Understanding Linear Regression

Machine Learning and Data Science Club

What is Linear Regression?

01	Definition	Regression is a supervised machine learning method used to model and analyze the relationship between one or more input/predictor/x variables and a target/explanatory/y variable. It helps identify patterns and quantify how changes in inputs affect the output.
02	Goal	The primary aim is to model a function that best describes the relationship between variables so we can predict future or unseen values. In practice, regression also helps with interpretation, showing which factors are most influential.
03	Output	Regression produces numerical predictions (not categories). Examples include predicting housing prices, estimating sales revenue, or forecasting temperature. The output can be a single value or multiple continuous values depending on the task.
04	Applications	Widely used across domains — finance (stock return modeling), healthcare (disease progression prediction), economics (demand forecasting), engineering (sensor data modeling), and environmental science (climate trend analysis).

Simple Linear Regression

Definition: A regression method that models the relationship between one predictor (independent variable) and one target (dependent variable).

Equation: $y = \beta 0 + \beta 1x + \varepsilon$

β0: intercept (value of y when x = 0)

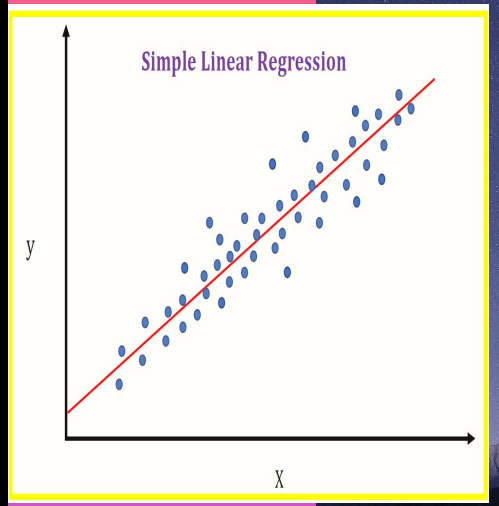
β1: slope (change in y for a one-unit change in x)

ε: error term (unexplained variation)

Goal: Find the line of best fit that minimizes the error between predicted and actual values.

Output: Predicts a continuous numerical value for y given x.

Applications:
-Predicting weight from height
-Estimating sales from advertising spend
-Forecasting temperature based on time of day



Multiple Linear Regression

Definition: A regression method that models the relationship between two or more predictors (independent variables) and one target (dependent variable).

Equation: $y = \beta 0 + \beta 1 x1 + \beta 2 x2 + \ldots + \beta k xk + \varepsilon$

β_0: intercept (value of y when all predictors = 0)

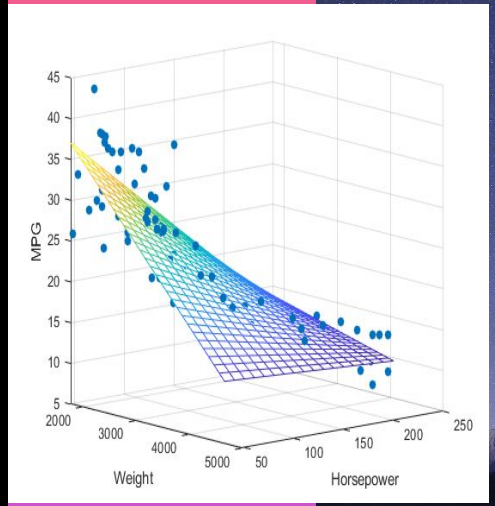
$\beta_1 \ldots \beta\square$: coefficients (effect of each predictor on y, holding others constant)

ε: error term (unexplained variation)

Goal:Find the best-fitting plane (or hyperplane in higher dimensions) that minimizes the error between predicted and actual values.

Output:Predicts a continuous numerical value for y given multiple predictors.

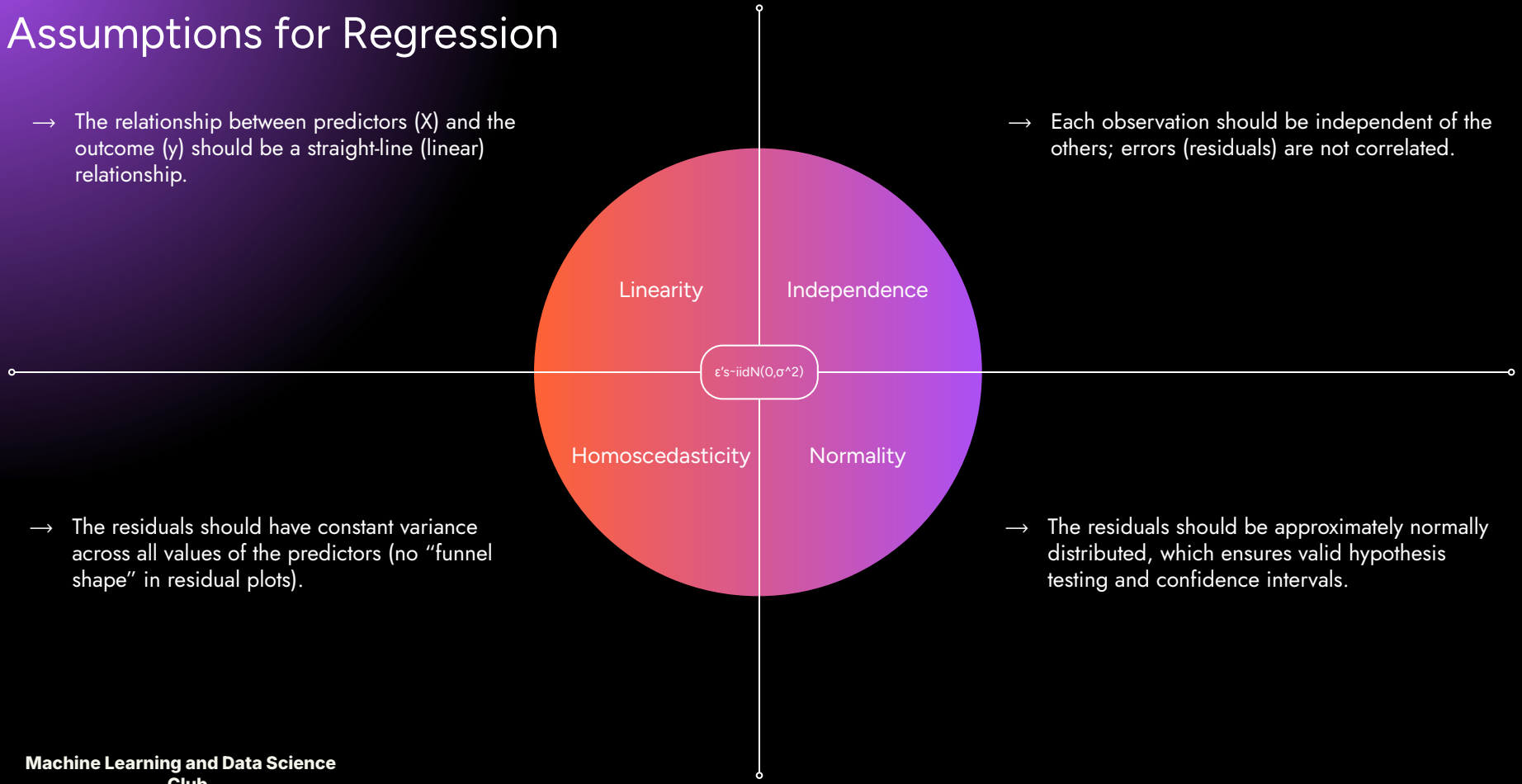
Applications:
-Predicting housing prices from size, location, and age
-Estimating exam scores from study hours, attendance, and sleep
-Forecasting demand from price, income, and advertising spend



Assumptions for Regression

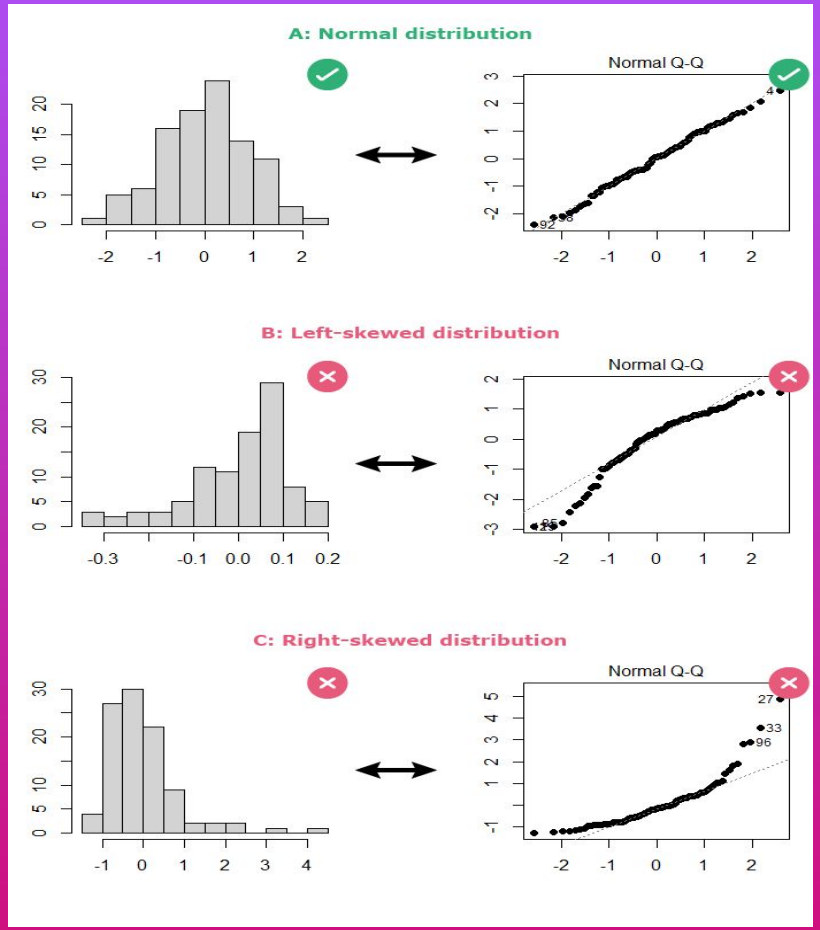
→ The relationship between predictors (X) and the outcome (y) should be a straight-line (linear) relationship.

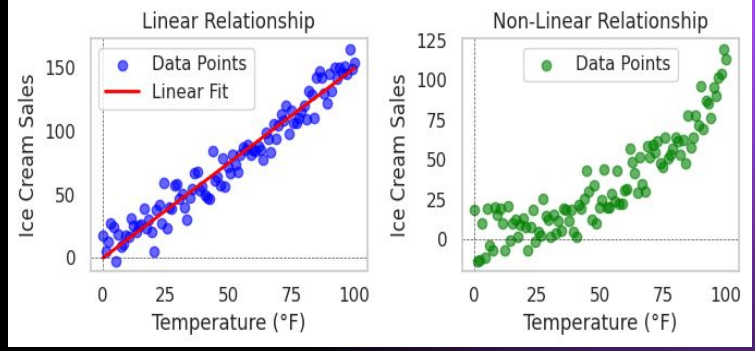
→ Each observation should be independent of the others; errors (residuals) are not correlated.

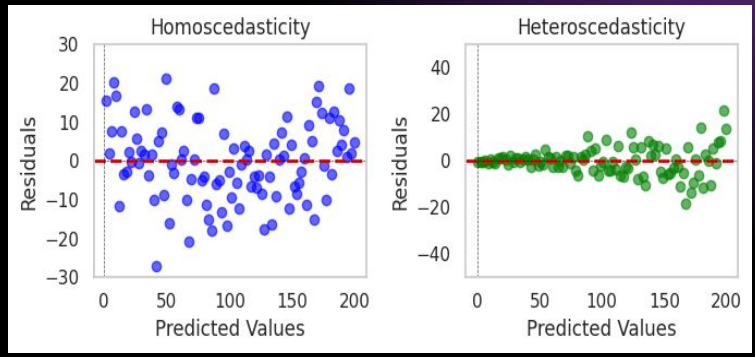


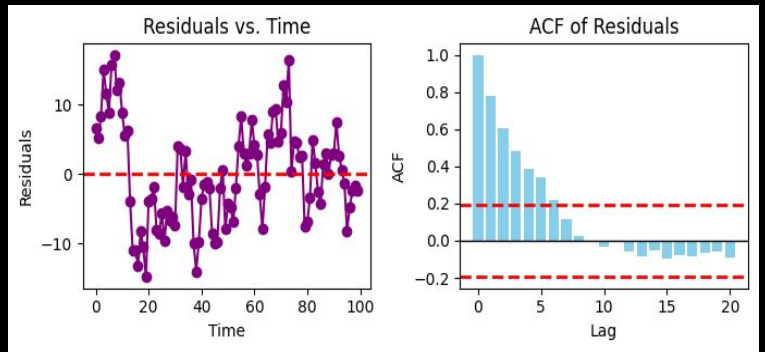
→ The residuals should have constant variance across all values of the predictors (no "funnel shape" in residual plots).

→ The residuals should be approximately normally distributed, which ensures valid hypothesis testing and confidence intervals.









Machine Learning and Data Science Club

How to measure model fit

Mean Absolute Error (MAE):

-Measures the average magnitude of errors in predictions.

-Units are the same as the target variable.

-Less sensitive to outliers than MSE.

Mean Squared Error (MSE):

-Squares the errors before averaging.

-Penalizes larger errors more heavily.

Root Mean Squared Error (RMSE):

-Converts MSE back to the same units as the target variable.

-Gives a sense of the "typical" error magnitude.

R² Score (Coefficient of Determination):

-Measures the proportion of variance in the dependent variable explained by the model.

-Ranges from 0 to 1 (closer to 1 is better).

1. Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

2. Mean Squared Error (MSE):

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

3. Root Mean Squared Error (RMSE):

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

4. R² Score (Coefficient of Determination):

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