Supervised Regression ML-Algo →

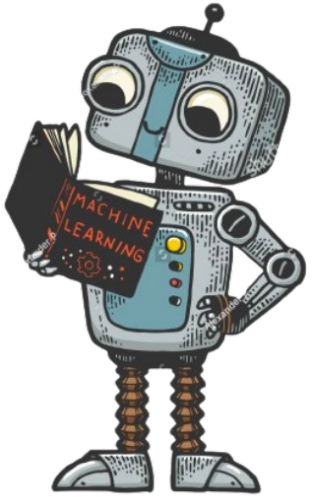
1. L.R
2. D.T.R
3. R.F.R
4. KNNR

ML: Linear Regression-1

Supervised ML-Algo

we may use this algo when the data comes with the target feature

Target feature → continuous data-type



Summary

Branches of ML

- **Supervised** (If data comes with the target)
  - **Regression** (continuous)
  - **Classification** (categorical)
- **unsupervised** (If data comes only with features)

Introduction to machine Learning

→ Experience (Historical Data)
→ Task (objective)
→ performance (Evaluation)

Independent var. (x)

age	Exp	Sex	Salary

→ Dependent variable (y) traget

"Historical Data"

Task → predict "salary"

continuous

Regression

→ ML-Algo → Model

(Learn the relationship between X's and y)

(mathematical eq. which connect X's with y)

$y = f(x)$

age = 27
exp = 7
sex = male

(new employee)

predictions

predicted Salary

x1 age	x2 income	x3 # of cars	y Buy/not Buy
			Y
			N
			Y
			N

Task → predict if a customer will purchase the product or not?

target → categorical

→ ML-Algo → ML-Model → predictions

supervised classification algo

$y = f(x)$ → Yes/NO

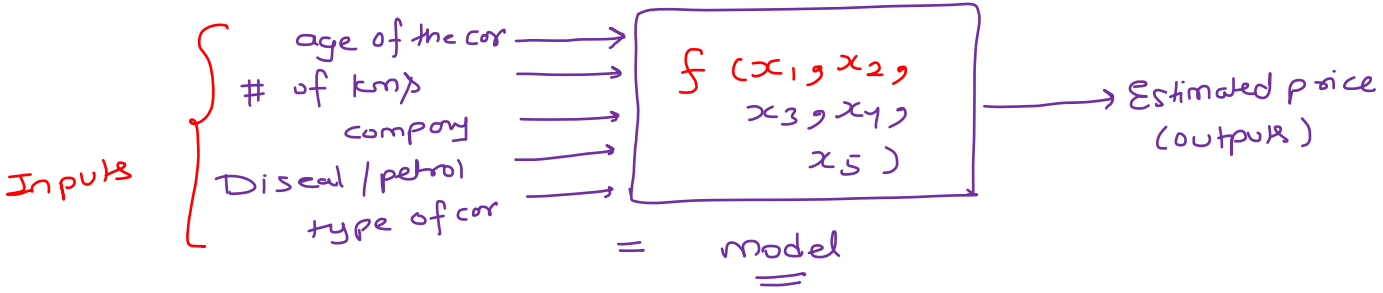
age = 27
income = 70k
# of cars = 1

Classification

Motivation

Can you think of the factors on which the price of a used car depends on?

1. Age of the car
2. # of kms
3. Company
4. Diseal / petrol
5. Type of car (Hatchback / sedan / SUV)



Which ML Algorithm To Choose?

2019

Cars24 [Last 8 years]

{Experts}

	x_1 age	x_2 # of kms	x_3 Model	x_4 P/D	y Price
C1					
C2					
C3					
⋮					

↳ continuous

(Historical)

ML-Algo

Linear-Reg

model

price = f(age, # of kms, model, P/D)

Data of new car

Estimated price

{ {x_1, x_2, x_3, x_4} → y } → multiple linear regressional

Independent feature

Mental Model

{multiple linear Regression}

1. Age of car
2. # of kms
3. company
4. Disel / petrol
5. Type of car

$\longrightarrow$ price

$$\text{price} = \underset{0.8}{(W_1)}(\text{age}) + \underset{0.2}{(W_2)}(\#\text{ of kms}) + \underset{0.01}{(W_3)}(\text{company}) + \overset{0.001}{(W_4)}(\text{Disel/petrol}) + \underset{0.005}{(W_5)}(\text{Type of car})$$

info
- $\longrightarrow$ age = 2
- $\longrightarrow$ # of kms = 20k
- $\longrightarrow$ company = Tata (EV)
- $\longrightarrow$ car = EV
- $\longrightarrow$ SUV

3 L
2.9 L
3.25 L
$\boxed{3.25\text{ L}}$
$\boxed{4\text{ L}}$

{ Equation of the line }

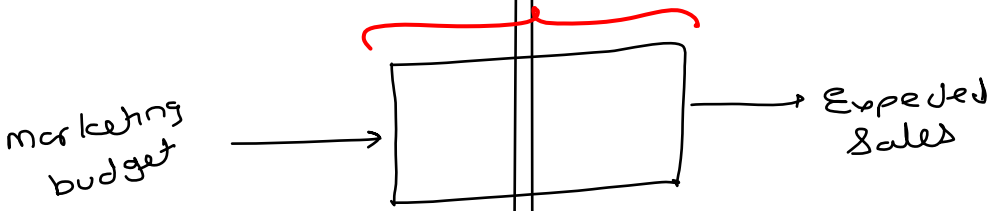
Simple Linear Regression

one independent variable → y

x — y

marketing_budget(thousands)	actual_sales(millions)
187.86	12.14
138.13	12.33
177.89	8.41
181.13	8.9
151.5	12.57
106.87	11.3
160.94	11.66
140.09	10.93
132.58	11.99

What we want to predict → Sales

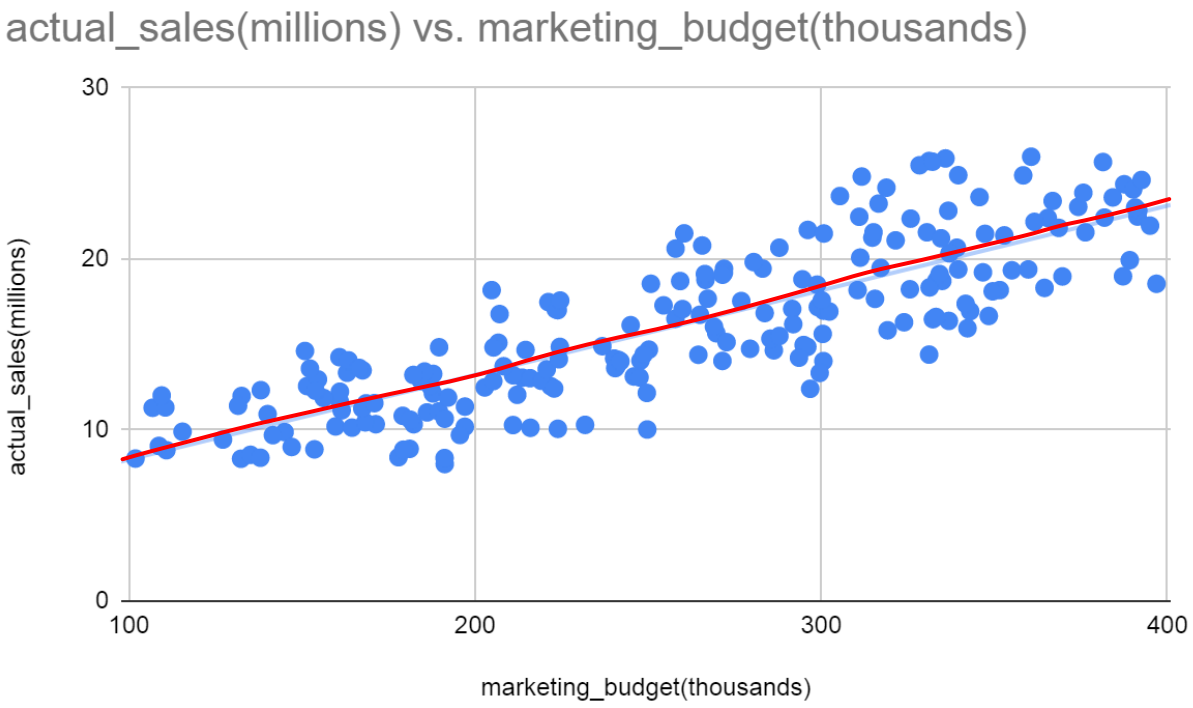


$(y = w_0 + w_1 x)$

$[\text{Sales} = w_0 + w_1 (\text{marketing budget})]$

Best value of w_0, w_1

Simple Linear Regression



Best fit line
$w_0, \bar{w}$

Simple Linear Regression

$(w_0, w_1) \longrightarrow$ coefficient

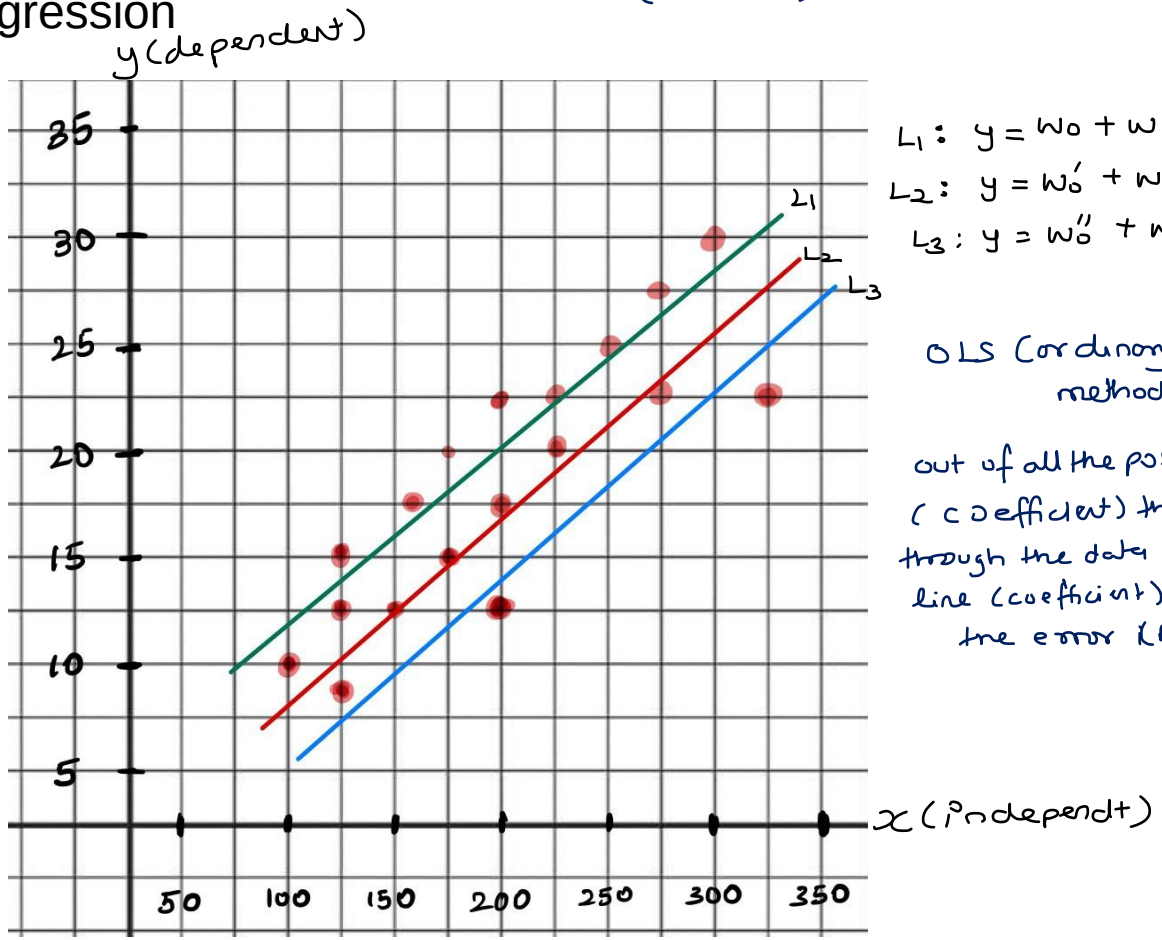
Task $\longrightarrow$ choose the best value of $w_0, \bar{w}$

Question $\rightarrow$ on what basis?

$y_a \lessgtr y_p$

or

$(y_a - y_p) \lessgtr 0$



$L_1: y = w_0 + w_1(x)$ RSS = 500

$L_2: y = w_0' + w_1'(x)$ RSS = 200

$L_3: y = w_0'' + w_1''(x)$ RSS = 700

OLS (ordinary Least Square method)

out of all the possible lines (coefficient) that can pass through the data choose that line (coefficient) for which the error (RSS) is minimum

Simple Linear Regression

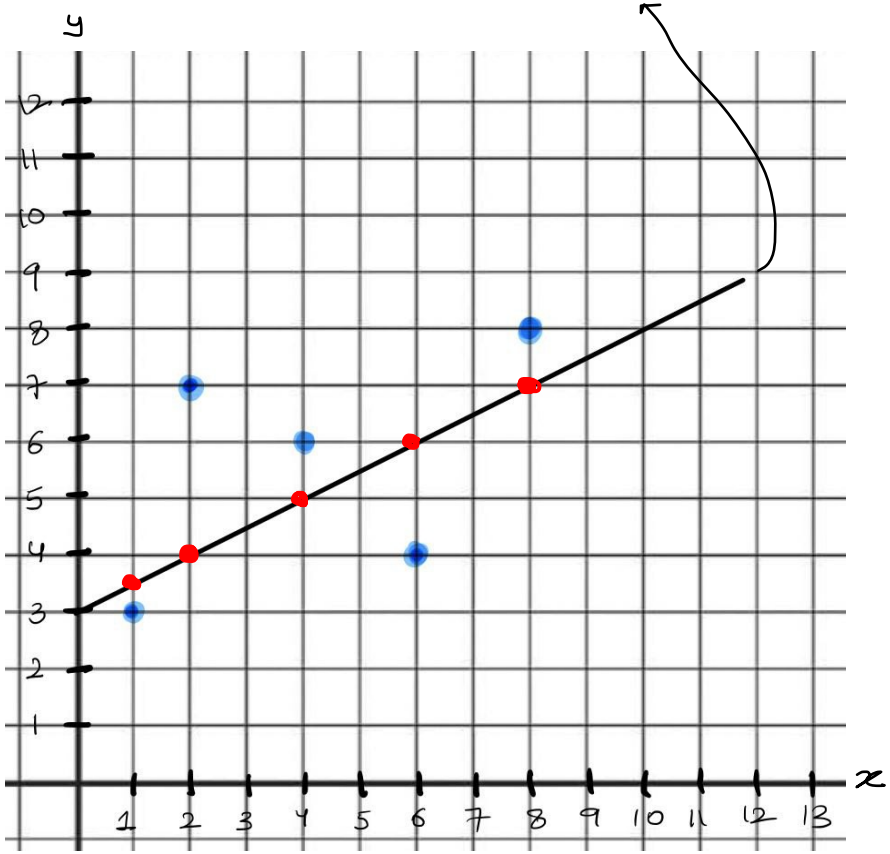
$y_p = y_{(predicted)} = \hat{y}$ (y estimate)

$y = w_0 + w_1(x)$

residual

x	y_a	y_p	$\varepsilon = (y_a - y_p)$	ε^2
1	3	3.5	-0.5	0.25
2	7	4	3	9
4	6	5	1	1
6	4	6	-2	4
8	8	7	1	1

Total 15.25 RSS (Residual Sum of Squares)



Simple Linear Regression

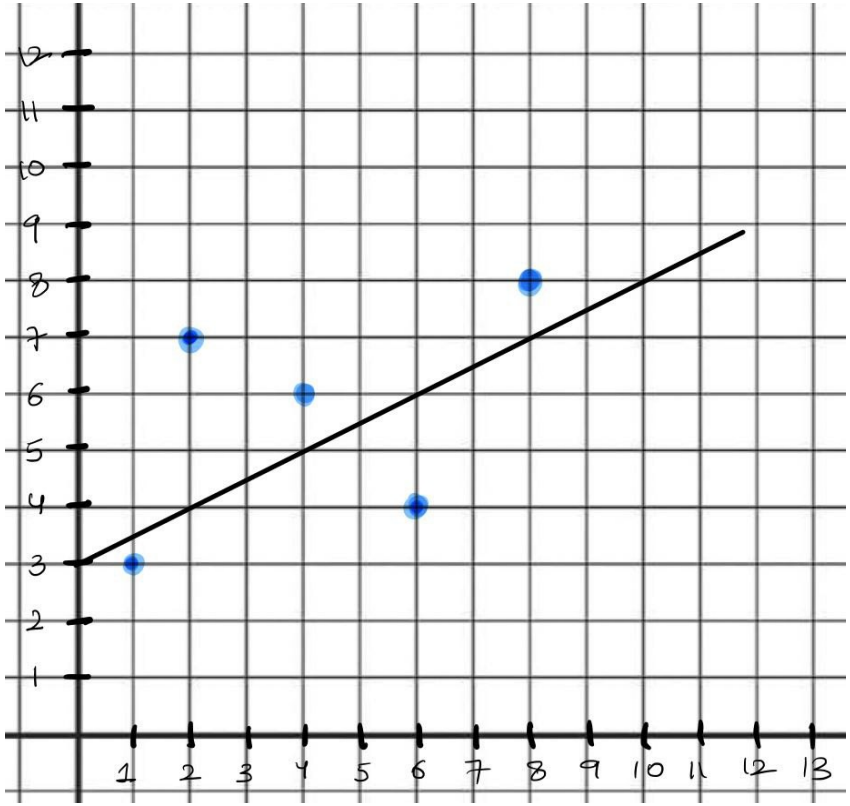
Initialisation → Random value of $w_0, \bar{w}$

① Random line

$w_0 + w_1(x)$ $[w_0 = 0, w_1 = 1]$

[RSS]

② Update $w_0, \bar{w}$ (gradient Descent)



Simple Linear Regression

objective $\longrightarrow$ minimize (RSS)

$$\hat{y} = w_0 + w_1 x$$

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

$$\mathcal{L}(w_0, w_1) = \underset{w_0, w_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x))^2 \Big\} \text{ Error minimisation}$$

objective : optimise the above function and find out that value of w_0, w_1 for which the function yields the minimum value $\Big\} \longrightarrow$ Gradient Descent

Slope of loss function w.r.t w_0

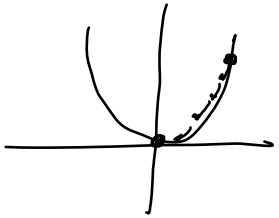
① initialise (w_0, w_1)

② update

$$w_0 = w_0 - \eta \overbrace{(\nabla_{w_0} \mathcal{L})}^{\text{Slope of loss function w.r.t } w_0}$$

$$w_1 = w_1 - \eta \underbrace{(\nabla_{w_1} \mathcal{L})}_{\text{Slope}}$$

η → Step size



Simple Linear Regression

x_1	x_2	x_3	y

$\longrightarrow$ $\boxed{\text{Linear Regression algo}}$ $\longrightarrow$ $\boxed{\text{Model}}$ $\longrightarrow$ $\boxed{\text{Predict}}$

Best w_0 w_1

Optimisation

$$\mathcal{L}(w_0, w_1) = \underset{w_0, w_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \overbrace{(w_0 + w_1(x))}^{\hat{y}})^2$$

Loss function

cost function

$\downarrow$

gradient Descent

objective: find that value of w_0 w_1 corrosponding to which the Loss function is minimum

$$\hat{y} = w_0 + w_1 x$$

$\hat{y}$

$x = 2$

Error Functions (Various Versions)

Problem with using Error function as Evaluation parameter

KRs

(y) K Income	$\hat{y}$	ε^2
10	5	25
20	15	25
25	25	0
15	15	0
20	20	0

50 = RSS

Rs Income	$\hat{y}$	ε^2
10,000	5000	
20,000	15000	
25,000	25000	
15,000	15000	
20,000	20,000	

RSS = 5 M

ERROR Is Dependent on the Scale of target variable

$$\text{MSE (mean Square error)} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$\text{MAE (mean-absolute error)} = \frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}|$$

$$\text{RMSE (Root mean square error)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y})^2}$$

Error

we can't use them for evaluating the goodness of the model

Motivation For R-2 Score

(Sharma ji ka ladka)
(sharma's son)

(Sumit)

(Sumit's father)

Compare your marks (good student)

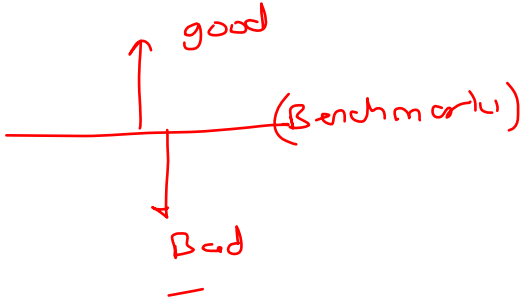
↓

Reference / Benchmark

Compare (performace)

↓

Benchmark

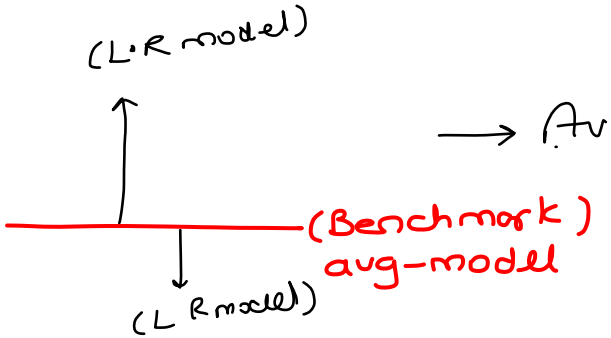


Simple Linear Regression

Assume that you work for a Marketing Firm, and your CEO requires you to provide her with a rough estimate of the sales that will occur next month. You already have the data (present on the right). Using the given data, try to determine the approximate sales for the upcoming month.

PS: Please remember that as of now, you do not have any knowledge of Linear Regression or complex Machine Learning algorithms. Try to solve this task using your logical understanding and basic mathematics.

avg (sales in last 3 months)



→ Average is the simplest possible prediction model.

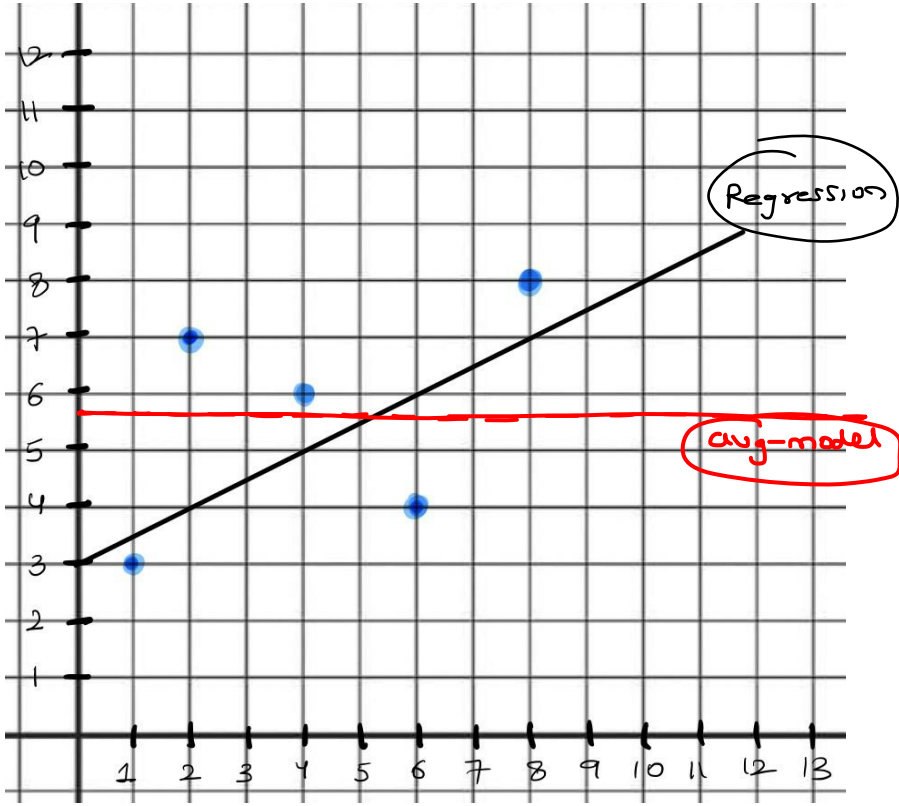
Metric → How much better your Regression model is from the avg-model

goodness

Simple Linear Regression

RSS = 15.25 (Squared Error by the regression model)
TSS = 17.20 (Squared Error by the avg-model)

x	y_a	$\hat{y}$	ε^2	$\bar{y}$	$\varepsilon^2 = (y_a - \bar{y})^2$
1	3	3.5	0.25	5.6	$(3-5.6)^2 = 6.76$
2	7	4	9	5.6	$(7-5.6)^2 = 1.96$
4	6	5	1	5.6	$(6-5.6)^2 = 0.16$
6	4	6	4	5.6	$(4-5.6)^2 = 2.56$
8	8	7	1	5.6	$(8-5.6)^2 = 5.76$
	5.6		15.25		17.2



$$R^2 = 1 - \left(\frac{RSS}{TSS}\right) = 1 - \left(\frac{15.25}{17.2}\right)$$

$$= 0.11$$

Industry ($R^2 = 85\%$)

Regression model is 11% better than the avg-model.