Simple Linear Regression (Least Square Method)

Approach - 2

In this approach calculation of θ_1 is different. Formula is dependent on Correlation coefficient. (r) (Pearson)

$$\boxed{\text{slope } \theta_1 = r \times \frac{\sigma_y}{\sigma_x}} \quad \text{—— (A)}$$

where, r is pearson's correlation

$\sigma_y \rightarrow$ standard deviation of y.

$\sigma_x \rightarrow$ standard deviation of x.

(r) Correlation coefficient :- It is a measure of how related two variables are in the range of -1 to 1.

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

where x is independent variable

y is dependent variable.

Example :- Approach - 2

Consider the same example of restaurant. We can compare the error occurred for both the approaches.

i	x	y	$x_i-\bar{x}$	$y-\bar{y}$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	100	5	−58	−2	116	3364	4
2	150	7	−8	0	0	64	0
3	200	10	42	−3	126	1764	9
4	180	5	22	−2	−44	484	4
5	160	8	2	1	2	4	1
	$\Sigma x =$ $\bar{x}=158$	$\Sigma y=$ $\bar{y}=7$			$\Sigma=200$	$\Sigma=5680$	$\Sigma=18$

calculate correlation coefficient (r)

$$r = \frac{\Sigma (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma (x_i-\bar{x})^2 \cdot \Sigma (y_i-\bar{y})^2}}$$

$$= \frac{200}{\sqrt{5680 \times 18}} = \frac{200}{319.74} = 0.625$$

$$\boxed{r = 0.625}$$

The value of r indicate that the variable x and variable y are positively correlated.

Now, calculate standard deviation for variables x & y.

$$\sigma x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N-1}} \Rightarrow \sqrt{\frac{5680}{4}}$$

$$\boxed{\sigma x = 37.68}$$

$$\sigma y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{N-1}} \Rightarrow \sqrt{\frac{18}{4}}$$

$$\boxed{\sigma y = 2.12}$$

Now, calculate slope θ_1

$$\theta_1 = r \cdot \frac{\sigma y}{\sigma x}$$

$$= 0.625 \cdot \frac{2.12}{37.68}$$

$$\boxed{\theta_1 = 0.0351}$$

Using second approach we got the same slope for the model. Therefore, other values like θ_2 and Error remains the same.

Correlation Coefficient

The extent to which the direction and size of deviation from mean in one variable are related to the direction and size of deviation from the mean in another variable.

if $r = 0$:- No correlation exist between variables

$r = -1$:- Strongest Correlation (-ve)

$r = 1$:- strongest Correlation (+ve)

Correlation can be calculated using Z-score.

$$r = \frac{\sum_{i=1}^{N} z_i^x z_i^y}{N}.$$

where, z_i^x → Measures how many standard deviations a data point i is from the mean of a data set. (x variable)

z_i^y → related to y variable.

N → no. of data points.

$$Z^x\text{-score}(i) = \frac{x_i - \bar{x}}{\sigma x} = z_i^x$$

$$z_i^y = \frac{y_i - \bar{y}}{\sigma y}$$

⇒ Calculation of 'r' using Z-score.

__Example__ : Same used in earlier approaches.

i	x	y	Z_x	Z_y	$Z_x \cdot Z_y$
1	100	5	−1.53	−0.94	1.43
2	150	7	−0.21	0	0
3	200	10	1.11	1.41	1.56
4	180	5	0.58	−0.94	−0.54
5	160	8	0.05	0.47	0.02
					$\Sigma = 2.47$

$\bar{x} = 158$
$\bar{y} = 7$ } taken from earlier example.

$$\sigma x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N}} = 37.68$$

$$\sigma y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{N}} = 2.12$$

$$i \Rightarrow 1 \quad z_i^x = \frac{100 - 158}{37.68} = -1.53$$

$$i \Rightarrow 1 \quad z_i^y = \frac{5 - 7}{2.12} = -0.94$$

Similarly, calculate all the values of Z-score with respect to x & y.
All the values of Z-scores are given in the above table.

$$r = \frac{\sum_{i=1}^{N} Z_x Z_y}{N - 1}$$

N−1, if std. deviation is calculated by N−1.

$$r = \frac{2.47}{4} = \underline{\underline{0.617}}$$

It is same as earlier two methods.

Now θ_1 can be calculated by

$$\boxed{\theta_1 = r \cdot \frac{\delta y}{\delta x}}$$

It will be same as earlier two methods.

<u>Additional Examples</u>.