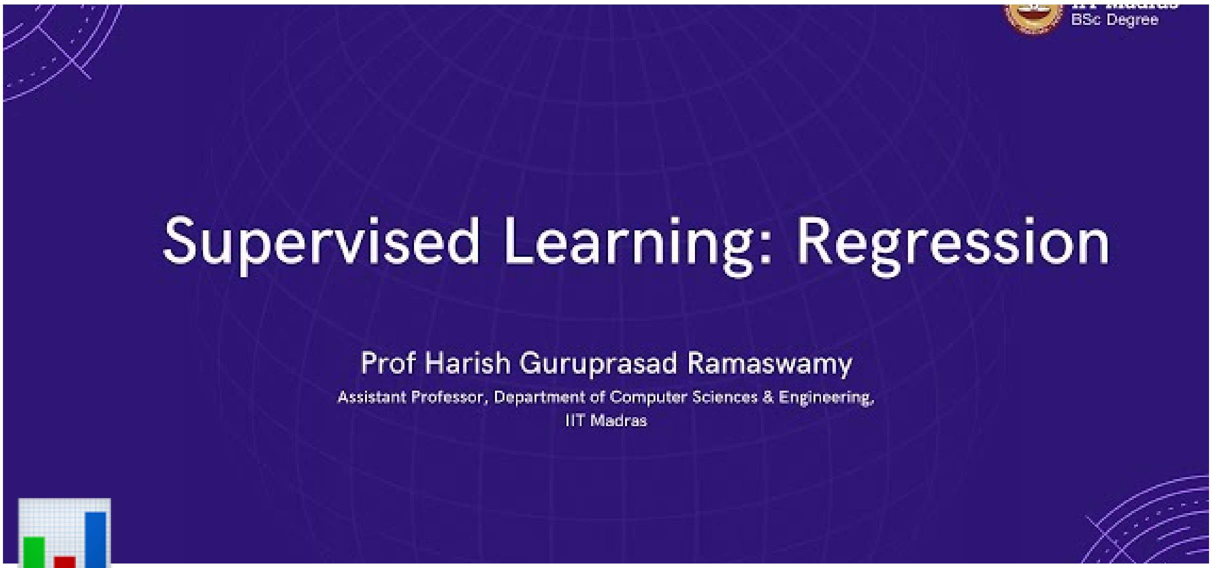


Module 3 | Linear Regression

🎯 1. Learning Goals

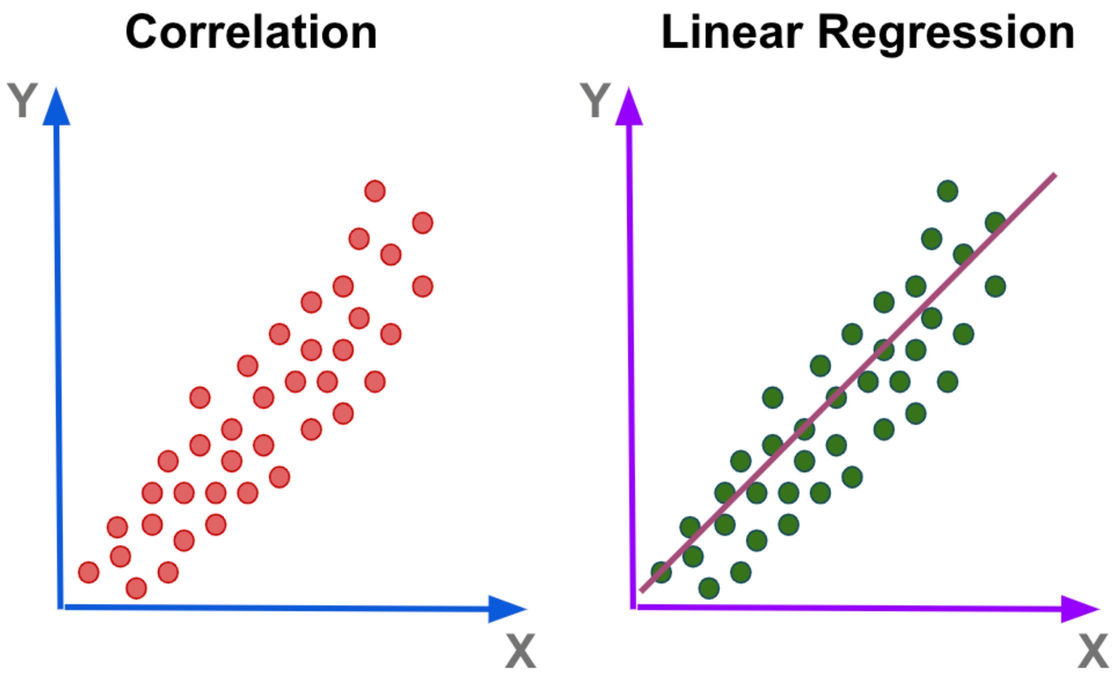
In this section, you will learn how to:

- Understand what **Linear Regression** is and how it works
- Know how to **measure error** and **model fit**
- Apply linear regression models using **Python (Scikit-learn)**

🧠 2. Introduction to Linear Regression

General Concept

Linear Regression is a statistical method that models a **linear relationship** between a dependent variable y and one or more independent variables $x_1, x_2, \ldots, x_n$



General Equation

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}$$

Parameter Explanation

Symbol	Name	Explanation
$\hat{y}_i$	Predicted value	The predicted value of the dependent variable (output) for observation i
β_0	Intercept	The average value of y when **all independent variables equal 0**
β_j	Coefficient	The average change in y when x_j increases by 1 unit, holding other variables constant
x_{ij}	Independent variable	Independent variable (input feature)
ε_i	Error term	The deviation between the actual value y_i and the predicted value $\hat{y}_i$

Example

Sample Data:

Hours Studied (x)	Exam Score (y)
2	50
4	60
6	70
8	80

→ **Trained model:** $\hat{y} = 40 + 5x$

Interpretation:

- β_0 = 40: If no hours studied, predicted score = **40 points**
- β_1 = 5: Each additional hour studied increases score by an average of **5 points**

⚙️ 3. Calculating the Residuals

Residual Formula:

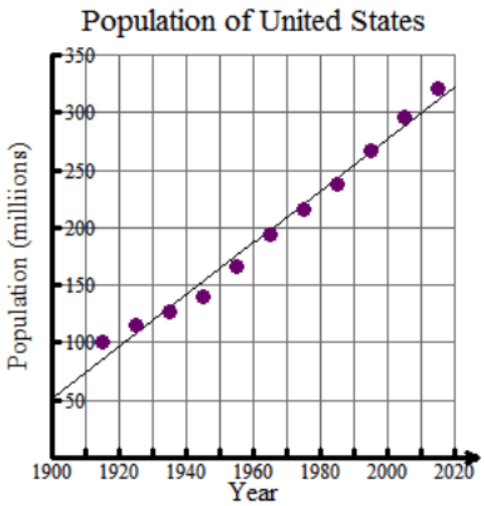
$$e_i = y_i - \hat{y}_i$$

The residual measures the **difference** between the actual and predicted values.

Explanation

Symbol	Meaning
y_i	Actual value (observed value)
$\hat{y}_i$	Predicted value from the model
e_i	**Residual** – error at data point $ `i` $, showing how much the model deviates from reality

---



$$Population = 2.26(Year) - 4235.96$$

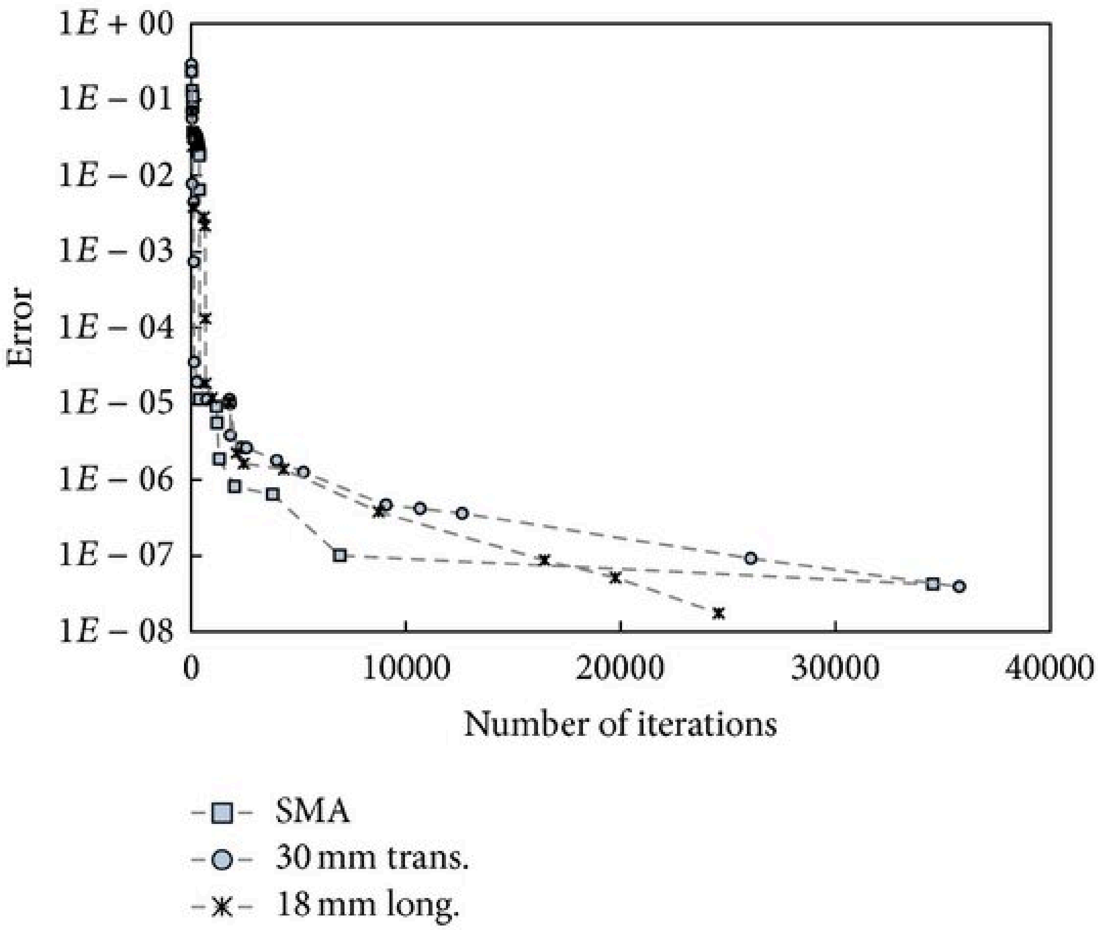
🧮 4. Minimizing the Error Function

Goal: Minimize the **Sum of Squared Errors (SSE)**

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

Key Terms

Symbol	Meaning
SSE	**Sum of Squared Errors** – total squared error
n	Number of observations (samples)
$y_i - \hat{y}_i$	Distance between actual and predicted values



We find the coefficients $\beta_0, \beta_1, \ldots, \beta_n$ such that:

$$\min_{\beta} SSE$$

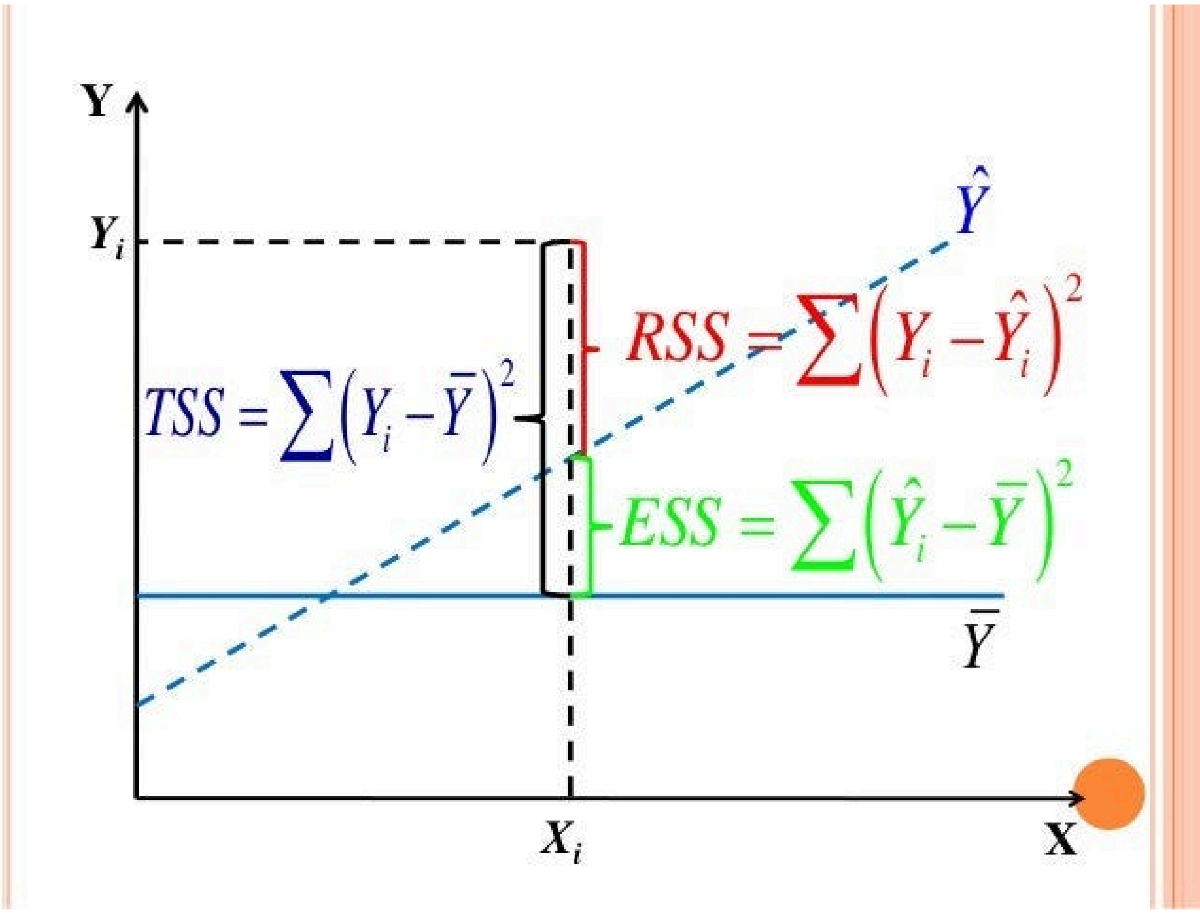
Example Calculation

x	y	ŷ	e	e²
2	52	50	2	4
4	58	60	-2	4
6	70	70	0	0

Result: $SSE = 4 + 4 + 0 = 8$

→ Total sum of squared errors = **8**

📏 5. Other Measures of Error



(1) Total Sum of Squares (TSS)

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

TSS measures the **total variation** in the actual data.

Symbol	Meaning
TSS	Total variation in the actual data
$\bar{y}$	Mean of actual values y_i

(2) Explained Sum of Squares (ESS)

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

ESS measures the **variation explained** by the model.

Symbol	Meaning
ESS	Variation explained by the model
$\hat{y}_i - \bar{y}$	Deviation between predicted value and actual mean

(3) Relationship Between the Three Metrics

The Fundamental Equation:

$$TSS = ESS + SSE$$

Total variation = Explained variation + Unexplained error

(4) Coefficient of Determination (R^2)

R^2 Formula:

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

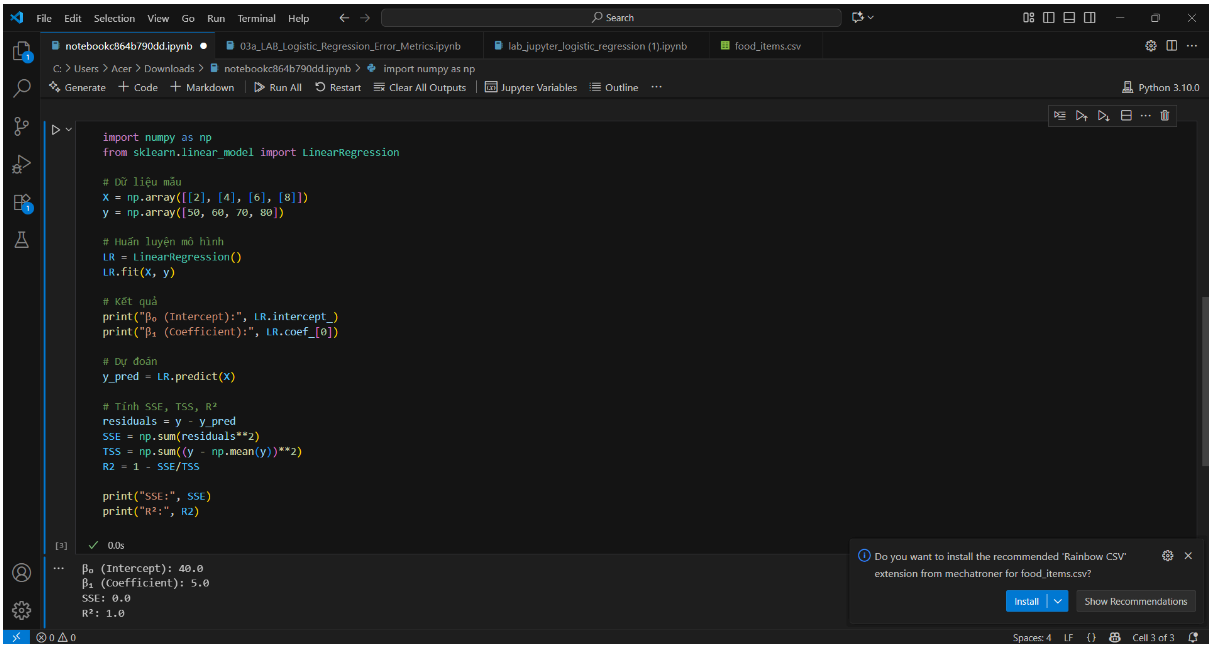
R^2 measures how well the model fits the data (ranges from 0 to 1).

Symbol	Meaning
R^2	**Coefficient of determination** – measures model fit quality
SSE/TSS	Proportion of variation **not** explained by the model

Example:
$SSE = 8, \quad TSS = 168 \Rightarrow R^2 = 1 - \frac{8}{168} = 0.952$
→ The model explains **95.2%** of the variation in the data!

💻 6. Code Implementation (Python)

```
import numpy as np
from sklearn.linear_model import LinearRegression

# Sample data
X = np.array([[2], [4], [6], [8]])
y = np.array([50, 60, 70, 80])

# Train the model
LR = LinearRegression()
LR.fit(X, y)

# Results
print("β₀ (Intercept):", LR.intercept_)
print("β₁ (Coefficient):", LR.coef_[0])

# Predictions
y_pred = LR.predict(X)

# Calculate SSE, TSS, R²
residuals = y - y_pred
SSE = np.sum(residuals**2)
TSS = np.sum((y - np.mean(y))**2)
R2 = 1 - SSE/TSS

print("SSE:", SSE)
print("R²:", R2)
```



📊 7. Summary Table of Formulas & Parameters

Symbol	Name	Formula	Explanation
$\hat{y}_i$	**Prediction**	$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_n x_{in}$	Linear regression function
e_i	**Residual**	$e_i = y_i - \hat{y}_i$	Error of observation i
SSE	**Sum of Squared Errors**	$SSE = \sum(y_i - \hat{y}_i)^2$	Total squared error
TSS	**Total Sum of Squares**	$TSS = \sum(y_i - \bar{y})^2$	Total variation in data
ESS	**Explained Sum of Squares**	$ESS = \sum(\hat{y}_i - \bar{y})^2$	Variation explained by model
MSE	**Mean Squared Error**	$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$	Average squared error
R^2	**Coefficient of Determination**	$R^2 = 1 - \frac{SSE}{TSS}$	Model fit assessment

🧠 8. Recap

Key Takeaways:

- **Linear Regression** is a linear model between x and y
- Parameters β are estimated by **minimizing SSE**
- Metrics SSE, MSE, R^2 are used to **evaluate accuracy**
- Higher R^2 → **better model fit**