Regression:

Example:

A company located in a certain region makes deliveries. The CEO of the company wants to study the relationship between the distance a shipment must travel and the length of time, in days, it takes the shipment to arrive at its destination. To investigate, a random sample of 10 shipments made last month is selected. The results are as follows:

Shipment	Distance (Miles)	Shipping time (days)
1	656	5
2	853	14
3	646	6
4	783	11
5	610	8
6	841	10
7	785	9
8	639	9
9	762	10
10	762	9

a- Develop an estimated Regression equation to Predict shipping time based on Distance

b- Obtain unexplained & total variation. Also

c- find coefficient of determination

c- Find correlation also

Sol:

Let 'x' be the independent variable (distance).

Let 'y' be the dependent variable (shipping time)

a- estimated regression equation is

$$\hat{Y} = a + bX$$

XY	X²	Y²
3280	430336	25
11942	727609	196
3876	417316	36
8613	613089	121
4880	372100	64
8410	707281	100
7065	616225	81
5751	408321	81
7620	580644	100
6858	580644	81
$\Sigma XY = 68295$	$\Sigma X^2 = 5453565$	$\Sigma Y^2 = 885$

$$\Rightarrow b = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{10(68295) - (7337)(91)}{10(5453565) - (7337)^2}$$

$$= \frac{15283}{704081} = \boxed{0.0217}$$

$$a = \bar{Y} - b\bar{X}$$

$$\bar{Y} = \Sigma Y/n = 91/10 = 9.1$$

$$\bar{X} = \Sigma X/n = \frac{7337}{10}$$

$$= 733.7$$

$$\boxed{\Sigma X = 7337}$$
$$\boxed{\Sigma Y = 91}$$

$$a = 9.1 - (0.0217)(733.7)$$

$$= 9.1 - 15.9213$$

$$a = -6.82$$

$$\hat{Y} = -6.82 + 0.0217 X.$$

b- unexplained variation: & total variation:

unexplained variation $= \Sigma(Y - \hat{Y})^2$

$$= \Sigma Y^2 - a\Sigma Y - b\Sigma XY$$

$$= 885 - (-6.82)(91) -$$
$$0.0217(68295)$$

$$= 1505.62 - 1482.0015$$

$$= 23.6185$$

Total variation: $\Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - \dfrac{(\Sigma Y)^2}{n}$

$$= 885 - \dfrac{(91)^2}{10}$$

$$= 56.9$$

Co-efficient of determination:

$$r^2 = 1 - \dfrac{\text{unexplained variation}}{\text{Total variation}}$$

$$= 1 - \dfrac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2} = 1 - \dfrac{23.6185}{56.9}$$

$$r = \sqrt{0.58} = 0.7616$$

$$= 0.58 = 58\%. \quad r = \dfrac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

58% of variability in dependent variable (time) is explained due to its relationship with independent variable (distance).

Multiple Regression

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots b_k X_k.$$

$b_1, b_2 \cdots b_k$ (Regression coefficients)

$X_1, X_2 \cdots X_k$ (Independent variables)

Y (dependent variables).

Prediction of CO_2 emission based on based on engine size & number of cylinders in a car.

It means guessing how much CO_2 (Carbon dioxide) a car will produce by looking at two things (size of its engine & no of cylinders).

How much pollution a car might make based on type of engine. It has & how big it is.