PN-notes

NOTES ON LINEAR REGRESSION

Topics:

- Linear regression equation
- formulae list
- Pearson's Correlation Geffiant
- making trendline.
- Evaluating Regression line using R-squared (R^2) (Co-efficient of Determination)
- Adjusted R^2
- Exercises

Simple linear regression

PN_NOTES

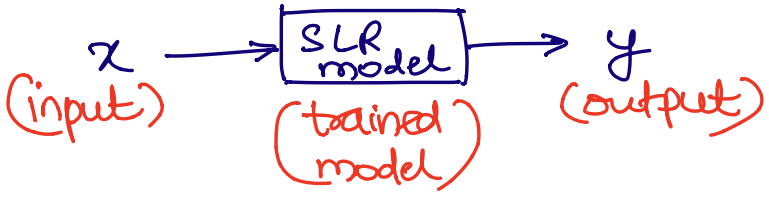
Simple linear regression allows us to predict the linear relationship between two variables.

So think in this way,

we are dealing with two variables,

x : feature : independent variable
y : label : dependent variable

So, in this case, take the value of x and predict y.



Now let's see the formulaes behind SLR.

⇒ Linear Regression function ;

PK NOTES

$$y = b_0 + b_1 x_1$$

b_0 : y intercept of the line

b_1 : slope of the Regression line for x_1.

⇒ To calculate the slope of the line (b_1) we have two formulae,

Formula 1 :-

$$b_1 = r \frac{S_y}{S_{x_1}}$$

S_y : standard deviation of y

S_{x_1} : standard deviation of x_1

r : Pearson's Correlation Coefficient

where,

Pearson's correlation co-efficient (r) can be calculated

x : value of x (feature)

$\bar{x}$: mean of x

y : value of y

$\bar{y}$: mean of y

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum (x - \bar{x})^2\right]\left[\sum (y - y^2)\right]}}$$

standard deviation of y and x is,

$$s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$$

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

number of samples

Formula 2 :

$$b1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

where,

$x \rightarrow$ feature , $y \rightarrow$ label

$\bar{x} \rightarrow$ mean of the feature

$\bar{y} \rightarrow$ mean of the label.

$\Rightarrow$ To calculate y-intercept of the line (b_0)

$$b_0 = \bar{y} - m\bar{x}$$

$\bar{y}$ ← mean of the y

m ← slope of line for x

$\bar{x}$ ← mean of the x

Let's solve one problem.

Q1 → Given the below dataset, calculate and plot the regression line.

Temperature (°C)	Electricity consumption (kW)
20	10
30	25
40	35
50	40

Soln: We will solve this problem using

Method 1 → using Pearson's Correlation Coefficient.

The Pearson's Correlation Coefficient

formula is

$$r = \frac{\sum [(x-\bar{x})(y-\bar{y})]}{\sqrt{\left(\sum (x-\bar{x})^2\right)\left(\sum (y-\bar{y})^2\right)}}$$

Lets calculate all the values required to substitute in above formula,

$$\bar{x} \text{ (mean of } x) = \frac{\sum_{i=1}^{4} x_i}{4} = \frac{20+30+40+50}{4} = 35$$

$$\bar{y} \text{ (mean of } y) = \frac{\sum_{i=1}^{4} y_i}{4} = \frac{10+25+35+40}{4} = 27.5$$

$$\boxed{\bar{x} = 35}$$

$$\boxed{\bar{y} = 27.5}$$

P.T.O

PN_NOTES

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
20	10	-15	-17.5	262.5	225	306.25
30	25	-5	-2.5	12.5	25	6.25
40	35	5	7.5	37.5	25	56.25
50	40	15	12.5	187.5	225	156.25
				500	500	525

$\sum(x-\bar{x})(y-\bar{y})$ $\quad \sum(x-\bar{x})^2$ $\quad \sum(y-\bar{y})^2$

$$\therefore r = \frac{500}{\sqrt{500 \times 525}} = 0.97590 \approx 0.976$$

$$\therefore r = 0.976$$

from the above result, we can say there exists positive correlation between feature and label.

Lets calculate the standard deviation of x & y

n = number of samples = 4

$$S_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{500}{3}} = 12.9$$

$$S_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}} = \sqrt{\frac{525}{3}} = 13.228$$

$\boxed{r = 0.976}$ $\boxed{S_x = 12.9}$ $\boxed{S_y = 13.228}$

Lets calculate the slope of the line.

$$\therefore b_1 \text{ (slope of } x) = r \frac{S_y}{S_x}$$

$$= (0.976)\left(\frac{13.228}{12.9}\right)$$

$$= 1 //$$

$\boxed{b_1 = 1}$ ——— (1)

PN Notes

Lets calculate the intercept of line

$$b_0 = \bar{y} - m\bar{x}$$

$$= 27.5 - 1(35)$$

$$= -7.5 //$$

$$\boxed{b_0 = -7.5} \quad \text{---------} \quad (2)$$

So from (1) & (2) the simple linear regression equation will be

$$y = b_0 + b_1 x_1$$

$$\boxed{\therefore \quad y = -7.5 + (1)\, x_1} \quad (3)$$

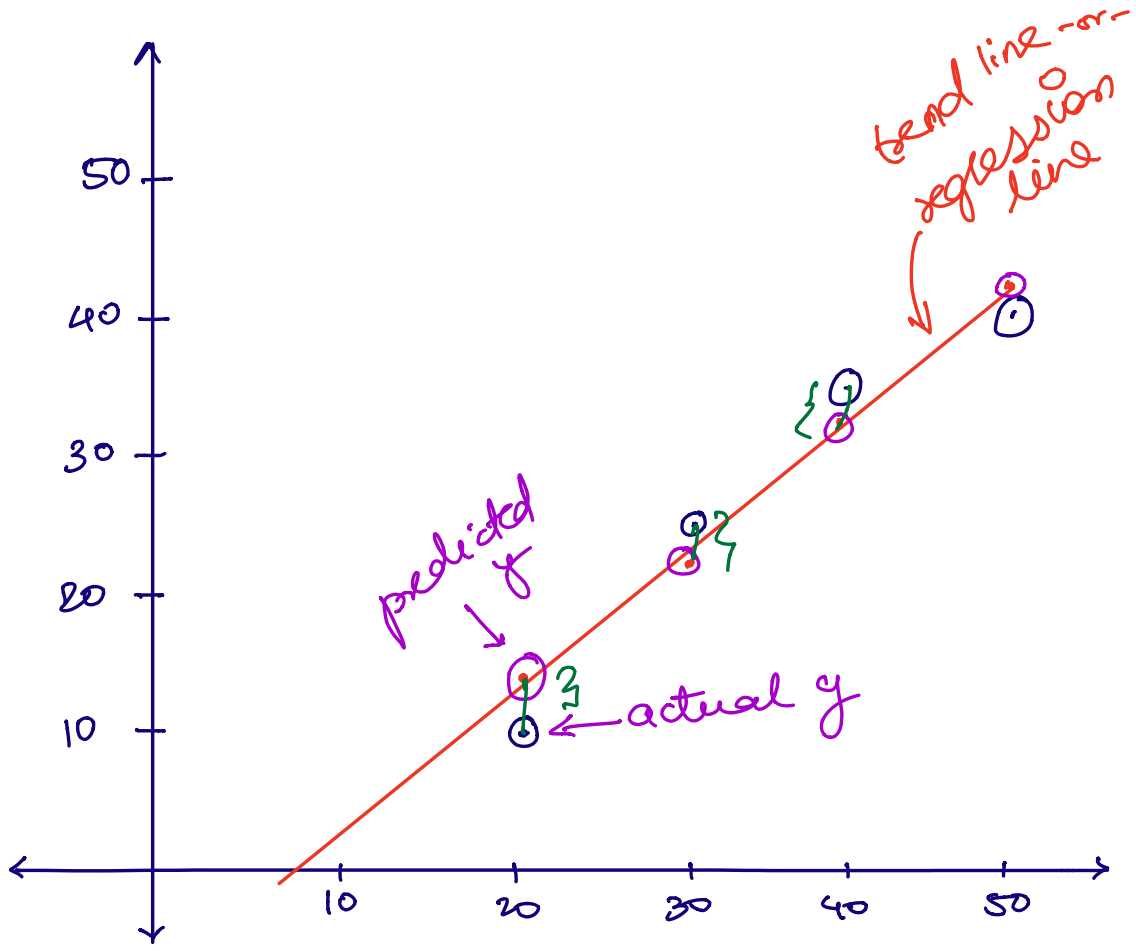
← equation of the line.

Now lets create a trendline.

PN notes

x	y	$\hat{y}$
20	10	12.5
30	25	22.5
40	35	32.5
50	40	42.5

Calculate $\hat{y}$ using ③



PN NOTES

Evaluating the Regression line

* METHOD - 1 → R-squared (co-efficient of determination)

The motto of R-squared is to figure out how good the line is fitted ensuring minimal errors of points (%).

To find R^2, you need to ask the following questions:

WHAT PERCENTAGE OF THE VARIATION IN y IS DESCRIBED BY THE VARIATION IN x

y ← label

x ← feature

∴ Total variations in y

= Squared error from mean of y

$= (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad //$$

This is also called SUM OF SQUARES OF RESIDUALS

PN NOTE

$$\therefore \text{Variance of } y = \frac{\text{Total variations in } y}{n}$$

$$= \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} \quad //$$

What percentage of total variation is not described by variation in x.

$$R^2 = 1 - \frac{SE_{line}}{SE_{\bar{y}}}$$

The above equation is also called CO-EFFICIENT OF DETERMINATION

PN Notes

If R^2 is greater value i.e. closer to 1, then the line is the best fitted line.

① SE_{line} is small $\Longrightarrow$ line is a good fit

$\Downarrow$

R^2 is close to 1.

② SE_{line} is large $\Longrightarrow$ line is not a good fit

$\Downarrow$

R^2 is close to 0.

Lets see an example to calculate R^2

x	y	$\hat{y}$	$(y-\hat{y})^2$ Squared errors with line	$(y-\bar{y})^2$ Squared from mean y
-2	-3	-2.1905	0.6553	10.5625
-1	-1	1.2143	0.0459	1.5625
1	2	0.7381	1.5924	3.0625
4	3	3.6666	0.4444	7.5625

PN notes

from table,

$$\sum_{i=1}^{n}(y-\hat{y})^2 = 2.738 \Rightarrow SE_{line}$$

$$\sum_{i=1}^{n}(y-\bar{y})^2 = 22.75 \Rightarrow SE_{\bar{y}}$$

$$\therefore R^2 = 1 - \left(\frac{SE_{line}}{SE_{\bar{y}}}\right)$$

$$= 1 - \left(\frac{2.738}{22.75}\right)$$

$$= 1 - (0.1203)$$

$$= 0.8797 \approx 0.88$$

$$\boxed{R^2 = 0.88}$$

Since R^2 is very near to 1, therefore its a good fit line.

PROBLEMS FOUND IN R^2 metric:

PWNotes

R^2 increases with increase in features (predictors) (independent variables) added to the model. This property of R^2 will MISLEAD the fact the model is best fit.

The above misleading also happens when Polynomial features are introduced.

Solution:

The above problem can be solved using

ADJUSTED R^2 metric

$$\text{ADJUSTED } R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

where,

n is number of observations (data pts)

k is no of independent variables in your model excluding the constant.