-axis, óbô is the slope of the line and e_i are the random error i.e. the effect of some unknown factors. The values of a and b are obtained by the method of least squares by which we find a and b such that the errors sum of squares $\Sigma e_i^2 = \Sigma(y_i - a-bx_i)^2$ is minimized. Mathematically it is minimized by differentiating it partially w.r.t. parameters to be estimated and putting them equal to zero. The two normal equations will be obtained as under:

$$\times Y = n a + b \times X \quad \text{---- (i)} , \qquad \times XY = a\times X + b\times X^2 \quad \text{---- (ii)}$$

Solving these two normal equations we can get the estimated values of a and b as

$$a = \bar{Y} - b\bar{X} \quad \text{and}$$

$$b_{yx} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n}$$

The regression equation Y on X becomes

$$Y = a + b_{yx} X$$

$$= Y - b_{yx}\bar{X} + b_{yx} X = Y + b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - \bar{Y} = b_{yx}(X - \bar{X}) \text{ where } b_{yx} \text{ is the regression coefficient of Y on X}$$

Similarly the regression equation of X on Y can be expressed as $X = a + b_{xy} Y$ and is obtained by interchanging X and Y, the normal equations will be:

$$\times X = n\,a + b \times Y \text{ ---- (i)}, \qquad \times XY = a \times Y + b \times Y^2 \text{ ---- (ii)}$$

and can again be solved for getting the estimated values of a and b, i.e.

$$a = \bar{X} - b\bar{Y} \quad \text{and}$$

$$b_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum Y_i^2 - (\sum Y_i)^2/n}$$

Similarly regression equation of X on Y can directly be written as:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

b_{xy} is the regression coefficient of X on Y.

Shortcut Method:

Since regression coefficients are independent of change of origin and not of change of scale, therefore, short cut method may be used as follows:

Consider $dx = X - A$ and $dy = Y - B$

then

$$b_{yx} = \frac{\sum dxdy - \sum dx \sum dy/n}{\sum dx^2 - (\sum dx)^2/n}$$

and $$b_{xy} = \frac{\sum dxdy - \sum dx \sum dy/n}{\sum dy^2 - (\sum dy)^2/n}$$

Regression Equation in terms of Correlation Coefficient:

The regression coefficients can be written in terms of correlation coefficient as follows:

$b_{yx} = r\sigma_y/\sigma_x$ and $b_{xy} = r\sigma_x/\sigma_y$

and the regression equations of Y on X and X on Y can be expressed as

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

Testing the Significance of an observed Regression Coefficient:

Consider the null hypothesis Ho: $b = b_0$

and the alternative hypothesis H_1: $b \neq b_0$

Choose a suitable level of significance, say $\alpha = 0.05$

Test statistic is $t = \frac{b - b_0}{SE(b)}$ follows t-distribution with (n-2) d.f.

Where $SE(b) = \frac{s^2}{s_{xx}}$ and $s^2 = \frac{SSE}{n-2} = \frac{s_{yy} - b\,s_{xy}}{n-2}$

If $|t_{cal}| \geq t_{\alpha/2,\, n-2}$ then we reject H_0.

Properties of Regression Coefficients:

i) Correlation coefficient is the geometric mean between two regression coefficients, i.e. $r = \pm\sqrt{b_{yx}\, b_{xy}}$

ii) If one of the regression coefficients is greater than one, the other must be less than one.

iii) Arithmetic mean of the regression coefficients is greater than the correlation coefficient.

iv) Regression coefficients are independent of change of origin but not of scale.

v) Both the regression coefficients are of the same sign and this sign depends on the sign of covariance.

Difference between Correlation and Regression:

- The coefficient of correlation measures the degree of linear relationship between two variables where as the regression coefficient gives the average change in dependent variable corresponding to a unit change in independent variable.
- The coefficient of correlation lies from -1 to 1. This can never exceed unity while the regression coefficient can exceed unity.
- The coefficient of correlation is always symmetrical for any two variables (r_{xy}= r_{yx}) but for the regression coefficient it is not so i.e. b_{yx} is not equal to b_{xy}.
- The coefficient of correlation is independent of change of scale and shift of origin but the regression coefficient is independent of the shift of origin only but not of scale.
- In regression analysis, the variables have cause and effect relation which is not required in correlation analysis.
- Correlation analysis is confined to study of the linear relationship between the variables and thus have limitations, while the regression analysis has much wider applications as it can study linear and non linear relationship between the two variables.
- Correlation coefficient has no unit while regression coefficient has the unit of dependent variable. It indicates the amount of change in dependent variable as per unit change in independent variable.

6.2 Non-linear Regression:

Sometimes it may happen that the original data is not in a linear form but can be reduced to linear form by some simple transformation of variables. This will be illustrated by considering the following curves:

Fitting of a Power Curve: $Y= aX^b$ to a set of n points $(x_1, y_1), (x_2, y_2),$ é .., (x_n, y_n)

Taking logarithm of both sides, we get

$$\log Y = \log a + b \log X$$

$\Rightarrow$ $U= A+ bV$, where $U= \log Y$, $A= \log a$ and $V =\log X$.

This is a linear equation in U and V.

Fitting of Exponential Curves: (i)$Y = ab^x$, (ii) $Y = ae^{bX}$ to a set of n points

i) $Y= ab^x$

Taking logarithm of both sides, we get

$\log Y = \log a + X \log b$

$\Rightarrow$ $U = A + BX$, where $U= \log Y$, $A = \log a$ and $B = \log b$

After solving the normal equations for estimating A and B, we finally get

a = antilog (A) and b= antilog (B)

ii) $Y= ae^{bX}$

$\log Y= \log a + bX \log e = \log a + (b \log e) X$

$\Rightarrow$ $U = A+ BX$, where $U= \log Y$, $A = \log a$ and $B = b \log e$.

After solving the normal equations for A and B, we have

a = antilog (A) and b= B/log e

Example-1: Using the following data, obtain the regression equation of Y on X

X: Additional Expenditure (000 Rs); Y = Sales (Crore Rs)

X:	14	19	24	21	26	22	15	20	19
Y:	31	36	48	37	50	45	33	41	39

Also estimate the sale for additional expenditure of Rs. 25000/-

Solution:

X	**Y**	**$x = (X - \bar{X})$**	**$y = (Y - \bar{Y})$**	**xy**	**x^2**
14	31	-6	-9	54	36
19	36	-1	-4	4	1
24	48	4	8	32	16
21	37	1	-3	-3	1
26	50	6	10	60	36
22	45	2	5	10	4
15	33	-5	-7	35	25
20	41	0	1	0	0
19	39	-1	-1	1	1
$\sum X=180$	**$\sum Y=360$**	**$\sum x=0$**	**$\sum y=0$**	**$\sum xy=193$**	**$\sum x^2=120$**

$\bar{X} = 180/9=20, \quad \bar{Y} = 360/9=40$

$$b_{yx} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = 193/120 = 1.608$$

Regression equation of Y on X is:

$$\hat{Y} - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\hat{Y} - 40 = 1.608 (X - 20)$$

$$\hat{Y} = 1.605 X + 7.84$$

$$\hat{Y}_{25} = 1.605(25) + 7.84 = 40.125$$

Thus estimated sales for additional expenditure of Rs. 25000 is Rs. 40.125 Crores

Example-2: The ages and blood pressures of 9 men are given below:

Age (X)	55	41	36	47	49	42	60	72	63
BP (Y)	142	124	117	127	144	138	154	157	148

i) Find the correlation coefficient between X and Y and test for its significance

ii) Find the regression equation of Y and X

iii) Estimate the blood pressure of a man of 40 years

Solution:

Age	BP	(X-50)		(Y-139)		
X	Y	dx	dx^2	dy	dy^2	dxdy
55	142	5	25	3	9	15
41	124	-9	81	-15	225	135
36	117	-14	196	-22	484	308
47	129	-3	9	-10	100	30
49	144	-1	1	5	25	-5
42	138	-8	64	-1	1	8
60	154	10	100	15	225	150
72	157	22	484	18	324	396
63	148	13	169	9	81	117
Total 465	**1253**	**15**	**1129**	**2**	**1474**	**1154**

i) Coefficient of correlation is given by

$$r = \frac{n\sum dxdy - \sum dx \sum dy}{\sqrt{n\sum dx^2 - (\sum dx)^2}\sqrt{n\sum dy^2 - (\sum dy)^2}}$$

$$= \frac{9(1154) - (15)(2)}{\sqrt{9(1129) - (15)^2}\sqrt{9(1474) - (2)^2}}$$

$$= \frac{10386 - 30}{\sqrt{10161 - 225}\sqrt{13266 - 4}} = \frac{10356}{11479} = 0.902$$

Test of Significance

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.902\sqrt{7}}{\sqrt{1-(0.902)^2}} = 5.53$$

Since $|t_{cal}| > t_{0.025,\ 7} = 2.305$. Hence there is positive and significant correlation in the population.

ii) The regression equation of Y on X is

$$\hat{Y} = \overline{Y} + \hat{b}_{yx}(X - \overline{X})$$

$$\overline{X} = \frac{465}{9} = 51.7\ ;\ \overline{Y} = \frac{1253}{9} = 139.2\ ;$$

$$\hat{b}_{yx} = \frac{n\Sigma dxdy - \Sigma dx \Sigma dy}{n\Sigma dx^2 - (\Sigma dx)^2} = 10356/9936 = 1.04$$

Hence $\hat{Y}$ =139.2+1.04(X-51.7) = 85.31 + 1.04 X

iii) For X = 40 $\hat{Y}$ = 85.31 + 1.04 (40) = 126.91

Hence, the estimated blood pressure for a man of 40 years is 127.

Example-3: For the following results

Variance of X = 9

Regression equations 8X ï 10Y + 66 = 0; 40X ï 18Y = 214, Find

i) The mean value of X and Y

ii) Coefficient of correlation between X and Y, and

iii) Standard deviation of Y

Solution:

i) Since regression lines pass through $(\overline{X}, \overline{Y})$, therefore, we have

$8\overline{X}$ ï $10\overline{Y}$ = -66 (1)

$40\overline{X}$ ï $18\overline{Y}$ = 214 (2)

Multiplying equation (1) by 5

$40\overline{X}$ -50$\overline{Y}$ = -330

$40\overline{X}$ -18$\overline{Y}$ = 214

-32$\overline{Y}$ = -544 hence $\overline{Y}$ = 17

Putting the value of $\overline{Y}$ in equation (1)

$8\overline{X}$ -10(17) = -66 we get $\overline{X}$ = 13

ii) Coefficient of correlation between X and Y

Let (i) is the regression equation of X and Y

8X = 10Y ï 66

$$X = \frac{10}{8}Y - \frac{66}{8} \text{ or } b_{xy} = \frac{10}{8}$$

From equation (2) ï 18Y = 214 ï 40X

$$Y = -\frac{214}{18} + \frac{40}{18}X \text{ or } b_{xy} = \frac{40}{18}$$

Since both regression coefficients are greater than 1, our assumption is wrong. Hence equation (1) is regression equation of Y on X.

-10Y = -66 ï 8X

$$Y = \frac{66}{10} + \frac{8}{10}X \qquad \text{or} \qquad b_{yx} = \frac{8}{10}$$

From equation (2) 40X = 214 + 18Y

$$X = \frac{214}{40} + \frac{18}{40}Y \qquad \text{or} \qquad b_{xy} = \frac{18}{40}$$

$$\text{Thus } r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{18}{40} \times \frac{8}{10}} = \sqrt{0.36} = 0.6$$

iii) The standard deviation of Y can be determined from any regression coefficient

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

Substituting the values

$$\frac{18}{40} = 0.6\frac{3}{\sigma_y}, \text{ we get } \sigma_y = 4$$

Example-4: The following data relate to marks obtained by 250 students in Economics and Statistics in an examination:

Subject	**Arithmetic Mean**	**Standard Deviation**
Economics	48	4
Statistics	55	5

Coefficient of correlation between marks in economics and statistics is +0.8. Draw the two lines of regression and estimate the marks obtained by a student in statistics who secured 50 marks in economics.

Solution: Let marks in economics be denoted by X and in statistics by Y.

Regression equation of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

$$\bar{X} = 48, \bar{Y} = 55, \ \sigma_x = 4; \sigma_y = 5, r = 0.8$$

$$X - 48 = 0.8\ (4/5)\ (Y-55)$$

$$X - 48 = 0.64\ (Y-55)$$

$$X = 0.64Y + 12.8$$

Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

$\hat{Y}$ ï 55 = 0.8(5/4) (X ï 48)

$\hat{Y}$ ï 55 = (X ï 48)

$\hat{Y}$ = 7 + X

For X = 50 $\hat{Y}$ = 7 + 50 = 57

Thus for marks in economics equal to 50, estimated marks in statistics shall be 57.

Example-5: Following Statistics were obtained in a study conducted for examination of relationship between yield of wheat and annual rainfall.

	Yield (kg/acre)	**Annual Rainfall (inches)**
Mean :	985.0	12.8
SD :	70.1	1.6

r = 0.52

Assuming a linear relationship between yield and rainfall obtain yield of wheat/acre when the rainfall of 9.2 inches.

Solution: Let the rainfall be denoted by X and yield by Y. The required yield can be obtained from the regression equation of Y on X, which is

$$\hat{Y} - \overline{Y} = r\frac{\hat{u}_y}{\hat{u}_X}(X - \overline{X})$$

$$\hat{Y} - 985 = 0.52\frac{70.1}{1.6}(X - 12.8) = 22.78\ (X - 12.8)$$

or $\hat{Y} - 985 = = 22.78\ X - 291.58$

or $\hat{Y} = 693.42 + 22.78X$

when X = 9.2, $\hat{Y}$ = 693.42 + 22.78 x 9.2 = 903 kg/acre

Example-6: Given

X :	36	28	38	42	44	46	30	34	32	40
Y :	128	156	142	135	177	184	149	191	163	170

i) Calculate the Karl-Pearson correlation coefficient between X and Y and interpret the results.

ii) Obtain the equations for two regression lines and estimate the values of Y for X= 30 and also the value of X for Y = 149

Solution: From the given data we obtain

n = 10, $\Sigma X = 370$, $\Sigma Y = 1595$

$\Sigma X^2 = 14020$, $\Sigma Y^2 = 258445$ and $\Sigma XY = 59274$

$$S_{xx} = \sum X^2 - \frac{(\sum X)^2}{n} = 14020 - \frac{(370)^2}{10} = 330$$

$$S_{yy} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 258445 - \frac{(1595)^2}{10} = 4042.5$$

and $$S_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 59274 - \frac{370 \text{ x } 1595}{10} = 259$$

i) $$r = \frac{S_{XY}}{\sqrt{S_{XX}\ S_{YY}}} = \frac{259}{\sqrt{330 \text{ x } 4042.5}} = \frac{259}{1155} = 0.224$$

Test of significance:

$$t_{cal} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.224}{\sqrt{1-0.224^2}}\sqrt{10-2} = 0.65$$

Let $\alpha = 0.05$ then $t_{0.025,\,8} = 2.31$

Since $t_{cal} < t_{0.025,\,8}$, so correlation between X and Y is statistically non-significant at 5% level.

ii) Let the equation of regression line of Y and X be.

$$Y = a + b_{yx} X$$

Then $$b_{yx} = \frac{S_{xy}}{S_{xx}} = \frac{259}{330} = 0.785$$

$$a = \overline{Y} - b_{yx}\,\overline{X} = 159.5 - 0.785 \text{ x } 37 = 130.46$$

Thus $$Y = 130.46 + 0.785\,X$$

When X = 30, $\hat{Y}$ = 130.46 + 0.785 x 30 = 154.01

Also, let the equation of the line of regression of X and Y be

$\hat{X} = a + b_{xy} Y$

Then $b_{xy} = \frac{S_{xy}}{S_{yy}} = \frac{259}{4042.5} = 0.064$

and $a = \bar{X} - b_{xy} \bar{Y}$ = 37 – 0.064 x 159.5 = 26.79

Then, $\hat{X} = 26.79 + 0.064Y$

When Y = 149, $\hat{X}$ = 26.79 + 0.064 x 149 = 36.33

Example-7: The length of panicles (x) in cm and the number of grains per panicles (y) for 15 plants from a field of paddy are given below. Fit a regression line of y on x and estimate the number of grains per panicle when panicle length in 25 cm.

x:	22.4	23.3	24.1	24.3	23.5	22.3	23.9	24	24.9	20	19.8	22	24.5	23.6	21.1
y:	95	109	133	132	136	116	126	124	137	90	107	108	143	127	92

Solution: From the given data, we have

n = 15, Σx = 343.7, Σy = 1775

Σx^2 = 7911.17, Σy^2 = 214247 and Σxy = 40998.6

$\bar{x}$ = 22.91 and $\bar{y}$ = 118.33

Also, $s_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 214247 - \frac{(1775)^2}{15} = 4205.33$

$s_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 7911.17 - \frac{(343.7)^2}{15} = 35.86$

$s_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 40998.6 - \frac{343.7 \text{ x } 1775}{15} = 327.43$

$b = \frac{s_{xy}}{s_{xx}} = \frac{327.43}{35.86} = 9.13$

$a = \bar{y} - b\,\bar{x} = 118.33 - 9.13 \text{ x } 22.91 = 118.33 - 209.17 = -90.84$

$\hat{y} = -90.84 + 9.13\,x$, $\hat{y}$ when x = 25 is 137.41

Example-8: From 18 pairs of observations on height of plant (X) and their respective produce (Y) following quantities were obtained.

$\overline{X} = 60$, $\overline{Y} = 10$, $\Sigma(X - \overline{X})^2 = 1500$, $\Sigma(Y - \overline{Y})^2 = 60$ and

$\Sigma(X - \overline{X})\ (Y - \overline{Y}) = 180$

Find the two regression coefficients and hence calculate the correlation coefficient between X and Y and test its significance

Solution: $b_{xy} = \dfrac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - \overline{X})^2} = \dfrac{180}{1500} = 0.12$

$b_{yx} = \dfrac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(Y - \overline{Y})^2} = \dfrac{180}{60} = 3.0$

$r_{xy} = \sqrt{b_{yx}\ b_{xy}} = \sqrt{0.12 \times 3} = \sqrt{0.36} = 0.6$

testing the hypothesis H_0 : $\rho = 0$ against H_1 : $\rho \neq 0$

$t_{cal} = \dfrac{r}{\sqrt{1 - r^2}}\sqrt{n-2} = \dfrac{0.6}{\sqrt{1-0.36}}\sqrt{18-2} = \dfrac{0.6 \times 4}{\sqrt{0.65}} = \dfrac{2.4}{0.8} = 3$

table $t_{0.025,\ 16} = 2.12$ $|t_{cal}| > t_{tab}$, therefore we reject H_0

Example-9: For 12 wheat earheads, the characters, length of earhead (x) and grain per earhead (y) were recorded and the following quantities were obtained:

$\overline{x} = 20,\ \overline{y} = 30$

$s_{xx} = 800$, $s_{yy} = 996$ and $s_{xy} = 760$

Fit a regression line and estimate the grains per earhead of length 10 units. Also test the significance of the regression coefficient.

Solution: The equation of the regression line of Y on X given by $\hat{y} = \hat{b}_0 + \hat{b}_1 x$

Given $n = 12,\ \overline{x} = 20,\ \overline{y} = 30$

$s_{xx} = 800,\ s_{yy} = 996$ and $s_{xy} = 760$

Hence, $\hat{b}_1 = \dfrac{s_{xy}}{s_{xx}} = \dfrac{760}{800} = 0.95$

and $\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x} = 30 - 0.95 \text{ x } 20 = 11$

The regression equation of y on x is

$$\hat{y} = 11 + 0.95x$$

when x = 10, $\hat{y}$ = 20.5. Thus we expect 20.5 grains per earhead of length 10 units on an average.

The regression coefficient $\hat{b}_1$ = 0.95, which indicates that for every unit increase of the length of the earhead, we can expect an increase of 0.95 grains per earhead on the average. For testing significance of the regression coefficient, we formulate and test the hypothesis.

H_0 : $b_1 = 0$ against H_1 : $b_1 \neq 0$

To test H_0, we need to obtain an estimate s^2 of σ^2, the variance of the error term. Hence

The sum of squares due to error (SSE) = s_{yy} - $\hat{b}_1$ s_{xy}

= 996 ï 0.95 x 760 = 274

and $s^2 = \dfrac{SSE}{n-2} = \dfrac{274}{10} = 27.4$

The test statistic, t, is calculated as

$$t_{cal} = \frac{\hat{b}_1}{SE(\hat{b}_1)} = \frac{\hat{b}_1}{\sqrt{s^2/S_{xx}}} = \frac{0.95}{\sqrt{27.4/800}} = \frac{0.95}{0.185} = 5.135$$

$t_{0.05,\,10}$ = 2.228

Since $|t_{cal}| > t_{tab}$, therefore, we reject the null hypothesis and conclude that the number of grains per earhead increases significantly with the increase in length of the earhead.

EXERCISES

1. Find the regression equations of X on Y and Y on X from the data given below

   Husbandôs age (X) 26 28 30 31 35

   Wifeôs age (Y) 20 27 28 30 25

   Also calculate the coefficient of correlation.

2. Obtain the two lines of regression and estimate the value of Y if X is 70 and that of X if Y is 90 from the following given data:

	X-series	Y-series
Arithmetic Mean	18	100
Standard deviation	14	20

   Coefficient of correlation between X and Y series = 0.8

3. From 18 pairs of observation on height of papaya plants (X) in inches and their produce (Y) in kg, the following quantities were calculated.

   $\overline{X} = 60$, $\overline{Y} = 10$, $\Sigma(X_i - \overline{X})^2 = 1500$

   $\Sigma (Y_i - \overline{Y})^2 = 60$ $\Sigma (X_i - \overline{X})(Y_i - \overline{Y}) = 180$

   Calculate the regression of yield on the height of plant and test its significance. Also calculate the coefficient of correlation between the two variates and test its significance.