Data Science and Artificial Intelligence

Machine Learning

Linear Regression

Lecture No. 05

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — loss fxn → matrix Rep
- Topic — loss fxn $\frac{\partial L}{\partial B}$
- Topic — Gradient descent
- Topic — H.W
- Topic

Topics to be Covered

- Topic — Homework
- Topic — R^2 : Coeff of det.
- Topic — MSE
- Topic — RMSE
- Topic — Assumptions in L.R


Mental Block
DON'T LIMIT YOUR
CHALLENGES
CHALLENGE YOUR LIMIT
LIONHEARTED

REVISION

Loss function in matrix format

$$m = \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow a^2 + b^2$$

$$m^T m \Rightarrow (a\ b)\begin{pmatrix} a \\ b \end{pmatrix} = a^2 + b^2$$

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

$$(Y - \hat{Y}) = \begin{bmatrix} y_1 - \hat{y}_1 \\ \vdots \\ \end{bmatrix}$$

$$L = (Y - \hat{Y})^T (Y - \hat{Y})$$

REVISION

Derivative of L by Beta

$$\frac{\partial L}{\partial \beta} = \begin{bmatrix} \partial y/\partial \beta_0 \\ \partial y/\partial \beta_1 \\ \partial y/\partial \beta_2 \\ \vdots \\ \partial y/\partial \beta_D \end{bmatrix} \Rightarrow -2 \begin{bmatrix} X^T Y - (X^T X)\beta \\ \\ \\ \end{bmatrix}$$

Gradient Descent Method...

we start with any Random β

$$\beta^{new} = \beta^{old} - \eta \left.\frac{\partial L}{\partial \beta}\right|_{\beta old}$$

η → learning

Basics of Machine Learning

Learning Rate →

- If it is large then the Solution is never reached → (.01 or .001)
- We keep it V. Small so that movement is precise and we reach the solution → minima location easily.

Linear Regression

What is a Hat Matrix

data

$$X = \begin{bmatrix} & \\ & \end{bmatrix}_{N \times DH}$$

Actual Values

$$Y = \begin{bmatrix} \\ \\ \end{bmatrix}_{N \times 1}$$

$$\hat{Y} = \begin{bmatrix} \\ \\ \end{bmatrix} \Rightarrow (X\beta) \Rightarrow \left(X(X^TX)^{-1}X^TY\right)$$

Matrix of Predicted Values.

So $\hat{Y} = \underbrace{X(X^TX)^{-1}X^T}_{\text{Hat matrix}} Y$

So this matrix $X(X^TX)^{-1}X^T$ is called hat matrix Bcoz $\hat{Y} \rightarrow Y$.

#Q. Let's consider regression in one dimension, so our inputs $x^{(i)}$ and outputs $y^{(i)}$ are in $\mathbb{R}$.

(a) (4 points) Linny uses regular linear regression. Given the following dataset, (x,y)

$D = \{((1), 1), ((2), 2), ((3), 4), ((3), 2)\}$

What value of θ and θ_0 optimize the mean squared error of hypotheses of the form $h(x; \theta, \theta_0) = \theta x + \theta_0$? $\theta_1 x + \theta_0 \rightarrow mx + c$

$\beta_1 x + \beta_0 \rightarrow$

1) 2 eq, Solve
2) $m = \frac{Cov(x,y)}{Var\,x}$, $c = \bar{y} - m\bar{x}$
3) Matrix

$$X=\begin{bmatrix}1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 3\end{bmatrix}\begin{matrix}\leftarrow data_1\\ \leftarrow data_2\\ \leftarrow data_3\\ \leftarrow data_4\end{matrix}\quad 4\times(1+1)$$

$$Y=\begin{bmatrix}1\\ 2\\ 4\\ 2\end{bmatrix}_{4\times1}$$

$$\beta=\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$$

$$\beta=(X^TX)^{-1}X^TY$$

Inv of matrix

$$\begin{bmatrix}a & b\\ c & d\end{bmatrix}\xrightarrow{inv}\frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$$

- $X^TX=\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 2 & 3 & 3\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 3\end{bmatrix}=\begin{bmatrix}4 & 9\\ 9 & 23\end{bmatrix}$

$$(X^TX)^{-1}\Rightarrow\frac{1}{23\times4-9\times9}\begin{bmatrix}23 & -9\\ -9 & 4\end{bmatrix}=\frac{1}{11}\begin{bmatrix}23 & -9\\ -9 & 4\end{bmatrix}$$

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

$$= \begin{bmatrix} 9 \\ 23 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} (X^T Y)$$

$$= \frac{1}{11} \begin{bmatrix} 23 & -9 \\ -9 & 4 \end{bmatrix} \begin{bmatrix} 9 \\ 23 \end{bmatrix}$$

$$= \frac{1}{11} \begin{bmatrix} 0 \\ 11 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{So} \begin{cases} \beta_0 = 0 \\ \beta_1 = 1 \end{cases}$$

The term β_0 in Linear Regression

$$Y = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 - - - -$$

β_0 is called Bias term or intercept

$y = mx + c$

this is called bias term/intercept

#Q. Consider a one-dimensional regression problem with training data $\{x_i, y_i\}$. We seek to fit a linear model with no bias term:

(a) Assume a squared loss $\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ and solve for the optimal value of ω^*.

→ data has single dimension

So L.R eq, ⇒ $y = \omega_1 x + \omega_0$ → no bias term

$y = \omega_1 x$ ⇐ Predicted Value

$$L = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$= \frac{1}{2}\sum_{i=1}^{N}(y_i - \omega_1 x_i)^2$$

So $\frac{\partial L}{\partial w_1} = -\frac{2}{2} \sum_{i=1}^{0} x_i^0 (y_i - w_1 x_i) = 0$

$$\sum_{i=1}^{N} x_i^0 y_i^0 - w_1 \sum_{i=1}^{N} (x_i)^2 = 0$$

$$w_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

#Q. Consider a one-dimensional regression problem with training data $\{x_i, y_i\}$. We seek to fit a linear model with ~~X~~ only bias term:

(a) Assume a squared loss $\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ and solve for the optimal value of ω^*.

$\omega^* =$ a) $\sum_{i=1}^{N} y_i x_i / N$

b) $\sum_{i=1}^{N} x_i / N$

c) $\sum_{i=1}^{N} y_i / N$

d) none

Sol $\hat{y} = \omega_0$

$L = \frac{1}{2}\sum_{i=1}^{N}(y_i - \omega_0)^2$

To minimize

$\frac{\partial L}{\partial \omega_0} = -\sum_{i=1}^{N}(y_i - \omega_0) = 0$

$\sum_{i=1}^{N} y_i - N\omega_0 = 0$

$\Rightarrow \omega_0 = \sum_{i=1}^{N} y_i / N$

#Q. Consider the following 4 training examples:

X	Y
–1	0.0319
0	0.8692
1	1.9566
2	3.0343

We want to learn a function $f(x) = ax + b$ which is parametrized by (a, b). Using squared error as the loss function, which of the following parameters would you use to model this function.

(a) (1, 1)

(b) (1, 2)

(c) (2, 1)

(d) (2, 2)

Linear Regression

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

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

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

So TSS → Total Sum of Squares

→ we take Simplest Solution
ie $\hat{y} = y_{avg}$

$$\Rightarrow TSS = \sum_{i=1}^{N} (y_i - y_{avg})^2$$

RSS ⇒

Residual Sum of Square of my model

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

Simplest model $\hat{y} = y_{avg}$.

$$TSS \Rightarrow \sum_{i=1}^{N} (y_i - y_{avg})^2$$

$$R^2 \Rightarrow \text{Coeff of det.} = \left(1 - \frac{RSS}{TSS}\right)$$

⇒ ideally for a very good model RSS = 0

↳ $R^2 = 1$

If RSS = TSS, then our model is "bekar", bcoz here simplest model perform similar to our model.

$R^2 = 0$ ⇒ "bekar" model.

- $R^2 = 1 - \frac{RSS}{TSS}$, If RSS > TSS, this means model performs much Poor wrt the avg model
  - → This condition never occur.

So Range of R^2 → 0 to 1.

Linear Regression

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

Linear Regression

Considering data of P Dimensions

R-squared in Regression Analysis in Machine Learning

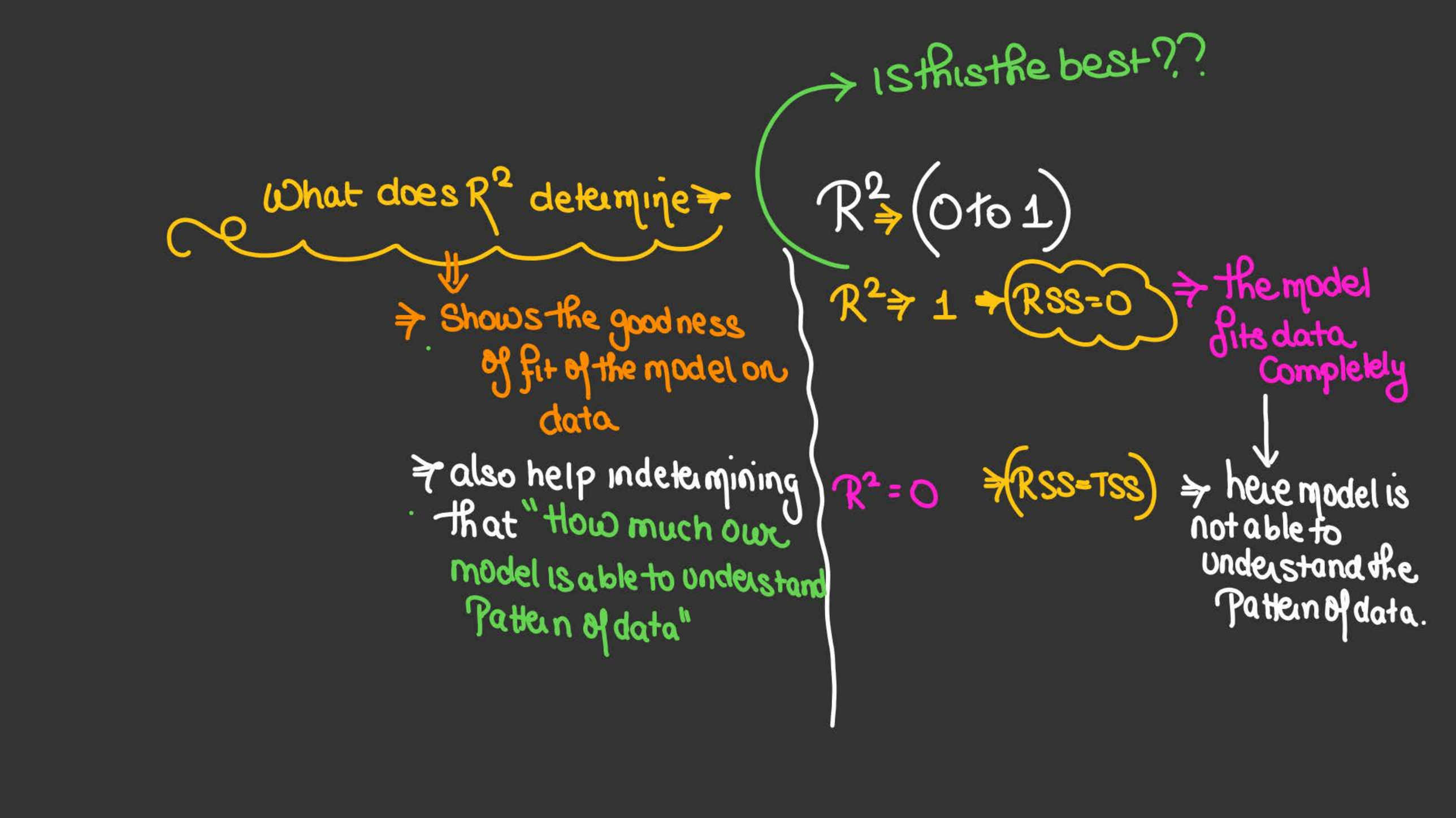
$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

TSS = total sum of squares

n = number of observations

y_i = value in a sample

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


Is this the best??
What does R^2 determine
$R^2 \Rightarrow$ (0 to 1)
$\Rightarrow$ Shows the goodness of fit of the model on data
$\Rightarrow$ also help in determining that "How much our model is able to understand Pattern of data"
$R^2 \Rightarrow 1$ $\Rightarrow$ (RSS=0) $\Rightarrow$ the model fits data Completely
$R^2 = 0$ $\Rightarrow$ (RSS=TSS) $\Rightarrow$ here model is not able to understand the Pattern of data.

$R^2 = 1 \Rightarrow$ is this the best model ??

→ overfitting $\Rightarrow$ noise in data is given importance

So $R^2 \neq 1$

But $R^2 \Rightarrow .8$ to $.9$ is a good Number.

because RSS $\neq 0$

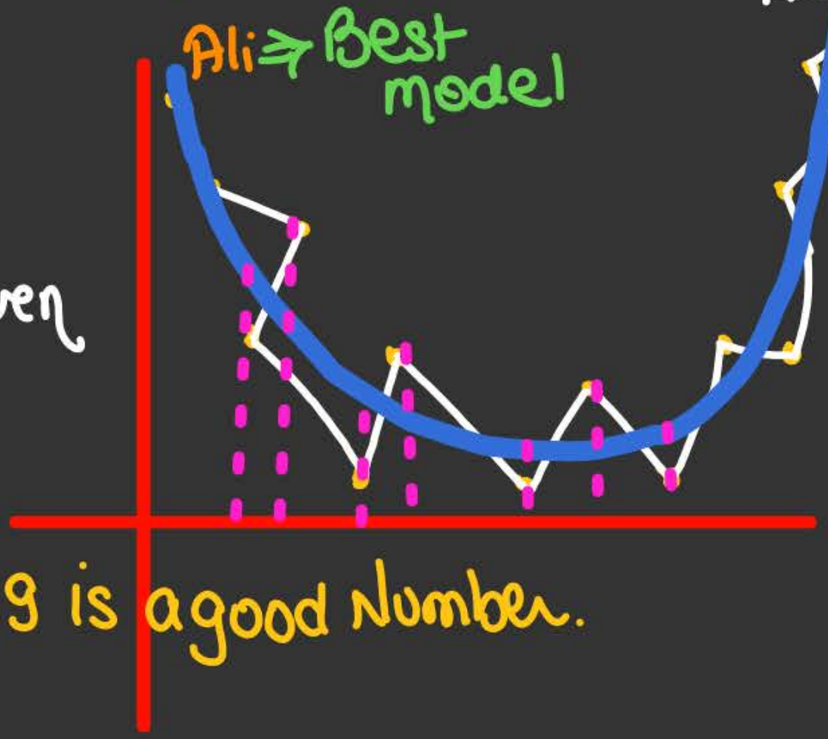
$R^2 \neq 1$

$R^2 = 1$

$y_i = \hat{y}_i$

RSS $= 0$

Ali $\Rightarrow$ Best model



Linear Regression

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. $R^2=1, RSS=0$, overfitting.
- However, we get R-square equals 0 when the model does not predict any variability in the model. ⇒ model donot understand data Pattern.

So in data we have x, y

In ML we find a function that relate y with x

$y = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 ----$

↓ dependent Variable

↓ Independent Variable

- (So using the Ind. Variable we try to predict the dep. Variable)
- If R^2 is close to 1 then the ^ function created by Indep. Variable is able to predict the dep. Variable.
- So R^2 determine how well the fxn of Ind. Variable understand variation, Pattern of y.

Linear Regression

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.

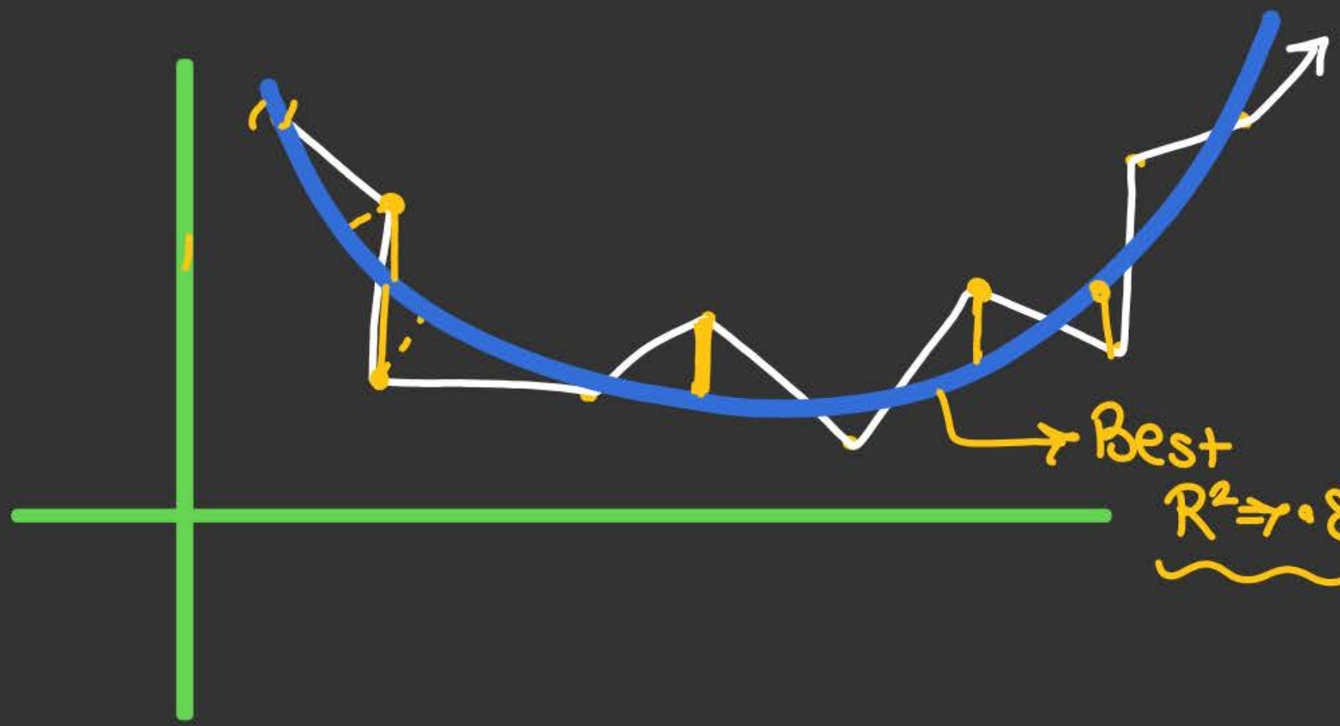
So we can say that

If $R^2 = 1$ ⇒ then owe model understand 100% of pattern of data

$R^2 = 0$ ⇒ " " " donot " the pattern

↳ understand 0% of the pattern.

$R^2 = .64$ ⇒ So model understand 64% of pattern of data.



100% data variation is follow

$RSS = 0$

$R^2 = 1$

overfitting

Linear Regression

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. .

Linear Regression

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

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

Linear Regression

Considering data of P Dimensions

Lets solve a question

Question 2: Given a simple linear regression model with an R-squared value of 0.64, what percentage of the variation in the dependent variable is explained by the predictor variable?

→ 64.1.

x^1, x^2, x^3, x^4

→ Independent Var

→ Predictor Var

Linear Regression

Considering data of P Dimensions

Lets solve a question

Question 6: In a simple linear regression model, if the coefficient of determination (R-squared) is 0.81 and the total sum of squares (SST) is 400, what is the sum of squared errors (SSE)?

a)76
b)77
c)54
d)33

SSE → RSS.

SST → TSS

$R^2 = .81$, $TSS = 400$

$RSS = ?$

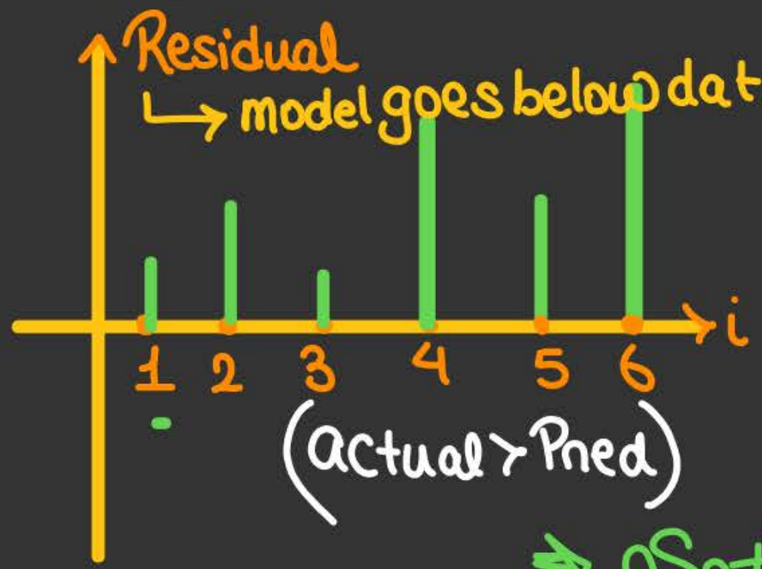
$$R^2 = 1 - \frac{RSS}{TSS}$$

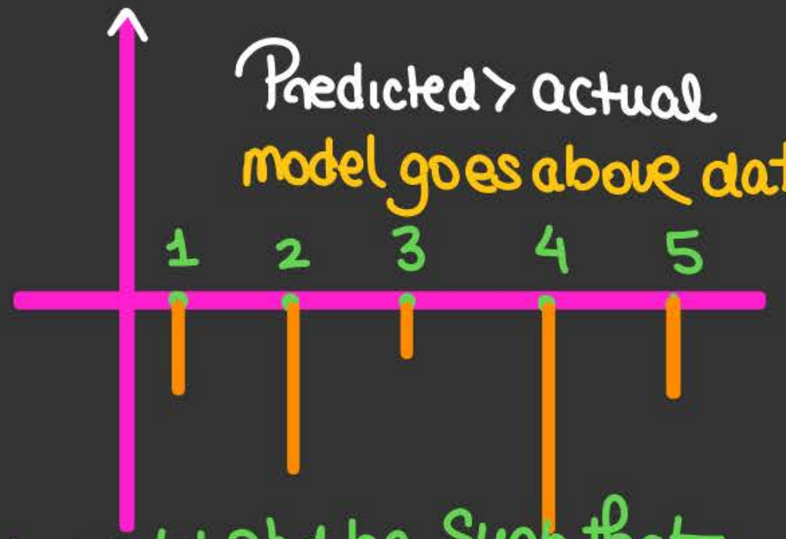
$$.19 \times 400 = RSS$$

$$76 = RSS$$

Residual Plot - Concept

$$\text{Residual} = (y_i - \hat{y}_i) = (\text{Actual Value} - \text{Predicted Value})$$





⇒ So the best model shd be such that the model shd pass through middle of data

⇒ Residual Plot shd be +ve and -ve both.

Linear Regression

What is Mean Square Error

$$\text{Mean} = \frac{\text{Sum of values}}{\text{Total No of values}}$$

Squared error $\Rightarrow (y_i - \hat{y}_i)^2$

- MSE $\Rightarrow \dfrac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}$

Linear Regression

What is Root Mean Square Error

$$RMSE = \sqrt{MSE}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Linear Regression

Question 5: In a simple linear regression analysis, if the mean of the dependent variable (Y) is 50, and the slope coefficient (a) is 3, what is the mean of the predictor variable (X) [illegible]

a) Cannot be determined without the value of the intercept (b).

b) Cannot be determined without the value of the intercept (a)

c)Can be determined without the value of the intercept (b)

$\bar{Y} = 50$,

$m = 3$

$y = ax + b$

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

$b = 50 - 3\bar{x}$

(we Cannot find $\bar{x}$ until we have b)

Linear Regression

Question 9: In a simple linear regression analysis, if the sum of squared errors (SSE) is 120 and the degrees of freedom for residuals is 15, what is the mean squared error (MSE)?

No of data points

H.W

$\frac{120}{15}$

Linear Regression

Question 15: What is the purpose of the coefficient of determination (R-squared) in simple linear regression?

R^2 (B) H.W

A. To determine the slope of the regression line
B. To measure the strength of the linear relationship
C. To calculate the p-value of the regression
D. To identify outliers in the dataset

Linear Regression

Question 19: If the R-squared value in simple linear regression is 0.75, what does it indicate?

.75 A

Siddharth Sir AI/ML ✓

A. A strong linear relationship between the variables
B. A weak linear relationship between the variables
C. No linear relationship between the variables
D. The model is overfitting

Linear Regression

Question 2: What does the coefficient of determination (R-squared) measure in multiple linear regression?

B

A. The correlation between predictor variables
B. The percentage of variance in the dependent variable explained by the model
C. The significance of the intercept term
D. The number of predictor variables in the model

Linear Regression

In multiple linear regression, what is the key difference between simple linear regression and multiple linear regression?

A) Simple linear regression has one independent variable, while multiple linear regression has two or more.
B) Simple linear regression uses categorical variables, while multiple linear regression uses continuous variables.
C) Simple linear regression is used for classification, while multiple linear regression is used for prediction.
D) There is no difference between simple and multiple linear regression.

Linear Regression

Which statistic is used to assess the strength and direction of the relationship between the dependent variable and each independent variable in multiple linear regression?

A) Mean absolute error (MAE)
B) R-squared (R^2)
C) Standard error
D) Confidence interval

Linear Regression

What is the purpose of the residual plot in multiple linear regression analysis?

A) To visualize the relationship between independent variables.
B) To check for homoscedasticity and the presence of outliers.
C) To calculate the correlation coefficient (r).
D) To assess multicollinearity.

Linear Regression

What is the main purpose of the intercept term in a multiple linear regression model?

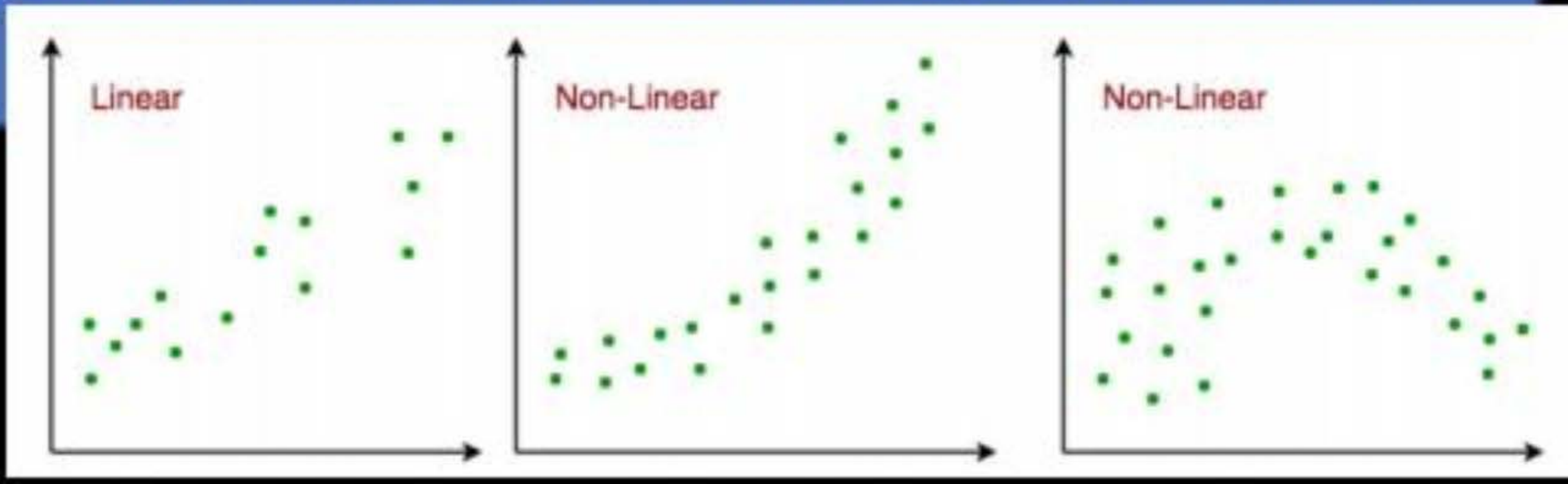
A) It represents the slope of the regression line.
B) It is used to control for multicollinearity.
C) It represents the expected value of the dependent variable when all independent variables are zero.
D) It is not used in multiple linear regression.

Linear Regression

Assumptions in Linear Regression

Linear regression needs to meet a few conditions in order to be accurate and dependable solutions.

1. **Linearity**: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then linear regression will not be an accurate model.



Linear Regression

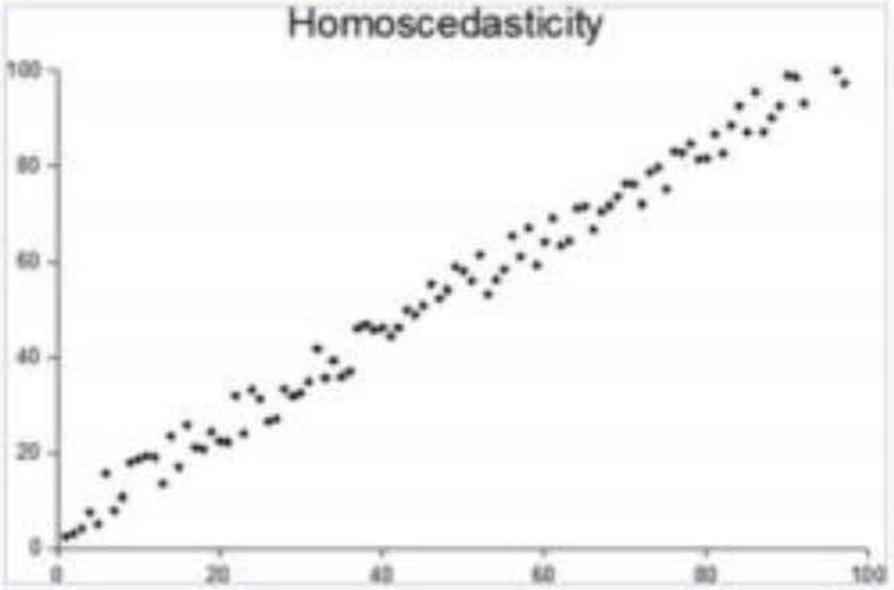
Assumptions in Linear Regression

2. Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation. If the observations are not independent, then linear regression will not be an accurate model.

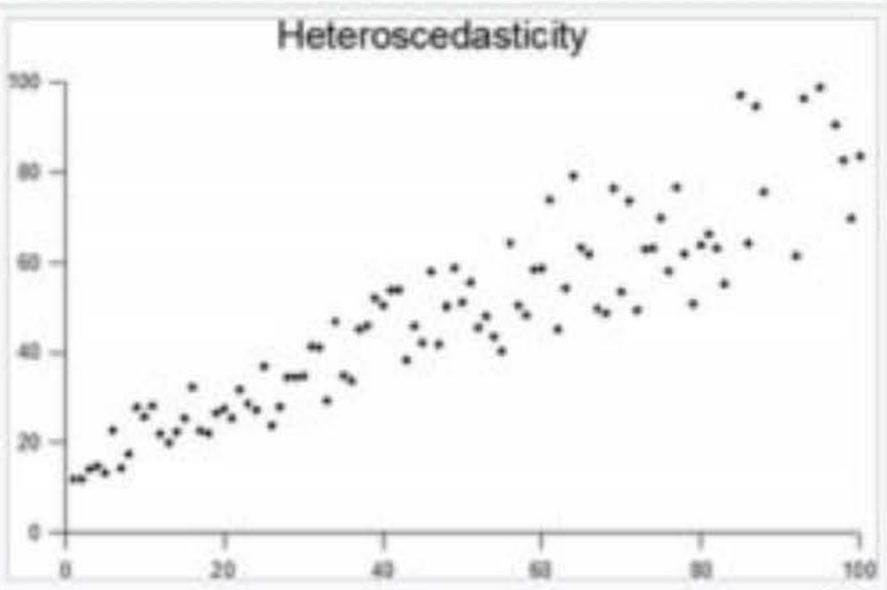
Linear Regression

Assumptions in Linear Regression

3. Homoscedasticity:



Plot with random data showing homoscedasticity: at each value of x, the y-value of the dots has about the same variance.



Plot with random data showing heteroscedasticity: The variance of the y-values of the dots increase with increasing values of x.

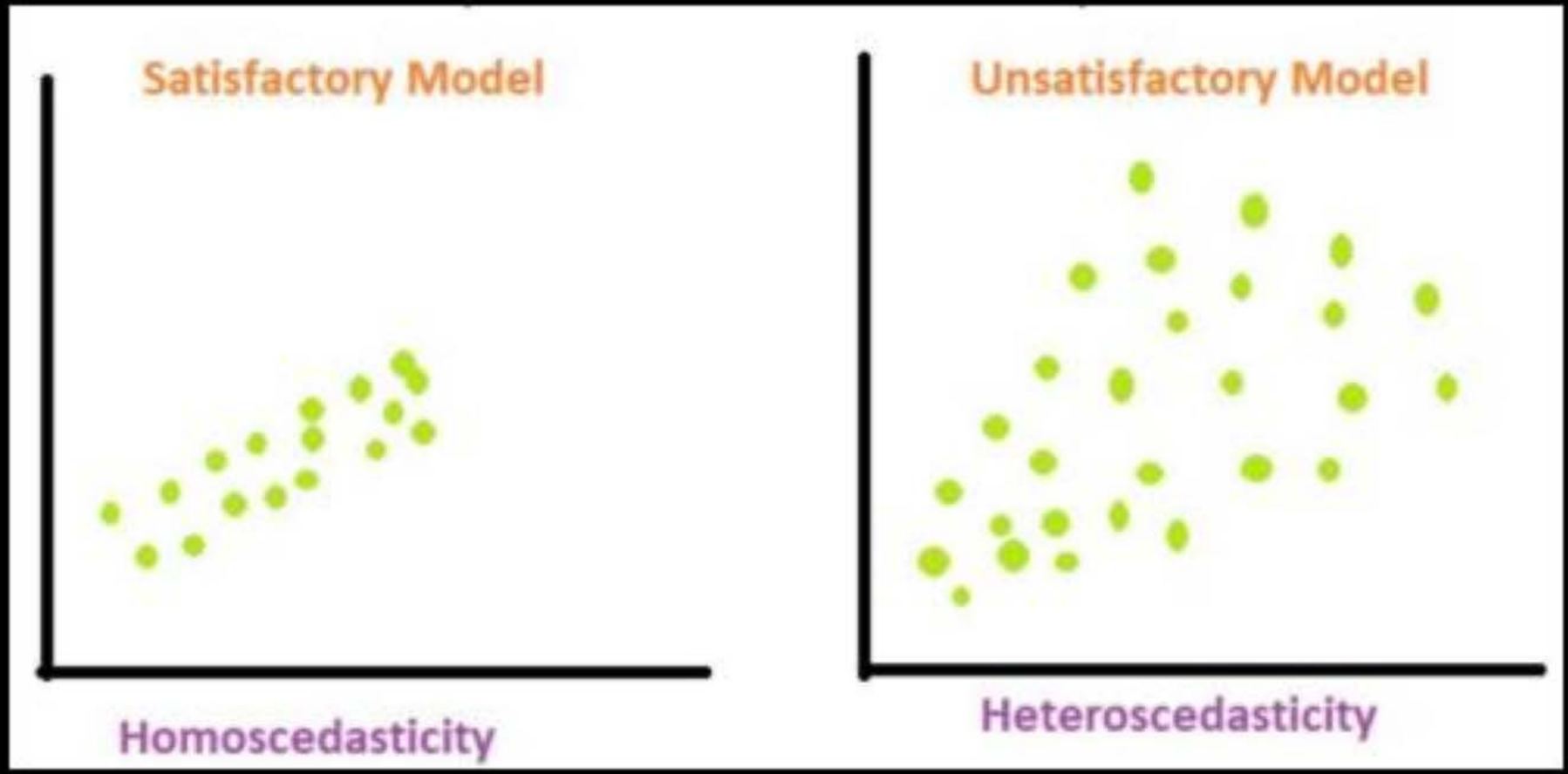
Linear Regression

Assumptions in Linear Regression

3. Homoscedasticity: Heteroscedasticity means unequal scatter. In regression analysis, we talk about heteroscedasticity in the context of the residuals or error term. Specifically, heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. Heteroscedasticity is a problem because ordinary least squares (OLS) regression assumes that all residuals are drawn from a population that has a constant variance (homoscedasticity).

Linear Regression

Assumptions in Linear Regression



Linear Regression

Assumptions in Linear Regression

4. No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable. If there is multicollinearity, then multiple linear regression will not be an accurate model.

Linear Regression

Assumptions in Linear Regression

Detecting Multicollinearity includes two techniques:

- Correlation Matrix: Examining the correlation matrix among the independent variables is a common way to detect multicollinearity. High correlations (close to 1 or -1) indicate potential multicollinearity.
- VIF (Variance Inflation Factor): VIF is a measure that quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. A high VIF (typically above 10) suggests multicollinearity.

Linear Regression

Assumptions in Linear Regression

Correlation between two variables :

$$\text{Correlation} = \rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Topics Left
- P value concept
- Spatial and time complexity of Linear Regression
- VIF
- Questions

2 mins Summary

- Topic
- Topic
- Topic
- Topic
- Topic

THANK - YOU