Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 02

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — Model
- Topic — optimisation
- Topic — Training
- Topic — data
- Topic

Basic

Topics to be Covered

{Gate exam}**

Topic

Linear regression

Topic

How We do optimisation

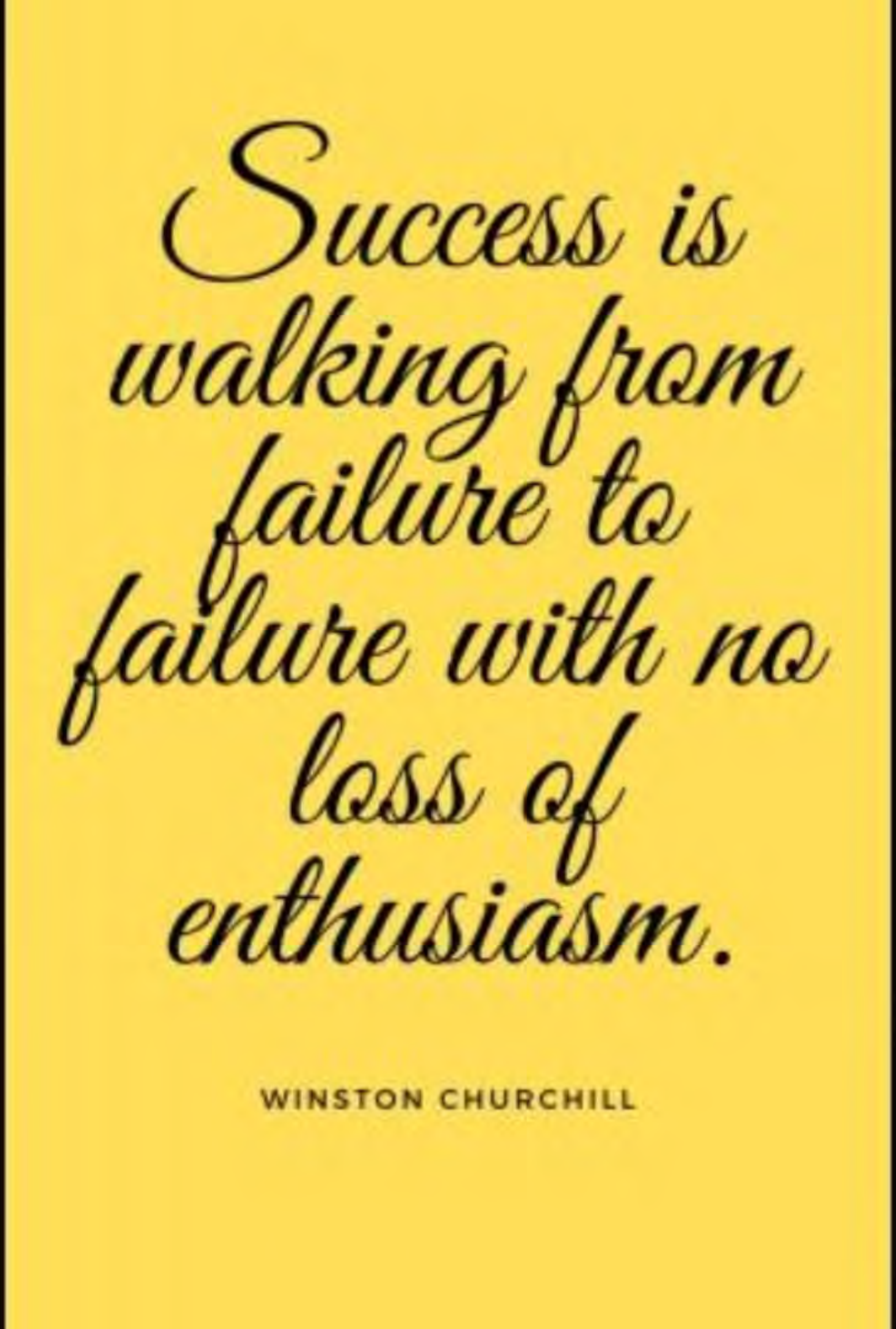
Topic

Loss function

Topic

Numerical Question.

Topic



Success is walking from failure to failure with no loss of enthusiasm.

WINSTON CHURCHILL

Patience

Fill in the blanks :

1. The target/Goal of the ML is "To predict y for new x" ____________

2. The best optimized model is that which minimize the error in "Training data" ____________

3. The problem with the simple model is and not learning Pattern of data ____________ ⇒ it has lot of error.

Basics of Machine Learning

REVISION

Fill in the blanks :

4. The problem with highly complicated model is
______________ noise in data is also included in Analysis / Rote learning

5. The data is used to ____Train____ the ML model

6. The data is collected from ____Survey / experiment____.

model ⇒ fxn
$y = f(x)$

Basics of Machine Learning

19. The output of training process in machine learning is

[✓ A. machine learning model]

B. machine learning algorithm

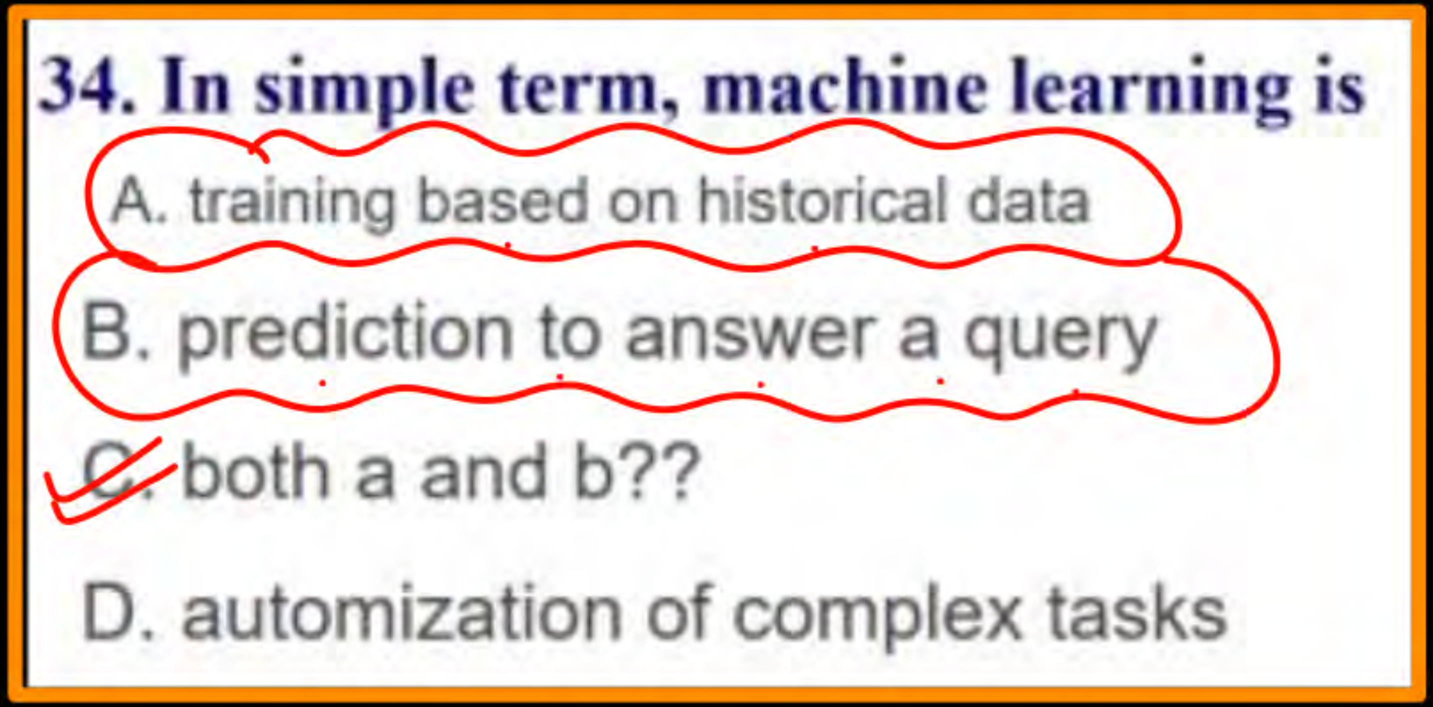
C. null

D. accuracy

- we get y and x Relation ⇒ we call this ML model.

Basics of Machine Learning

REVISION



34. In simple term, machine learning is

A. training based on historical data

B. prediction to answer a query

[✓ C. both a and b??]

D. automization of complex tasks

Basics of Machine Learning

Problem 2 – Predict Sale of I-phone based on Age of customer

We must create a model with following data

x

Age	Sale of I-Phone (in a month)
30	300
40	400

y

Predict $\underbrace{\text{Sale of iphone}}_{y}$ @ $\underbrace{\text{age}}_{x}$=20

Now predict the Sale of I-Phone at Age = 20

Problem 2 – Predict Sale of I-phone based on Age of customer

We don't have any expert now, and data has only two Points.
So ____________________

What is the best model now ?

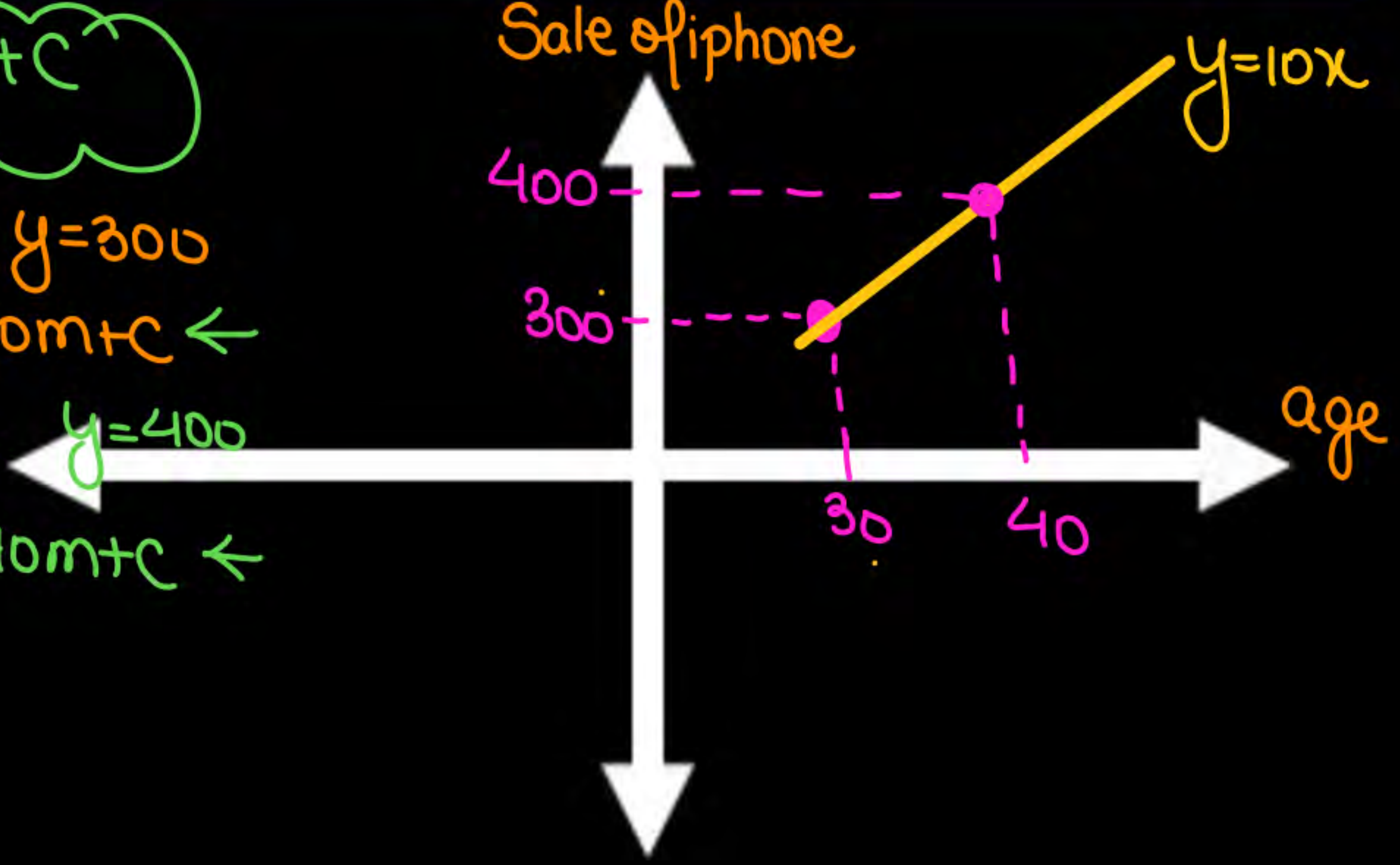
$y = mx + c$

$@x = 30 \quad y = 300$

$300 = 30m + c \leftarrow$

$@x = 40 \quad y = 400$

$400 = 40m + c \leftarrow$



$$30m + c = 300$$

$$40m + c = 400$$

$$10m = 100$$

$$m = 10 \checkmark$$

$$30 \times 10 + c = 300$$

$$c = 0 \checkmark$$

So we will try to fit a line on the data ⇒

So the Best line is that which has min gap b/w y_{actual} and $y_{predicted}$ values by model.

Since only 2 points are given hence ⇒ we can draw a line that can pass through both the points ⇒ $y_{actual} = y_{predicted}$

Basics of Machine Learning

Problem 2 – Predict Sale of I-phone based on Age of customer

Now we have to find the best parameters..

So Simple Case ⇒ 2 point in data easily we can draw a line passing through points

$y_{act} = y_{pred}$

Problem 2 – Predict Sale of I-phone based on Age of customer

Now we have to find the best parameters..

(done)

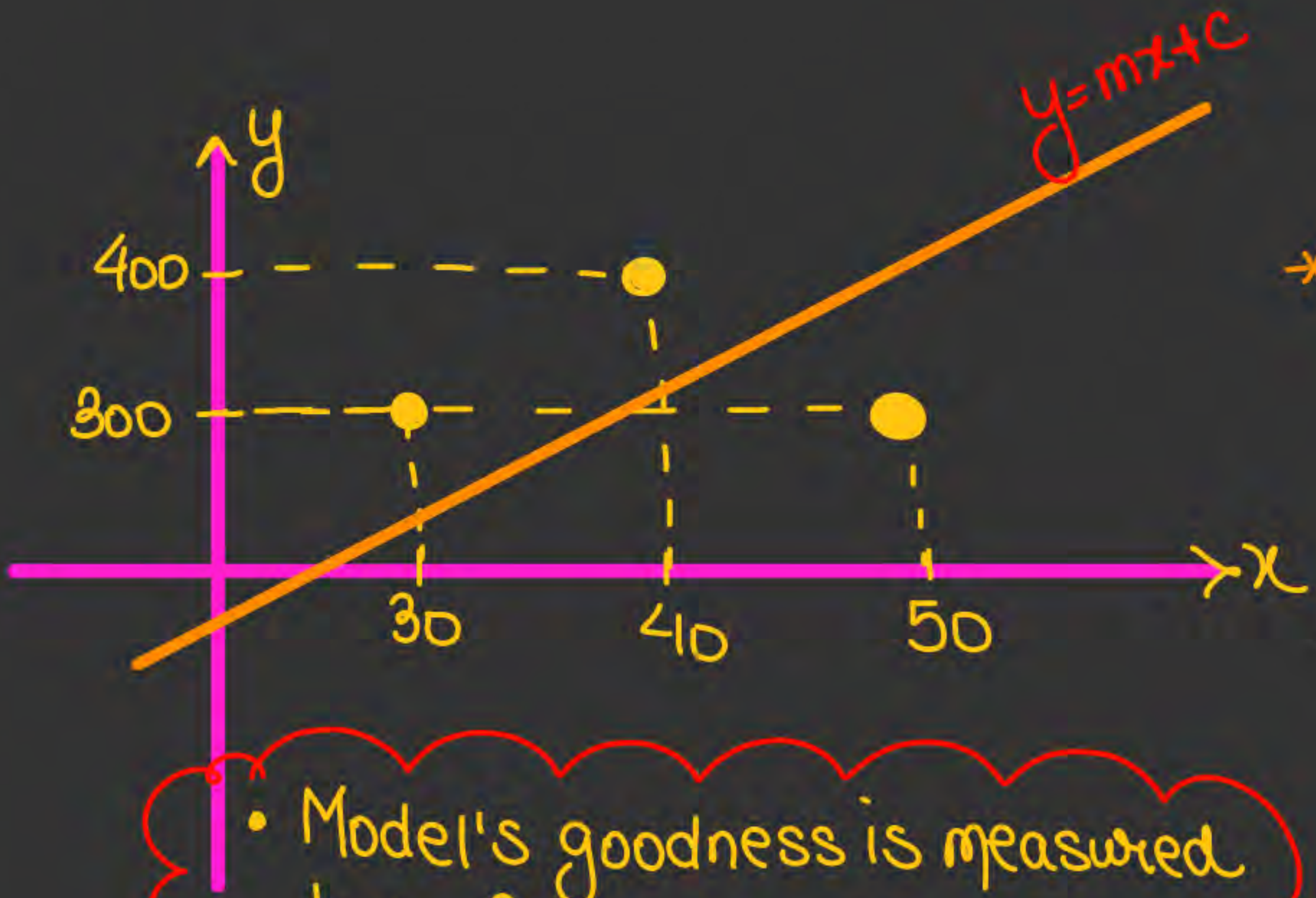
Problem 3 – Predict Sale of I-phone based on Age of customer

We must create a model with following data

Age	Sale of I-Phone (in a month)
30	300
40	400
50	300

- Predict Sale of iphone ⇒ y
  age ⇒ x

Now predict the Sale of I-Phone at Age = 20

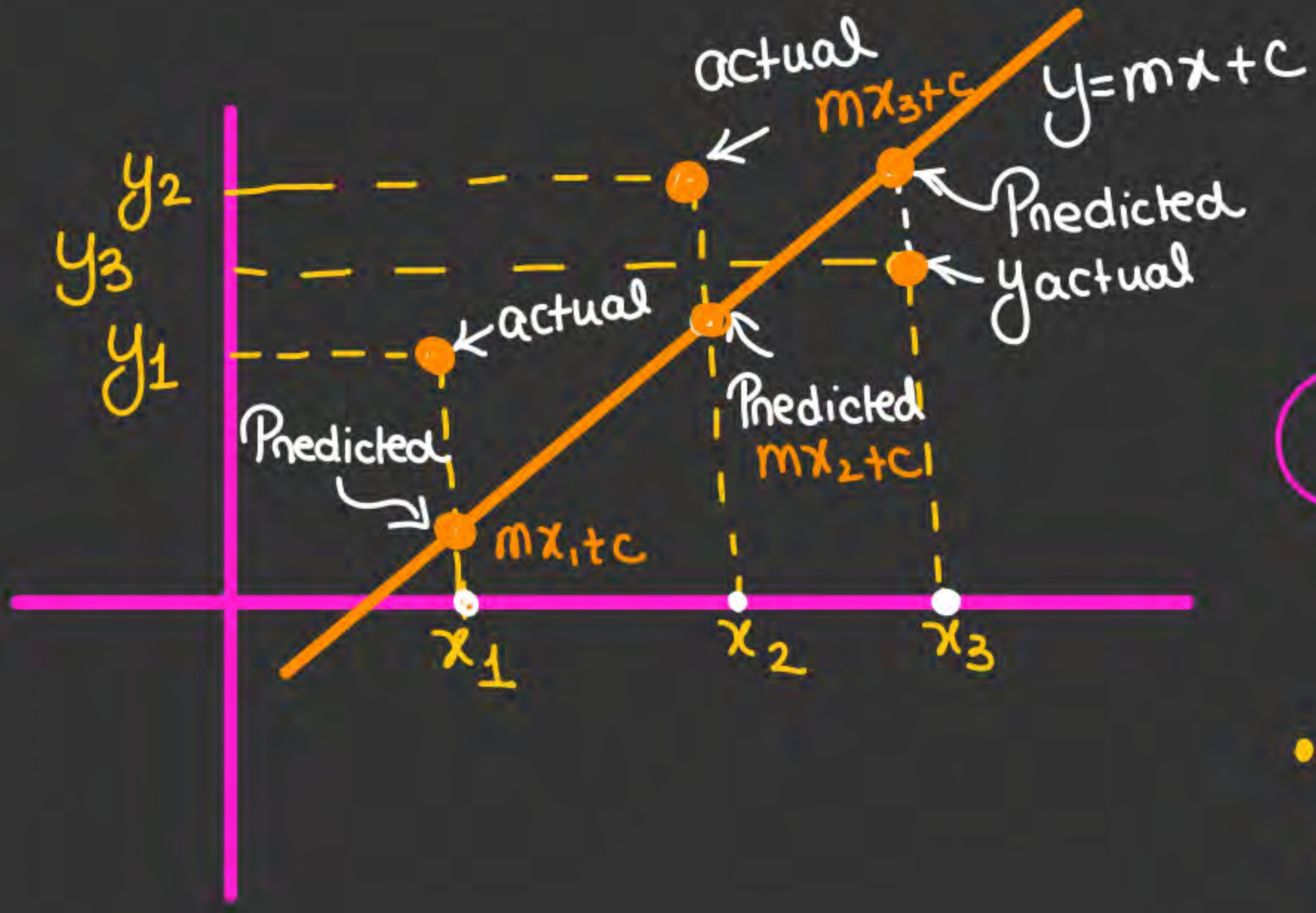


- Model's goodness is measured by ⇒ the gap b/w y_{actual} and $y_{predicted}$.

* So now data has 3 Points (x_1,y_1) (x_2,y_2) (x_3,y_3)

* So we can see that no St. line can pass through all 3 points.

* Now we have to find best Straight line.



So Shall we min $[y_{actual} - y_{predicted}]$

$\Downarrow$

- Here we have a problem that is this value can be +ve/ -ve

$\Rightarrow$ If we add all error then -ve error will Reduce the total error, may lead to confusion

So 2 options

$$\min \sum_{i=1}^{3} \left| y_i^{act} - y_i^{pred} \right|$$

⇓

absolute error

OR

$$\min \sum_{i=1}^{3} \left(y_i^{act} - y_i^{pred} \right)^2$$

- Residual Sum of Square.
- RSS

To find the min location we differentiate the fxn

↳

So Residual Sum of Square ⇒ The loss function

So data 3 point ⇒ $(x_1, y_1), (x_2, y_2), (x_3, y_3)$

The y_1, y_2, y_3 ⇒ These are actual 'y' Values of data

The st. line model ⇒ $y = mx + c$

$$\left\{\begin{array}{l} x_1 \xrightarrow{\text{Predicted}} \hat{y}_1 = mx_1 + c \\ x_2 \xrightarrow{\text{Predicted}} \hat{y}_2 = mx_2 + c \\ x_3 \xrightarrow{\text{Predicted}} \hat{y}_3 = mx_3 + c \end{array}\right\}$$

Now Best model ⇒ that minimize the loss function

⇓

By this maths we will get m, c

⇓

The Best model $\boxed{y=mx+c}$

The loss function ⇒

$$\mathcal{L} = \sum_{i=1}^{3} \left(y_i^{actual} - y_i^{pred.}\right)^2$$

RSS ⇐ $\mathcal{L} = \sum_{i=1}^{3} \left(y_i - (m x_i + c)\right)^2$

So x_i, y_i are given
But m, c are unknown

Now we have to minimize '$\mathcal{L}$' loss fxn ⇒

$\min \mathcal{L} \Rightarrow \min \sum_{i=1}^{3} \left(y_i - (m x_i + c)\right)^2$ ⇒ for minimizing ⇒ $\frac{\partial L}{\partial m} = 0$, $\frac{\partial L}{\partial c} = 0$

$$L = \sum_{i=1}^{3} \left(y_i - (m x_i + c)\right)^2$$

$$\frac{\partial L}{\partial m} = 2 \sum_{i=1}^{3} \left(y_i - (m x_i + c)\right)(-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{3} x_i y_i - m x_i^2 - c x_i = 0 \tag{1}$$

$$\frac{\partial L}{\partial c} = 2 \sum_{i=1}^{3} \left(y_i - (m x_i + c)\right)(-1) = 0$$

$$\sum_{i=1}^{3} \left(y_i - m x_i - c\right) = 0 \tag{2}$$

$$\frac{\partial}{\partial m} \left(f(m)\right)^2$$

$$\downarrow$$

$$2 f(m) \cdot f'(m)$$

$$2 f(m) = 0$$

$$\Rightarrow f(m) = 0$$

2 eq, 2 variable we can find m, c

Q $(30,300), (40,400), (50,300)$

(x_i, y_i)

of
find out best line for the data ⇒

$$\boxed{y = mx + c}$$

$$L = \sum_{i=1}^{3} \left(y_i - y_{pred_i}\right)^2$$

$$\min L = \min \sum_{i=1}^{3} \left(y_i - (mx_i + c)\right)^2$$

$$\frac{\partial L}{\partial m} = 0 \Rightarrow \sum_{i=1}^{3} \left(y_i - (mx_i + c)\right) x_i = 0$$

$$\left\{ \begin{array}{l} \Rightarrow \displaystyle\sum_{i=1}^{3} \left(y_i x_i - m x_i^2 - c x_i\right) = 0 \\ \dfrac{\partial L}{\partial c} = 0 \Rightarrow \displaystyle\sum_{i=1}^{3} \left(y_i - m x_i - c\right) = 0 \end{array} \right\}$$

$$\underbrace{\sum_{i=1}^{3}(y_i x_i)}_{30\times300+40\times400+50\times300} - m\underbrace{\sum_{i=1}^{3} x_i^2}_{30^2+40^2+50^2} - C\underbrace{\sum_{i=1}^{3} x_i}_{30+40+50} = 0$$

$$\boxed{40{,}000 - 5000m - 120C = 0}$$

$$\sum_{i=1}^{3} y_i - m\sum_{i=1}^{3} x_i - C\sum_{i=1}^{3} 1 = 0$$

$$300+400+300 \qquad 30+40+50$$

$$\boxed{1000 - 120m - 3C = 0}$$

$$5000m + 120C = 40{,}000$$

$$120m + 3C = 1000$$

$$\boxed{m = 0,\ C = 1000/3}$$

Q $\underline{\sum_{i=1}^{3}(3i+c)} \Rightarrow (3(1)+c)+(3(2)+c)+(3(3)+c)$

$$\Downarrow$$

$$\left(3\sum_{i=1}^{3} i + c\sum_{i=1}^{3} 1\right)$$

$$\underbrace{\Rightarrow 3\sum_{i=1}^{3} i + 3c}$$

If we had N points ⇒ $\mathcal{L} = \min\left(\sum_{i=1}^{N}\left(y_i - (mx_i + c)\right)^2\right)$

$$\frac{\partial L}{\partial m} = 0 \; \checkmark$$

$$\frac{\partial L}{\partial c} = 0$$

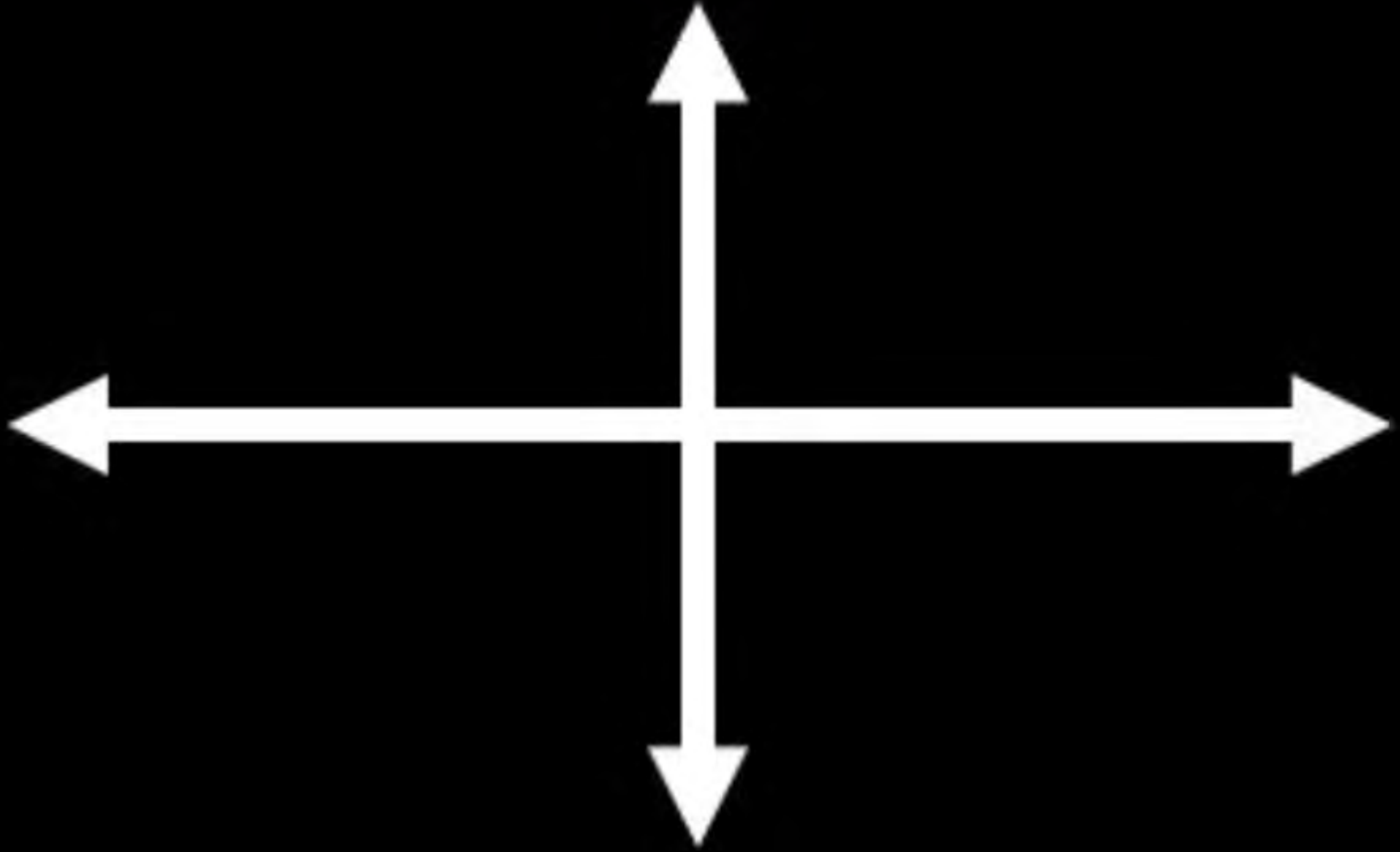
⇒ 2 equation and we get m, c.

Problem 3 – Predict Sale of I-phone based on Age of customer

We don't have any expert now, and data has only two Points. So

____

What is the best model now ?



Problem 3 – Predict Sale of I-phone based on Age of customer

We don’t have any expert now, and data has only two Points. So

____

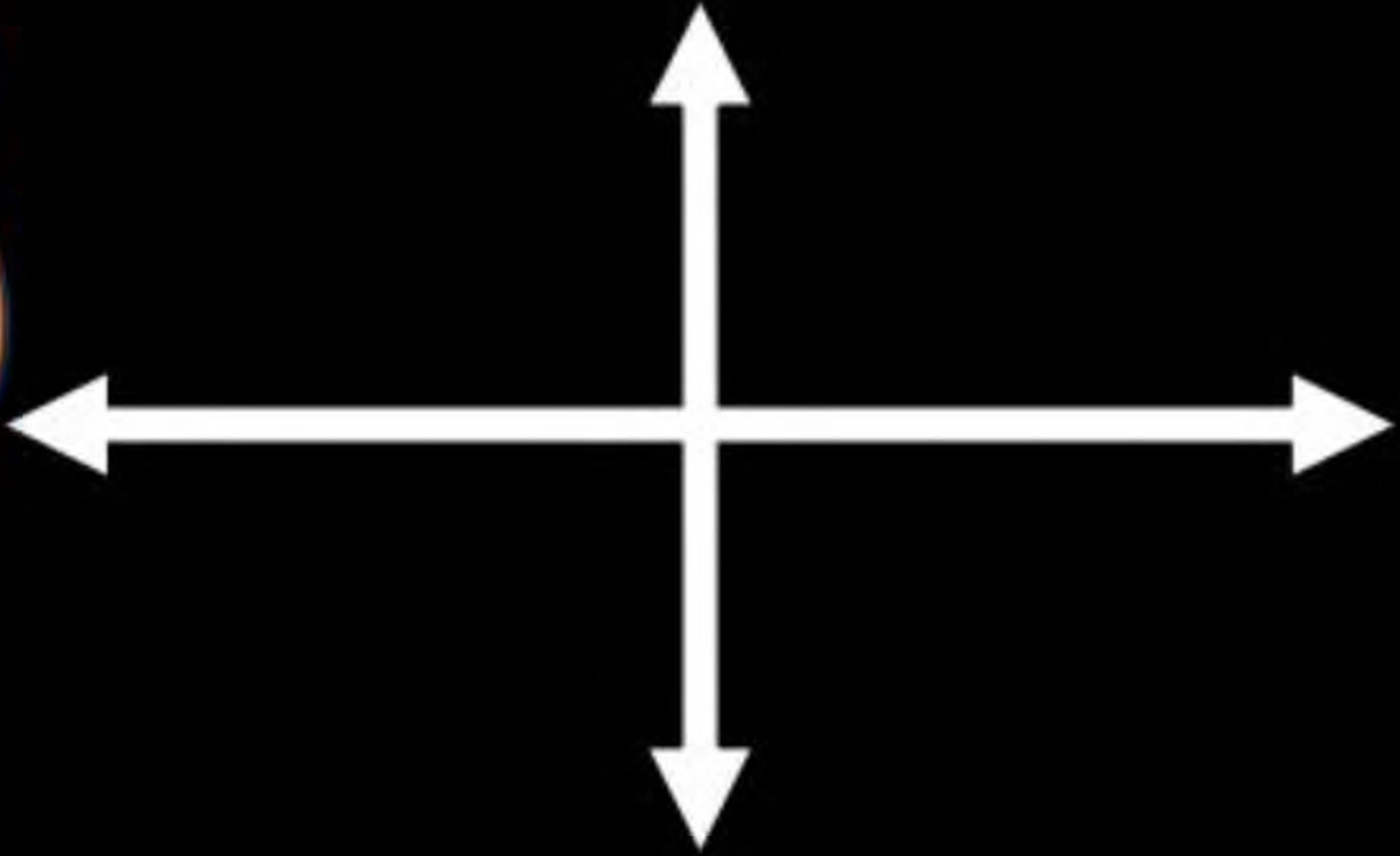
But we will try to find the linear model only.

So, we must find the model that try to

Problem 3 – Predict Sale of I-phone based on Age of customer

Creating the best model

Now we have that is called the predicted value of input.



Problem 3 – Predict Sale of I-phone based on Age of customer

Creating the best model

Loss Functions ?? (RSS- Residual Sum of Squares)

(done)

Reading

The residual sum of squares (RSS), also known as the sum of squared residuals (SSR) or the sum of squared estimate of errors (SSE), is the sum of the squares of residuals

Basics of Machine Learning

Now how to find the best parameters ??

Variance and mean...

mean of any variable

x_{i}^{o} ⇛ are some values

$$\overline{x_{i}^{o}} \Rrightarrow \frac{\sum_{i=1}^{N} x_i^{o}}{\text{Number of Values}}$$

$x^{o} = 3, 5, 7, 8$

$$\overline{x_{i}^{o}} = \frac{3+5+7+8}{4}$$

Now how to find the best parameters ??

Variance and mean...

Variance of a variable

$$(\sigma_x^2) = \frac{\sum_{i=1}^{N} (x_{i0} - \bar{x}_{i0})^2}{\text{Number of Values}}$$

$\sigma_{x_i} \Rightarrow$ Standard deviation.

$x_{i0} = 3, 8, 2, 7$

mean $\bar{x}_{i0} \Rightarrow \frac{3+8+2+7}{4} \Rightarrow 5$

$$\sigma_{x_{i0}}^2 \Rightarrow \frac{(3-5)^2+(8-5)^2+(2-5)^2+(7-5)^2}{4}$$

$\Rightarrow 6.5$

Now how to find the best parameters ??

Variance and mean...

If X and Y are two variables

$$\Rightarrow \mathrm{Cov}(XY) \Rightarrow \left\{ \frac{\sum_{i=1}^{N}(x_i-\overline{x}_i)(y_i-\overline{y}_i)}{N} \right\}$$

in Machine learning ⇛

mean of Variable X

↳ $$\frac{\sum_{i=1}^{N} x_i}{\text{No of Values}} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Variance of X ⇛

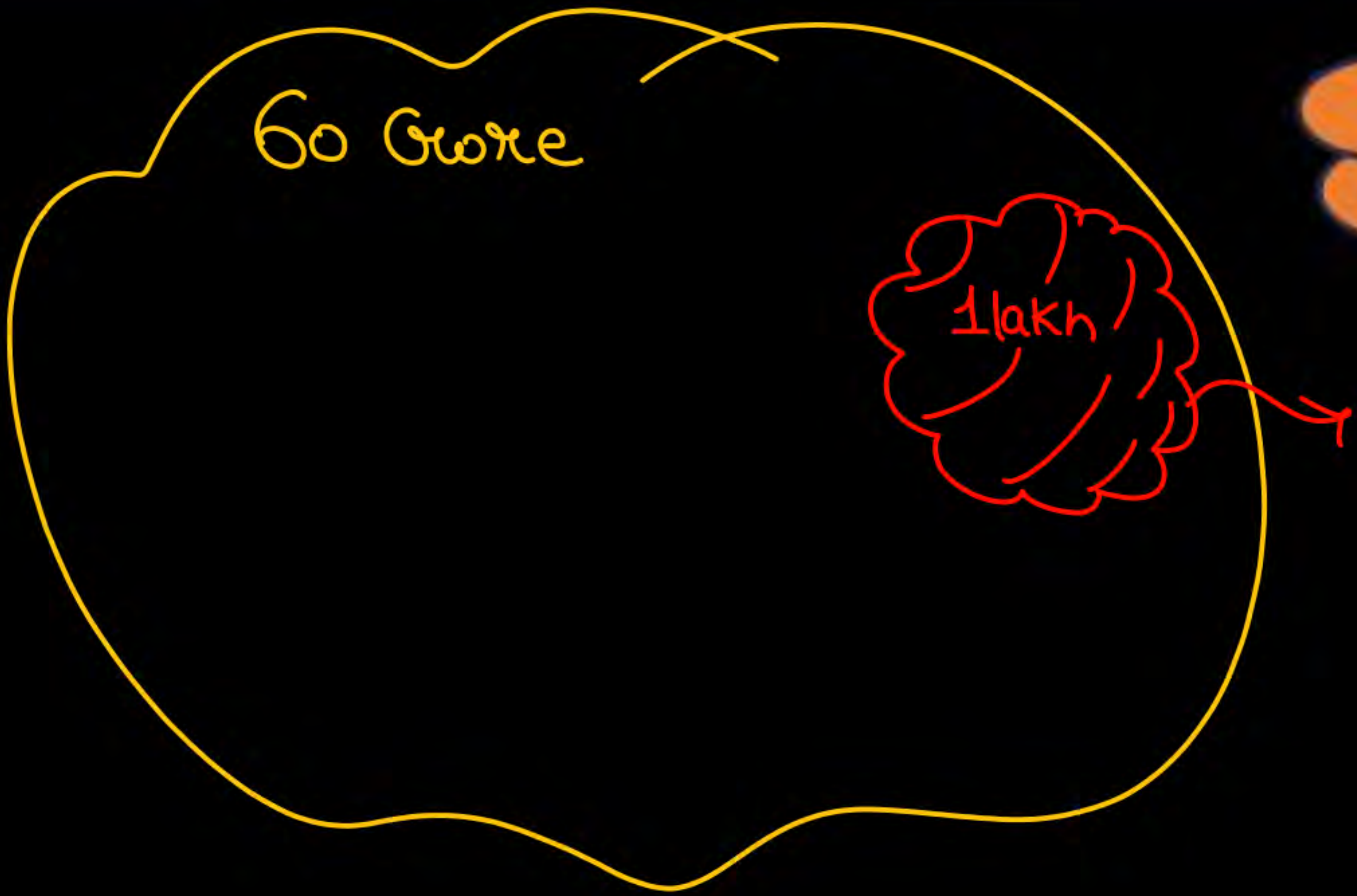
$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_i)^2}{(N-1)}$$

Covariance of X, Y

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{(N-1)}$$

Now how to find the best parameters ??

Formulae to find direct value of m and c



data is only a Sample.

Basics of Machine Learning

Now how to find the best parameters ??

Formulae to find direct value of m and c

So if we have any data

$(x_1, y_1)\ (x_2, y_2)\ (x_3, y_3)\ (x_4, y_4)\ ---$

So let $(y = mx + c)$ is our model

So $x = x_1, x_2, x_3, x_4 ---$

$y = y_1, y_2, y_3, y_4 ----$

$$\Rightarrow \left(m = \frac{Cov(x,y)}{Var\ x} \right)$$

after calculating m

$$\bar{y} = m\bar{x} + c$$

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

Now how to find the best parameters ??

Formulae to find direct value of m and c

Example

Obtain a linear regression for the data in below table assuming that y is the independent variable.

$x =$	1	2	3	4	5
$y =$	10	15	18	20	25

$$\Rightarrow y = mx + c$$

$$m = \frac{Cov(x,y)}{Var(x)}$$

$$\bar{x} \Rightarrow 3 \quad \bar{y} = 17.6$$

$$Var(x) = \frac{\sum (x_i - \bar{x})^2}{4} \Rightarrow 2.5$$

$$\mathrm{Cov}(x,y) = \frac{\sum_{i=1}^{5}(x-\bar{x})(y-\bar{y})}{4}$$

$$\Rightarrow \frac{(1-3)(10-17.6)+(2-3)(15-17.6)+(3-3)(18-17.6)+(4-3)(20-17.6)+(5-3)(25-17.6)}{4}$$

$$\Rightarrow \boxed{35/4}$$

$$m = \frac{35/4}{2.5} \Rightarrow 3.5 \checkmark$$

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

$$C = 17.6 - 3.5 \times 3$$

$$\boxed{C = 7.1}$$

Basics of Machine Learning

A set of observations of independent variable (x) and the corresponding dependent variable (y) is given below.

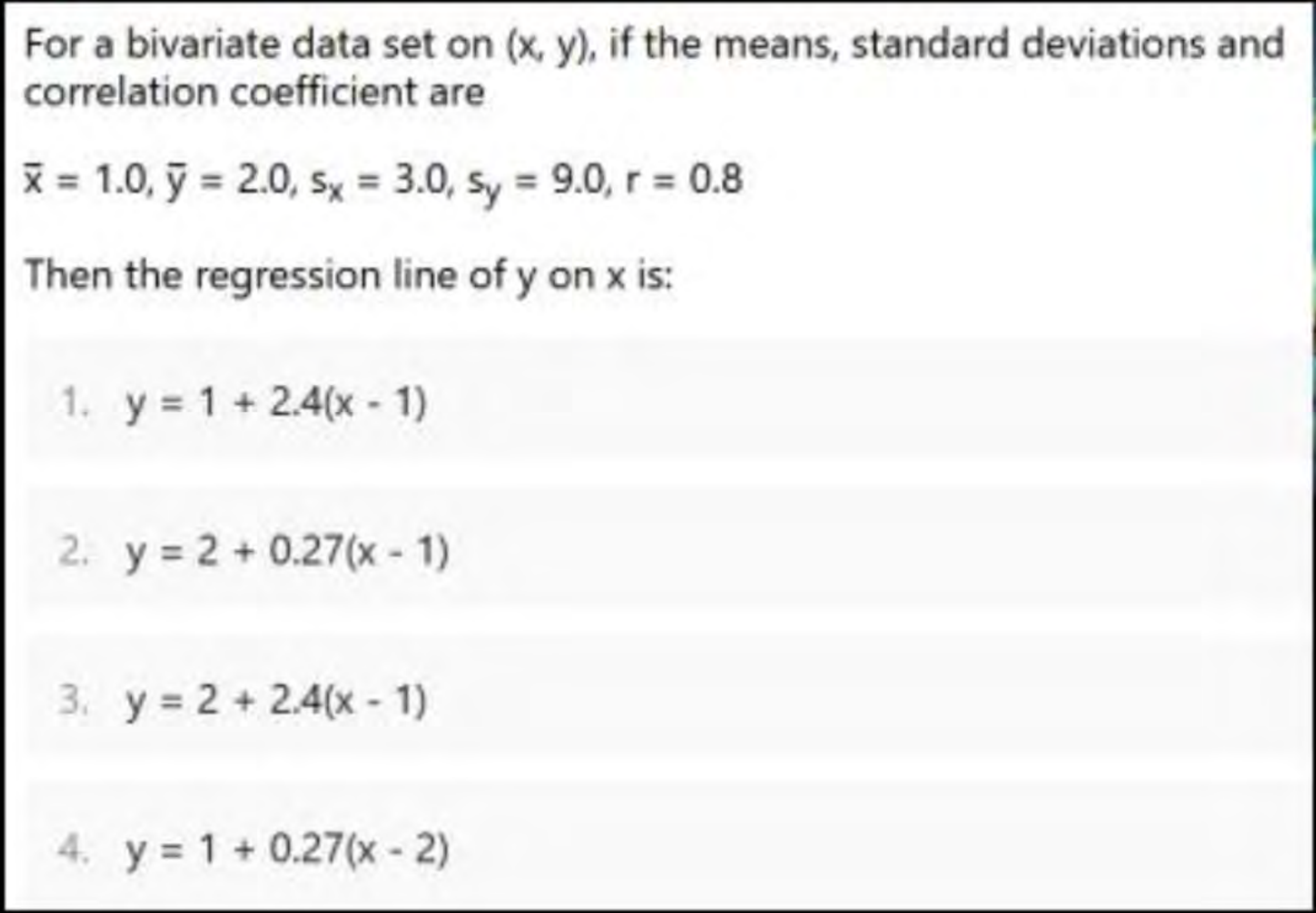
x	5	2	4	3
y	16	10	13	12

Based on the data, the coefficient a of the linear regression model

$y = a + bx$ is estimated as 6.1

The coefficient b is ____________ . (round off to one decimal place)

Basics of Machine Learning

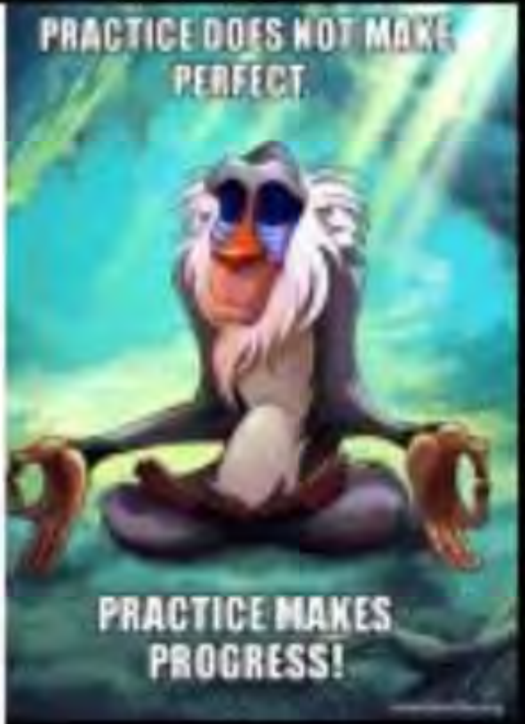


For a bivariate data set on (x, y), if the means, standard deviations and correlation coefficient are

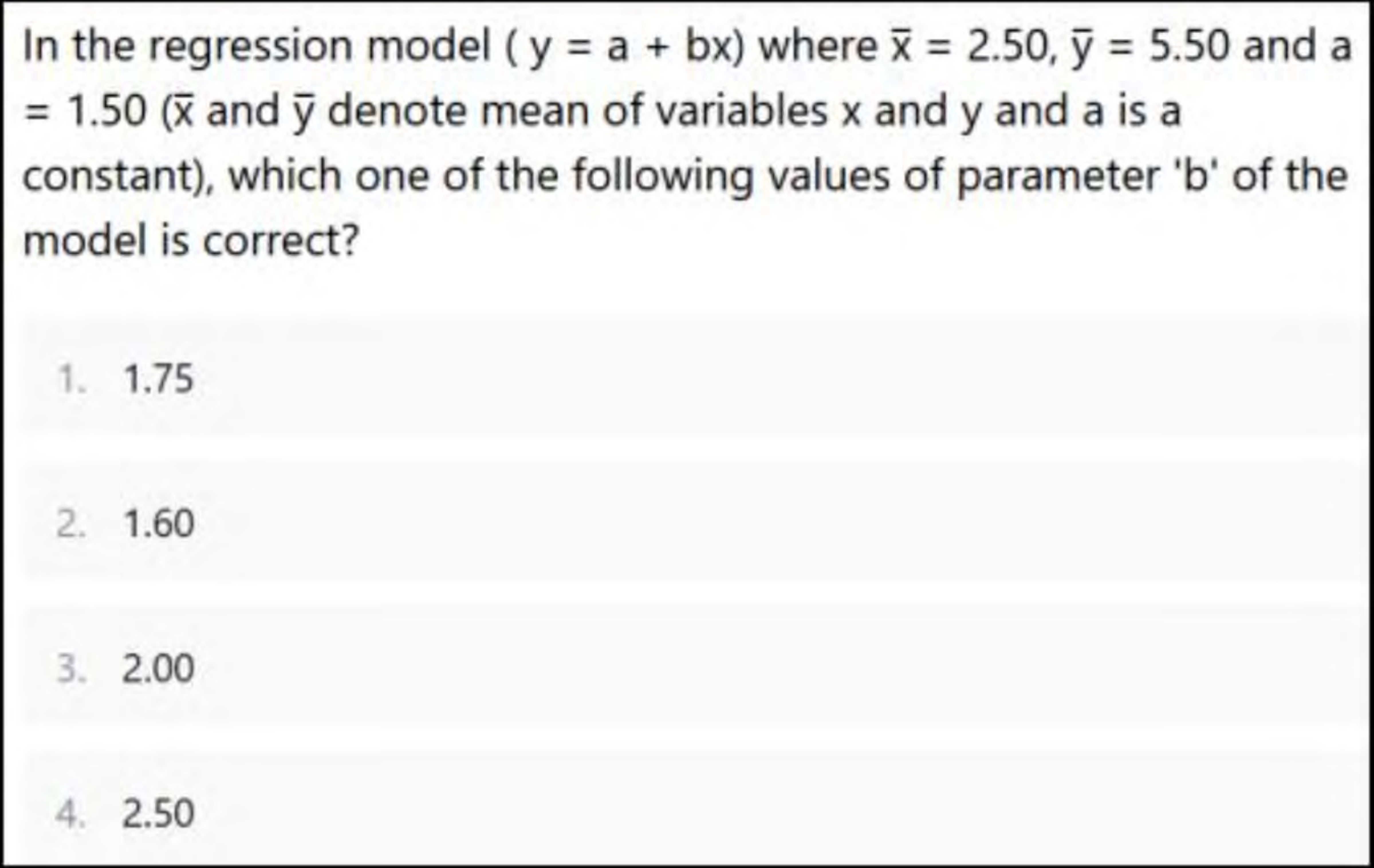
$\bar{x} = 1.0, \bar{y} = 2.0, s_x = 3.0, s_y = 9.0, r = 0.8$

Then the regression line of y on x is:

1. $y = 1 + 2.4(x - 1)$
2. $y = 2 + 0.27(x - 1)$
3. $y = 2 + 2.4(x - 1)$
4. $y = 1 + 0.27(x - 2)$



Basics of Machine Learning



In the regression model (y = a + bx) where $\bar{x}$ = 2.50, $\bar{y}$ = 5.50 and a = 1.50 ($\bar{x}$ and $\bar{y}$ denote mean of variables x and y and a is a constant), which one of the following values of parameter 'b' of the model is correct?

1. 1.75
2. 1.60
3. 2.00
4. 2.50

Basics of Machine Learning

There is no value of x that can simultaneously satisfy both the given equations. Therefore, find the ‘least squares error’ solution to the two equations, i.e., find the value of x that minimizes the sum of squares of the errors in the two equations. ________

$2x = 3$

$4x = 1$

Basics of Machine Learning

We can expect one Question from here in GATE exam

Considering data of 2 Dimensions

Attributes, Features, Dimensions...

Till now we have seen a simple case of 1 D data, now let's see 2 D Data

Income (LPA)	Age	Sale of I-Phone (in a month)
20	30	300
50	40	400
70	50	300
We have N Data points		

Now the input data is 2 D (age and income)

Linear Regression

The representation of D dimensional data

2 mins Summary

Topic

Topic

Topic

Topic

Topic

$y_{pred} = mx + c$

Loss function $\Rightarrow$ RSS $\Rightarrow \sum_{i=1}^{N}(y_i - y_{pred})^2 \Rightarrow \frac{\partial L}{\partial m}, \frac{\partial L}{\partial c}$

- $m = \frac{Cov(x, y)}{Var(x)}$
- $c = \bar{y} - m\bar{x}$

THANK - YOU