SNS COLLEGE OF TECHNOLOGY

Kurumbapalayam (Po), Coimbatore – 641 035

An Autonomous Institution

Accredited by NAAC – UGC with 'A++' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

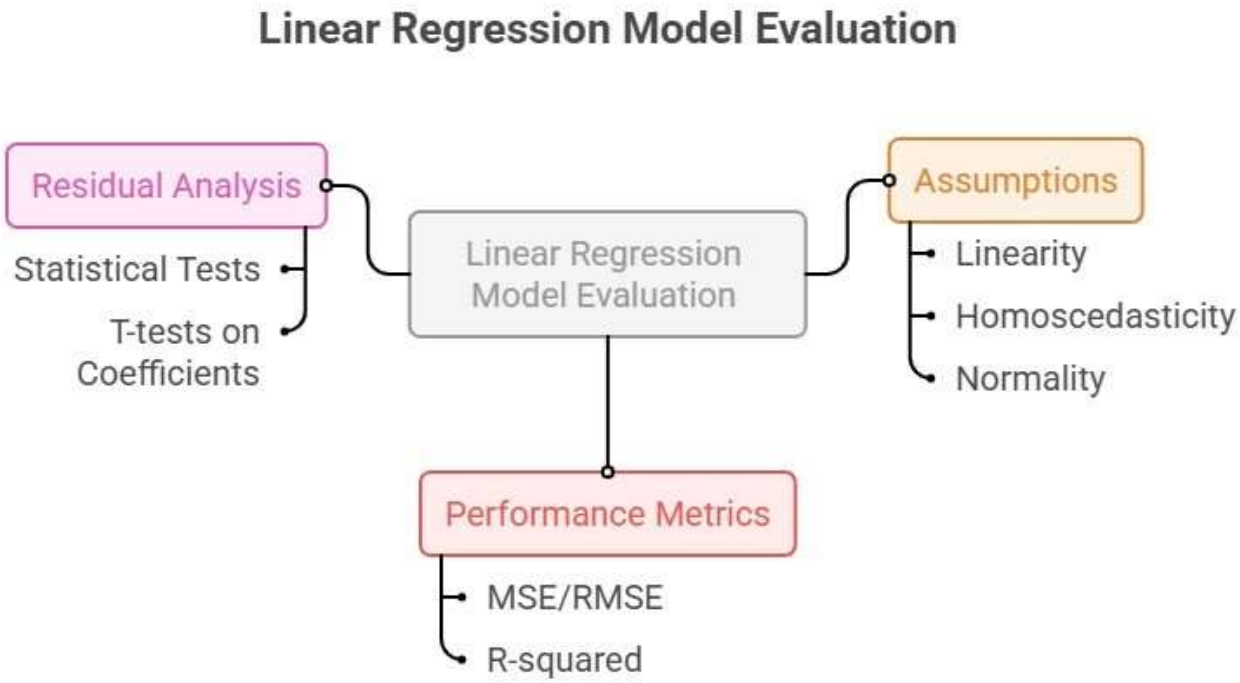
COURSE NAME: 23ADO201- DATA SCIENCE FUNDAMENTALS (OPEN ELECTIVE)

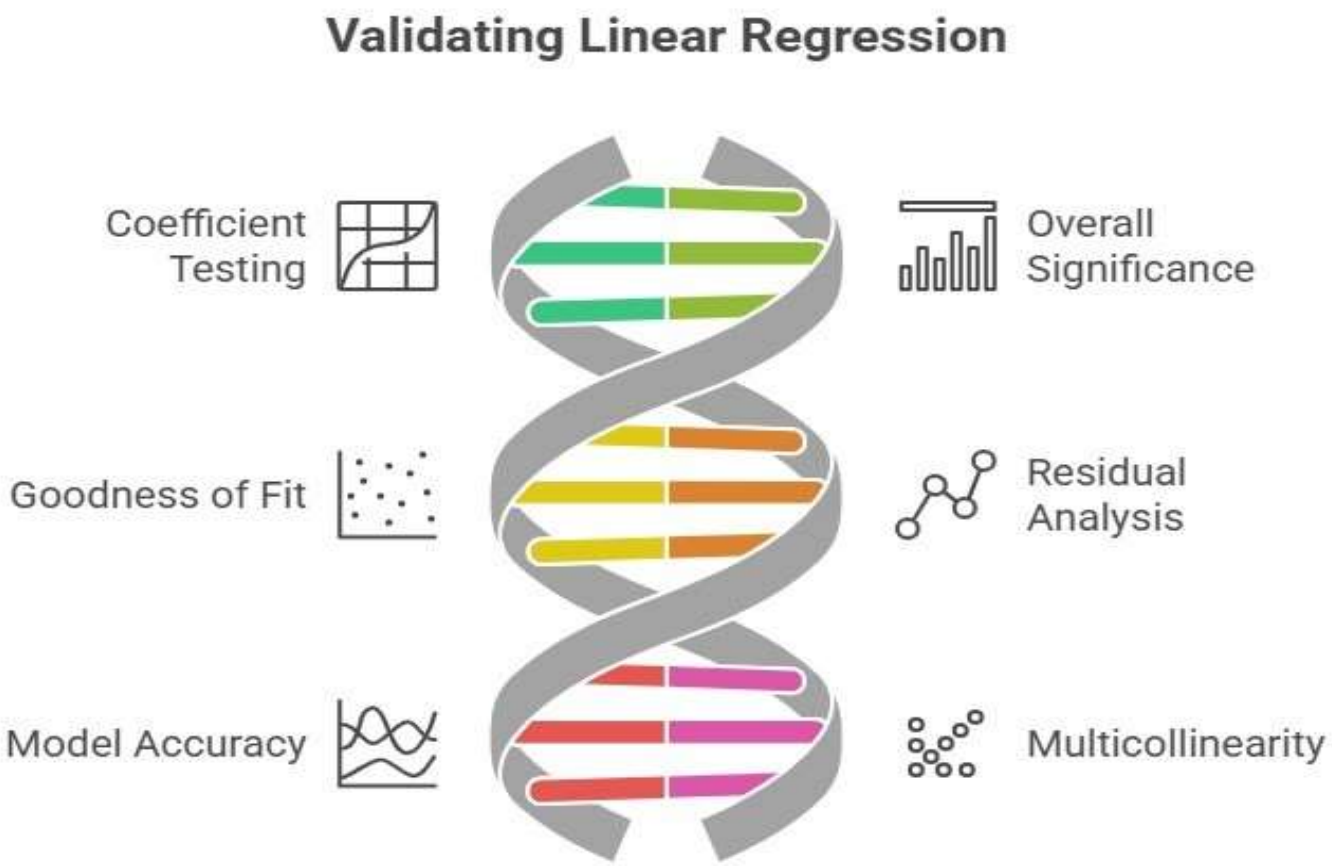
III YEAR / VI SEMESTER

Unit 5 - PREDICTIVE ANALYTICS

Topic : TESTING A LINEAR MODEL

Empathy


Linear Regression Model Evaluation
Residual Analysis
Statistical Tests
T-tests on Coefficients
Linear Regression Model Evaluation
Assumptions
Linearity
Homoscedasticity
Normality
Performance Metrics
MSE/RMSE
R-squared


Validating Linear Regression
Coefficient Testing
Overall Significance
Goodness of Fit
Residual Analysis
Model Accuracy
Multicollinearity

Python Implementation of Linear Regression

1. Import the necessary libraries

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
```

2. Generating Random Dataset

```
np.random.seed(42)

X = np.random.rand(50, 1) * 100

Y = 3.5 * X + np.random.randn(50, 1) * 20
```

3. Creating and Training Linear Regression Model

```
model = LinearRegression()
model.fit(X, Y)
```

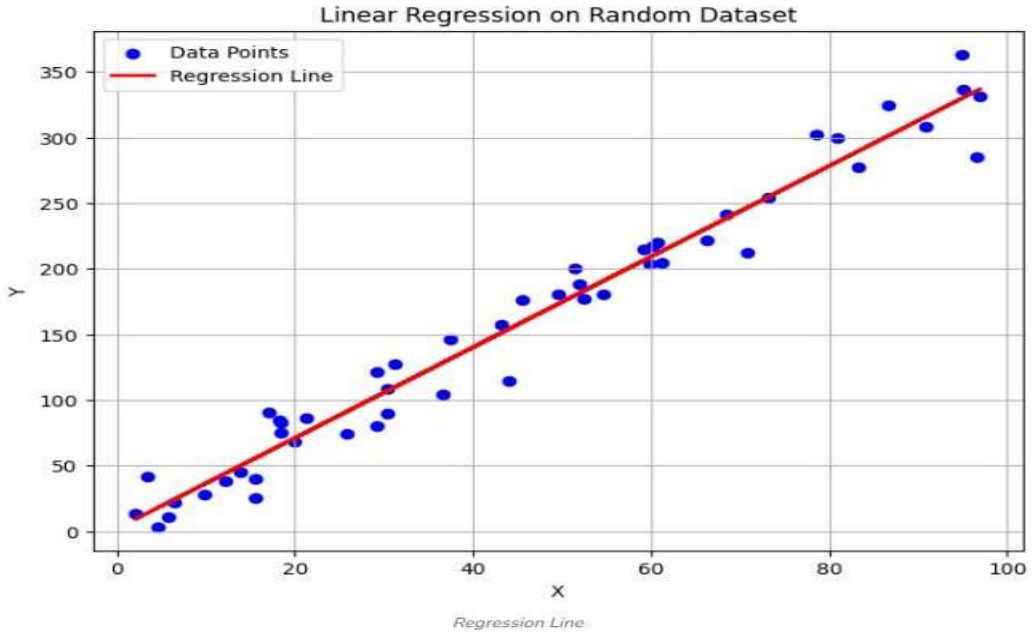
4. Predicting Y Values

```
Y_pred = model.predict(X)
```

5. Visualizing the Regression Line

```
plt.figure(figsize=(8,6))
plt.scatter(X, Y, color='blue', label='Data Points')
plt.plot(X, Y_pred, color='red', linewidth=2, label='Regression Line')
plt.title('Linear Regression on Random Dataset')
plt.xlabel('X')
plt.ylabel('Y')
plt.legend()
plt.grid(True)
plt.show()
```

Output:



Regression Line

ACTIVITY: Testing a Linear Model

🎯 **Objective**

To test the performance of a linear regression model using **test data, error metrics, and R^2 score**.

Given Data

x	y
1	2
2	3
3	5
4	4
5	6

Student Tasks

1.Split dataset into training and testing
2.Find slope (m) and intercept (c)
3.Predict test values
4.Compute MSE & RMSE
5.Interpret model performance

MCQ

MCQs on Testing a Linear Model (based on Linear Regression & Least Squares):

1. Testing a linear model primarily checks:

A) Data size

B) Model accuracy and validity

C) Number of variables

D) Execution time

Answer: B) Model accuracy and validity

2. Which dataset is used to evaluate model performance?

A)Training data

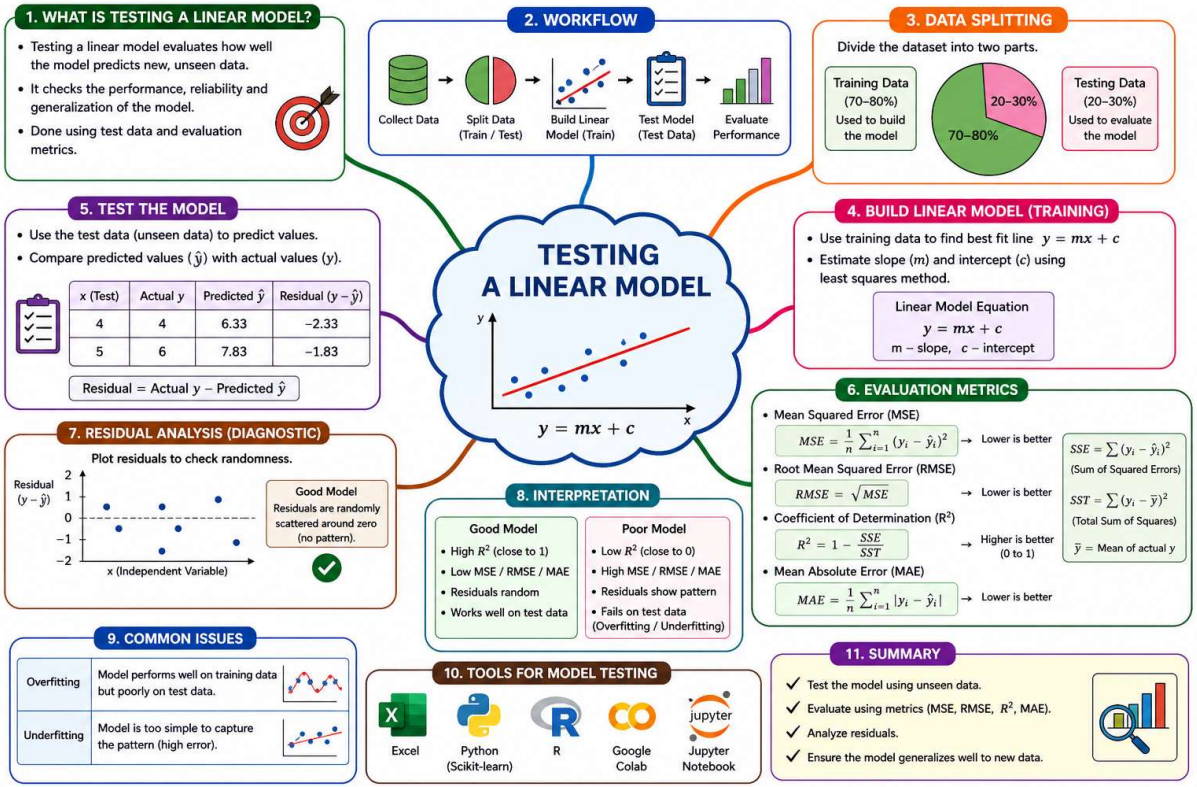
B) Testing data

C) Random data

D) Sample data

Answer: B) Testing data

MINDMAP


1. WHAT IS TESTING A LINEAR MODEL?
• Testing a linear model evaluates how well the model predicts new, unseen data.
• It checks the performance, reliability and generalization of the model.
• Done using test data and evaluation metrics.
2. WORKFLOW
Collect Data
Split Data (Train / Test)
Build Linear Model (Train)
Test Model (Test Data)
Evaluate Performance
3. DATA SPLITTING
Divide the dataset into two parts.
Training Data (70–80%) Used to build the model
20–30%
70–80%
Testing Data (20–30%) Used to evaluate the model
TESTING A LINEAR MODEL
y = mx + c
5. TEST THE MODEL
• Use the test data (unseen data) to predict values.
• Compare predicted values ($\hat{y}$) with actual values (y).
x (Test) | Actual y | Predicted ŷ | Residual (y − ŷ)
4 | 4 | 6.33 | −2.33
5 | 6 | 7.83 | −1.83
Residual = Actual y − Predicted ŷ
4. BUILD LINEAR MODEL (TRAINING)
• Use training data to find best fit line $y = mx + c$
• Estimate slope (m) and intercept (c) using least squares method.
Linear Model Equation
$y = mx + c$
m – slope, c – intercept
6. EVALUATION METRICS
• Mean Squared Error (MSE) $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ → Lower is better
• Root Mean Squared Error (RMSE) $RMSE = \sqrt{MSE}$ → Lower is better
• Coefficient of Determination (R^2) $R^2 = 1 - \frac{SSE}{SST}$ → Higher is better (0 to 1)
• Mean Absolute Error (MAE) $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ → Lower is better
$SSE = \sum(y_i - \hat{y}_i)^2$ (Sum of Squared Errors)
$SST = \sum(y_i - \bar{y})^2$ (Total Sum of Squares)
$\bar{y}$ = Mean of actual y
7. RESIDUAL ANALYSIS (DIAGNOSTIC)
Plot residuals to check randomness.
Residual (y − ŷ)
x (Independent Variable)
Good Model
Residuals are randomly scattered around zero (no pattern).
8. INTERPRETATION
Good Model
• High R^2 (close to 1)
• Low MSE / RMSE / MAE
• Residuals random
• Works well on test data
Poor Model
• Low R^2 (close to 0)
• High MSE / RMSE / MAE
• Residuals show pattern
• Fails on test data (Overfitting / Underfitting)
9. COMMON ISSUES
Overfitting | Model performs well on training data but poorly on test data.
Underfitting | Model is too simple to capture the pattern (high error).
10. TOOLS FOR MODEL TESTING
Excel
Python (Scikit-learn)
R
Google Colab
Jupyter Notebook
11. SUMMARY
✓ Test the model using unseen data.
✓ Evaluate using metrics (MSE, RMSE, R^2, MAE).
✓ Analyze residuals.
✓ Ensure the model generalizes well to new data.

TEXT BOOKS	
1.	Srinivasan Desikan and Gopalaswamy Ramesh, ―Software Testing – Principles and Practices, Pearson Education, 2006.
REFERENCES	
1.	Ilene Burnstein, ―Practical Software Testing, Springer International Edition, 2003
2.	Edward Kit, Software Testing in the Real World – Improving the Process‖, Pearson Education, 1995
3.	Boris Beizer, Software Testing Techniques – 2nd Edition, Van Nostrand Reinhold, New York, 1990.
4.	Aditya P. Mathur, ―Foundations of Software Testing _ Fundamental Algorithms and Techniques, Dorling Kindersley (India) Pvt. Ltd., Pearson Education, 2008.

THANK YOU!

10