Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 03

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — Mean
- Topic — Variance
- Topic — Covariance
- Topic
- Topic

- linear regression
- 1D data ka LR

$$y = mx + c$$

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

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

Topics to be Covered

Topic Questions

Topic data representation

Topic LR for data with more than 1D.

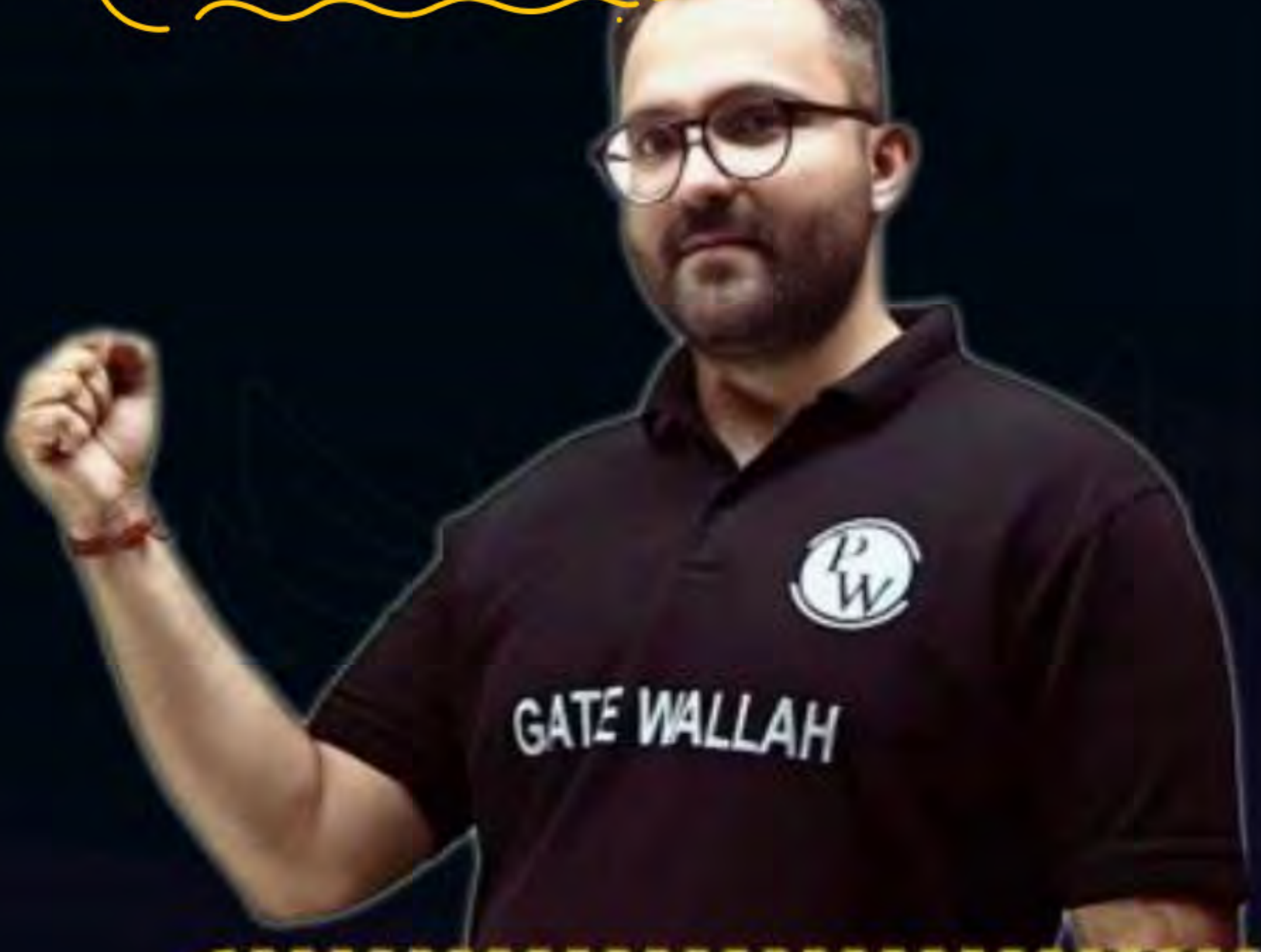
Topic

Topic

About the Faculty

- AIR 1 GATE 2021, 2023 (ECE).
- AIR 3 ESE 2015 ECE.
- M.Tech from IIT Delhi in VLSI.
- Published 2 papers in field of AI-ML.
- Paper 1 : Feature Selection through Minimization of the VC dimension.
- Paper 2 : Learning a hyperplane regressor through a tight bound on the VC dimension.

Siddharth Sir PW



By- SIDDHARTH SABHARWAL SIR

Inner Motivation

1. What is the Loss Function

Loss fxn ⇒ for linear Regression

$$\Rightarrow \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad \underline{RSS}$$

3. Direct formulae for M and C.

done

REVISION

4. Covariance :

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

5. Variance :

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

Linear Regression

REVISION

6. So the data that we were using has 1 ____ number of dimensions and the straight line obtained by liner regression has 2. ____ number of parameters.

line ka equation $y = mx + c$

$\downarrow \quad \downarrow$

2 parameters

(m, c)

What is Correlation Coefficient

$\rightarrow$ x and y are two variables.

x	y
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮

$Cov(x, y)$

$\sigma_x = \sqrt{Var(x)}$

$\sigma_y = \sqrt{Var(y)}$

ρ_{xy} : Correlation Coefficient of x, y

$$\rho_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

- $\rho_{xy} \Rightarrow$ Value always -1 to $+1$

Covariance $=0$

$\rho_{xy}=0 \Rightarrow$ x, y are uncorrelated

$\rho_{xy}=\pm 1 \Rightarrow$ x and y are highly correlated

$\rho_{xy}=+1 \Rightarrow$ x, y are likely related
x inc then y inc
x dec the y dec
Viceversa

$\rho_{xy}=-1 \Rightarrow$ then x, y have opposite relation
x inc then y dec
y inc then x dec

GAATE 2026 FODENGE

Representing the two equations in Matrix format

2 unknown 2equation

$$3x + 2y = 15$$
$$8x + 9y = 20$$

$$\begin{bmatrix} 3 & 2 \\ 8 & 9 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 15 \\ 20 \end{bmatrix}$$

$$\begin{bmatrix} 3x + 2y \\ 8x + 9y \end{bmatrix} = \begin{bmatrix} 15 \\ 20 \end{bmatrix}$$

$$4x + 9y = 20$$
$$9x + 3y = 50$$

$$\begin{bmatrix} 4 & 9 \\ 9 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 20 \\ 50 \end{bmatrix}$$

Representing the two equations in Matrix format

1D data

x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
⋮	⋮

Linear regression

model $y = \beta_1 x + \beta_0$

N data points

* β_1, β_0 are parameters.

$\hat{y}_i = \beta_1 x_i + \beta_0$

So loss function $\Rightarrow$ $L = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

$$= \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2$$

$$\mathcal{L} = \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2$$

Not Imp
Only result imp

To find β_1, β_0 (min L)

$$\frac{\partial L}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{N} \cancel{2}(y_i - \beta_1 x_i - \beta_0)\cancel{(-1)} = 0 \Rightarrow \sum_{i=1}^{N}(y_i - \beta_1 x_i - \beta_0) = 0$$

$$\frac{\partial L}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{N} 2(y_i - \beta_1 x_i - \beta_0)(-x_i) = 0 \Rightarrow \sum_{i=1}^{N}(y_i - \beta_1 x_i - \beta_0)x_i = 0$$

$$\boxed{\begin{array}{l|l} \sum_{i=1}^{N}\beta_1 x_i + \sum_{i=1}^{N}\beta_0 = \sum_{i=1}^{N} y_i & \left(\sum_{i=1}^{N} x_i\right)\beta_1 + \left(\sum_{i=1}^{N} 1\right)\beta_0 = \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N}\beta_1 x_i^2 + \sum_{i=1}^{N}\beta_0 x_i = \sum_{i=1}^{N} x_i y_i & \left(\sum_{i=1}^{N} x_i^2\right)\beta_1 + \left(\sum_{i=1}^{N} x_i\right)\beta_0 = \sum_{i=1}^{N} x_i y_i \end{array}}$$

- $\left(\sum_{i=1}^{N} x_i\right)\beta_1 + \left(\sum_{i=1}^{N} 1\right)\beta_0 = \sum_{i=1}^{N} y_i$

- $\left(\sum_{i=1}^{N} x_i^2\right)\beta_1 + \left(\sum_{i=1}^{N} x_i\right)\beta_0 = \sum_{i=1}^{N} x_i y_i$

$$\begin{bmatrix} \sum_{i=1}^{N} 1 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}$$

1D data Representation

x	y
x_1	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$

Created X, Y from available data

4 data Points

- $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$

- $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$

$$\begin{bmatrix} \sum_{i=1}^{N} 1 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix}$$

$$X^T X \Rightarrow \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} \sum_{i=1}^{4} 1 & \sum_{i=1}^{4} x_i \\ \sum_{i=1}^{4} x_i & \sum_{i=1}^{4} x_i^2 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} \sum_{i=1}^{4} y_i \\ \sum_{i=1}^{4} x_i y_i \end{bmatrix}$$

1D data Representation

x	y
x_1	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$

Created X, Y from available data

4 data Points

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

$$\begin{bmatrix} \sum_{i=1}^{N} 1 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}$$

model $y = \beta_1 x + \beta_0$

To find $\beta_1, \beta_0 \Rightarrow$

$$(X^T X)\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T Y)$$

$$\Rightarrow \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \left[(X^T X)^{-1}(X^T Y)\right]$$

Representing the two equations in Matrix format

Inverse of 2x2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \Rightarrow \boxed{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \frac{1}{ad-bc}}$$

- a and d interchange
- c and b sign change
- divide by determinant ad-bc

Linear Regression

GAATE 2026 FODENGE

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

$y = \beta_1 x + \beta_0$

$$X = \begin{bmatrix} 1 & 5 \\ 1 & 2 \\ 1 & 4 \\ 1 & 3 \end{bmatrix} \quad Y = \begin{bmatrix} 16 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 2 & 4 & 3 \end{bmatrix}$$

$$X^T X \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = X^T Y$$

$$\begin{bmatrix} 4 & 14 \\ 14 & 54 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 51 \\ 188 \end{bmatrix}$$

PRACTICE DOES NOT MAKE PERFECT

PRACTICE MAKES PROGRESS!

$$X = \begin{bmatrix} 1 & 5 \\ 1 & 2 \\ 1 & 4 \\ 1 & 3 \end{bmatrix} \quad Y = \begin{bmatrix} 16 \\ 10 \\ 13 \\ 12 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 2 & 4 & 3 \end{bmatrix}$$

$$X^T X \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = X^T Y$$

$$\begin{bmatrix} 4 & 14 \\ 14 & 54 \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{bmatrix} 51 \\ 188 \end{bmatrix}$$

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \frac{1}{20} \begin{bmatrix} 54 & -14 \\ -14 & 4 \end{bmatrix} \begin{bmatrix} 51 \\ 188 \end{bmatrix}$$

$$= \begin{bmatrix} 122 \\ 38 \end{bmatrix} \frac{1}{20}$$

$$\boxed{\begin{aligned} \beta_0 &= 122/20 \\ \beta_1 &= 38/20. \end{aligned}}$$

- $y = \beta_1 x + \beta_0$

$$y = \frac{38}{20} x + \frac{122}{20}$$

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

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

Line eq
$y = 2.4x - 0.4$

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. ~~$y = 1 + 0.27(x - 2)$~~

$\bar{x} = 1$
$\bar{y} = 2.0$
$\sigma_x = 3$
$\sigma_y = 9$
$\boxed{\rho_{xy} = .8} = \dfrac{Cov(x, y)}{\sigma_x \sigma_y}$

Find Linear Reg line

$\Rightarrow y = mx + c$

$m = \dfrac{Cov(x, y)}{Var(x)} = \dfrac{0.8 \times 3 \times \cancel{9}}{\cancel{(3)^2}} = 2.4$

$\rightarrow Cov(x, y) = .8 \times 3 \times 9$

$c = \bar{y} - m\bar{x} = 2 - 2.4 \times 1 = -0.4$

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?

done ✓✓

1. 1.75
2. $\boxed{1.60}$
3. 2.00
4. 2.50

$y = a + bx$

$a = 1.5 \quad \bar{x} = 2.5$

$\bar{y} = 5.5$

$a = \bar{y} - b\bar{x}$

$1.5 = 5.5 - b\,2.5$

$b \times 2.5 = 4$

$b = 4/2.5 = 1.60$

Linear Regression

There is no value of x that can simultaneously satisfy both the given equations. Therefore, find the $\boxed{\text{'least squares error' solution}}$ to the two equations, i.e., find the value of x that minimizes the sum of squares of the errors in the two equations. ________

$2x = 3$

$4x = 1$

$$\boxed{\begin{matrix} 2x = 3 \\ 4x = 1 \end{matrix}}$$

If α is value of x Chahiye

Error in Term1 ⇒ actual − predicted ⇒ $3 - 2\alpha$

Term2 ⇒ actual − predicted ⇒ $1 - 4\alpha$

Square error ⇒ $(3-2\alpha)^2 + (1-4\alpha)^2$

To min error ⇒ $\frac{\partial L}{\partial \alpha} = 0$

$\cancel{2}(3-2\alpha)(+2) + \cancel{2}(1-4\alpha)(+4) = 0$

$6 - 4\alpha + 4 - 16\alpha = 0$

$\alpha = 1/2$

$y = 2/3$

$y = 9/5$

$(y = \alpha)$

$\left[\left(\frac{2}{3} - \alpha\right)^2 + \left(\frac{9}{5} - \alpha\right)^2\right]$

$3y = 2$

$5y = 9$

$y = 1.2$

Both Cannot Satisfy together

y aisa chahiye

3y close to 2

5y close to 9

5y chahiye tha → 9

5y mila kitna $5 \times 1.2 = 6$

3

H.W

$5z = 1$

$2z = 5$

$10z = 3$

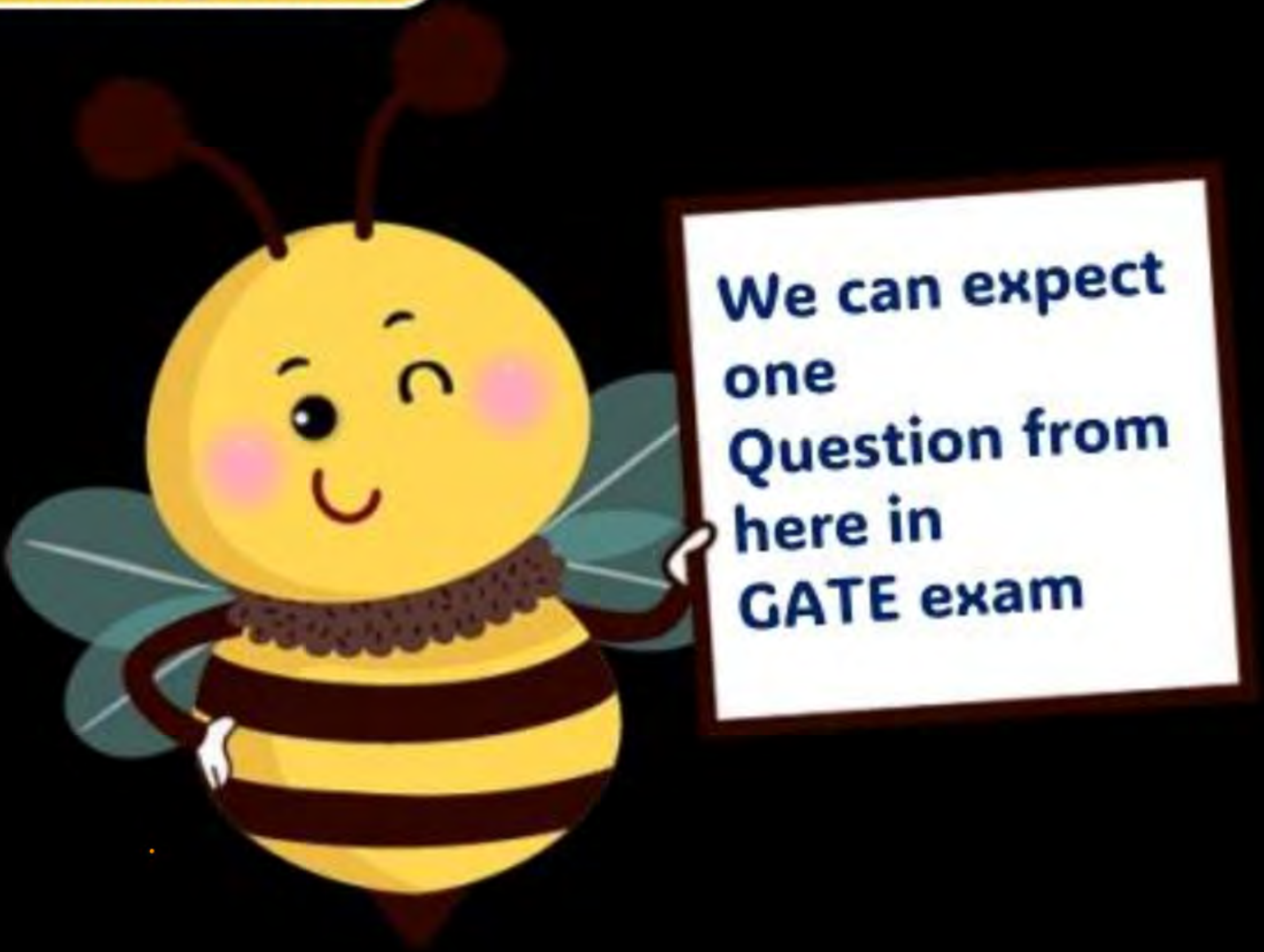
H.W

VV.

Find z that minimize the RSS or SSE

SSE $(5\alpha-1)^2 + (2\alpha-5)^2 + (10\alpha-3)^2$

Ans $\Rightarrow \frac{15}{43}$


We can expect
one
Question from
here in
GATE exam

GAATE 2026 FODENGE

Hamesha ek hi label

Considering data of 2 Dimensions

So now data has Two dimension attribute features.

Y label

Attributes, Features, Dimensions...

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)

Two dimension
Model ⇒ $\left[y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 \right]$
3 Parameters.

How to write the 2 D inputs ??

	x^1	x^2	y.
	1st dimension	2nd dimension	label
1	x_1^1	x_1^2	y_1
2	x_2^1	x_2^2	y_2
3	x_3^1	x_3^2	y_3
4	x_4^1	x_4^2	y_4
	⋮	⋮	⋮

$$X = \begin{array}{c} \begin{array}{ccc} & 1 & 2 \end{array} \\ \begin{bmatrix} 1 & x_1^1 & x_1^2 \\ 1 & x_2^1 & x_2^2 \\ 1 & x_3^1 & x_3^2 \\ 1 & x_4^1 & x_4^2 \end{bmatrix} \end{array}$$

$x_{\text{Subscript}}^{\text{Superscript}}$

Superscript ⇒ Show dimension

Subscript ⇒ datapoint Number

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Linear Regression

Linear model will have ________ number of parameters

Solution

$$\boxed{(X^TX)\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix}=(X^TY)}$$

Linear Regression

Considering data of P Dimensions

The loss function for P dimensions case

Loss function in Matrix Form

We do partial differentiation in terms of all variables to get the optimized variable values

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

RSS = residual sum of squares

y_i = i^th value of the variable to be predicted

$f(x_i)$ = predicted value of y_i

n = upper limit of summation

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

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

TSS = total sum of squares

n = number of observations

y_i = value in a sample

$\bar{y}$ = mean value of a sample

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- The most important thing we do after making any model is evaluating the model.
- R-squared is a statistical measure that represents the goodness of fit of a regression model.
- The value of R-square lies between 0 to 1.
- Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value.
- However, we get R-square equals 0 when the model does not predict any variability in the model.

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.
- The most common interpretation of r-squared is how well the regression model explains observed data. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model.

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

- The goodness of fit of regression models can be analyzed on the basis of the R-square method. The more the value of the r-square near 1, the better the model is.

- Note: The value of R-square can also be negative when the model fitted is worse than the average fitted model. .

Considering data of P Dimensions

Adjusted R - Squares

- Adjusted R-Squared is an updated version of R-squared which takes account of the number of independent variables while calculating R-squared.
- n is the total number of observations in the data
- k is the number of independent variables (predictors) in the regression model

$$Adjusted R^2 = 1 - \frac{(1-R^2)\cdot(n-1)}{n-k-1}$$

THANK - YOU