Data Science

Unit 3-01: Machine Learning – Linear Regression

AKADEMI GA

GENERAL ASSEMBLY

COURSE CONTENT

Week 1 : Data Science Foundations

Installation and Github, Python fundamentals, Introduction to Pandas

Congratulations! 👏

Week 2 : Working with Data

Congratulations! 👏

More pandas, basics of probability and statistics, Exploratory Data Analysis (EDA), working with data, use statistical analysis and visualisation

Week 3 : Data Science Modeling

Linear regression Train/Test/Split, Classification, Logistic Regression

Week 4 : Data Science Applications

Using APIs, Natural Language Processing, Time Series Analysis

Week 5: Final Presentation

Present your capstone project

Review of Week1: Data Science Foundations

Previously, we have covered:

- a review of Python fundamentals

- Introduction to Pandas

- Data Joining, cleaning, manipulation with Pandas

Week 1 Units
1-01 Installation and Github
1-02 Python Review and Practice
1-03 List Comprehension
1-04 Introduction to Pandas
1-05 Data Wrangling

Review of Week2: Data Visualization

Previously, we have covered:

- Data visualization using seaborn, matplotlib
- Data transformation
- Hypothesis testing using confidence interval

Week 1 Units
2-01 Data Visualization
2-02 Data Transformation
2-03 Probability Distributions
2-04 Confidence Interval
2-05 Hypothesis Testing

Week 3: Data Science Modeling

- *In Unit 2, we will use machine learning Python modules.*
- *We will review the theory of machine learning and hands-on practice for classification regression modelling.*

Week 3 Units
3-01 Linear Regression
3-02 Regression Evaluation
3-03 Intro to classification
3-04 Logistic Regression
3-05 Grid searching & Decision Tree

Schedule

Time	Topics
5:00 - 6:30	Lesson 1: Introduction to machine learning
6:30 - 6:45	Break
6:45 - 7:45	Lesson 2: Linear Regression
7:45 - 8:00	Wrap up and Q&A

Our Learning Goals

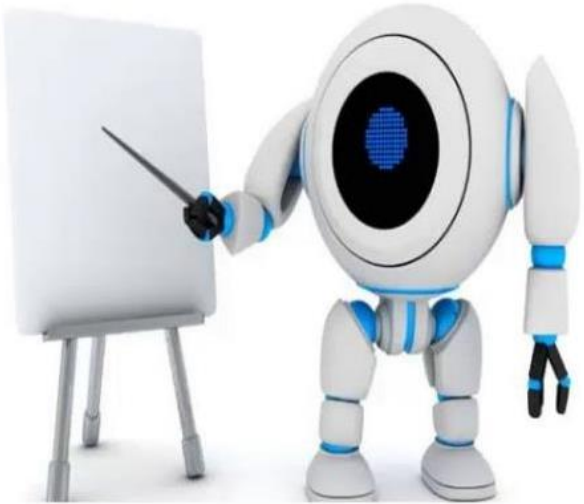
In this lesson, we will learn how to:

- Understanding Linear Regression Fundamentals
- Differentiating Simple and Multiple Linear Regression
- Evaluating Linear Regression Models and Handling Overfitting

Machine Learning

Learn From Experience

Learn From **Data**



Follow Instructions

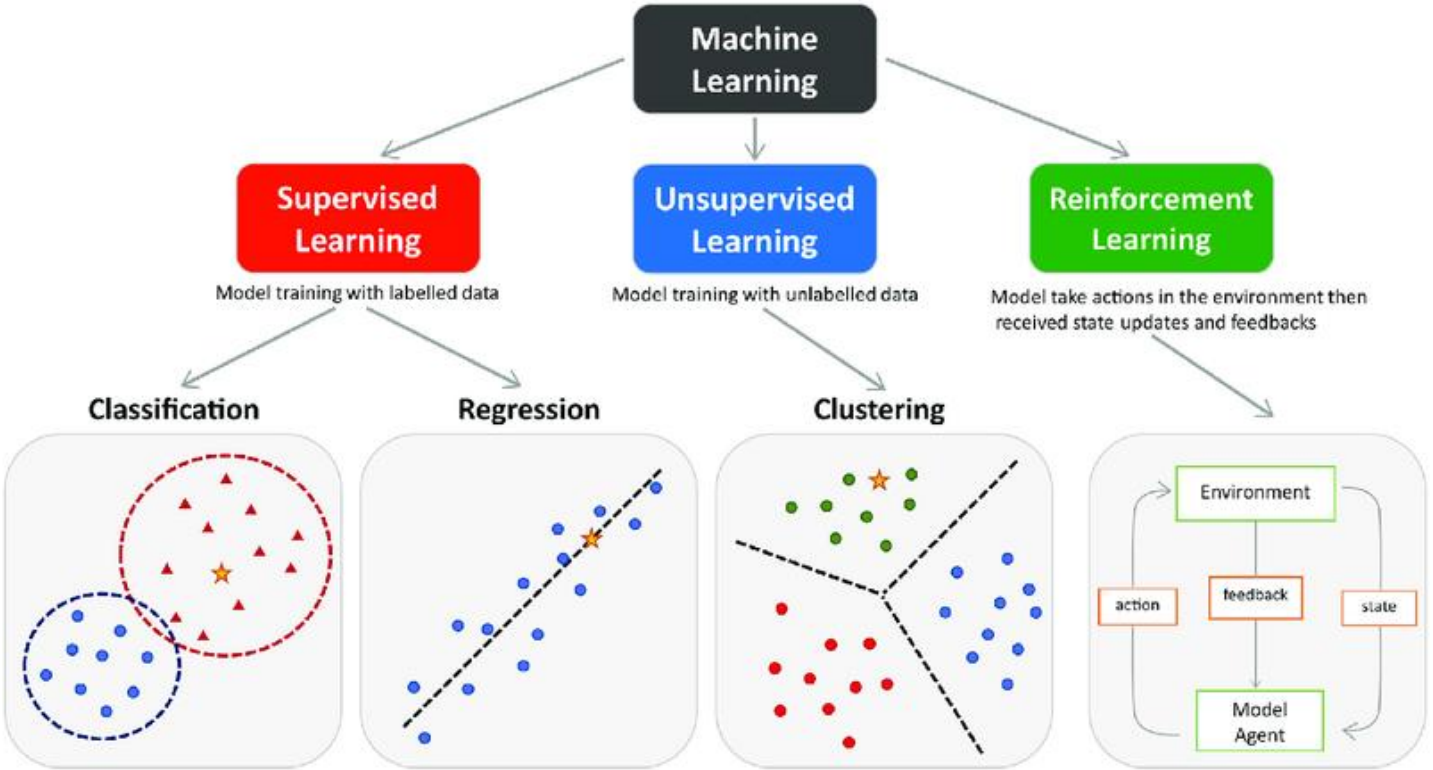
Machine Learning

Definition:
Machine learning is a subset of artificial intelligence that involves the development of algorithms and models that enable computers to learn patterns and make predictions from data without being explicitly programmed.

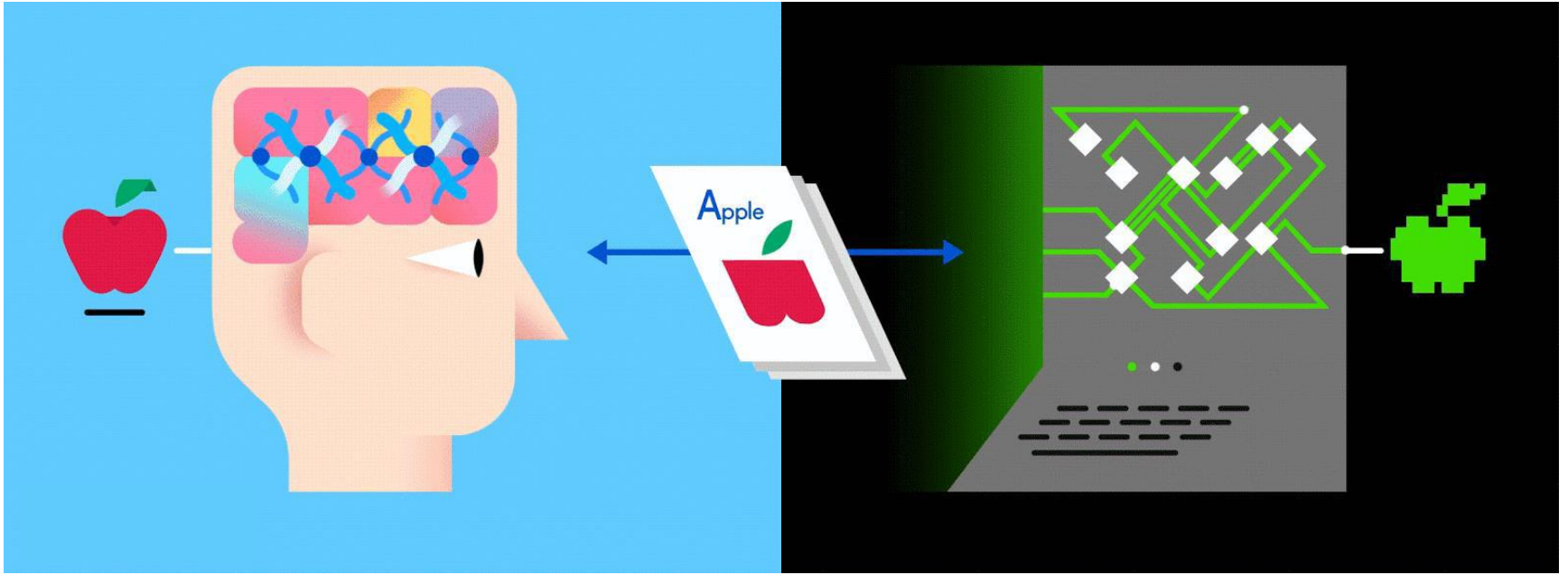
Key Components:

- Data: Input information used to train and test the models.
- Algorithms: Set rules and statistical techniques used to identify patterns and make predictions.
- Models: Trained algorithms that can generalize and make predictions on new, unseen data.

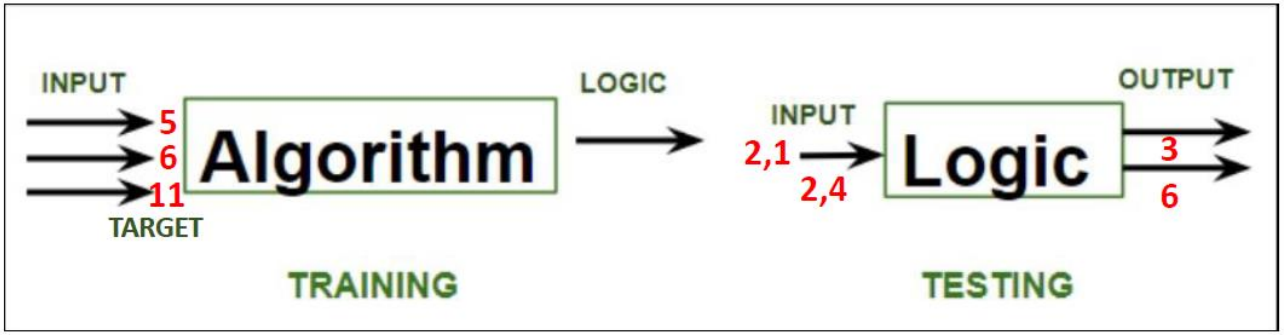
Types of Machine Learning

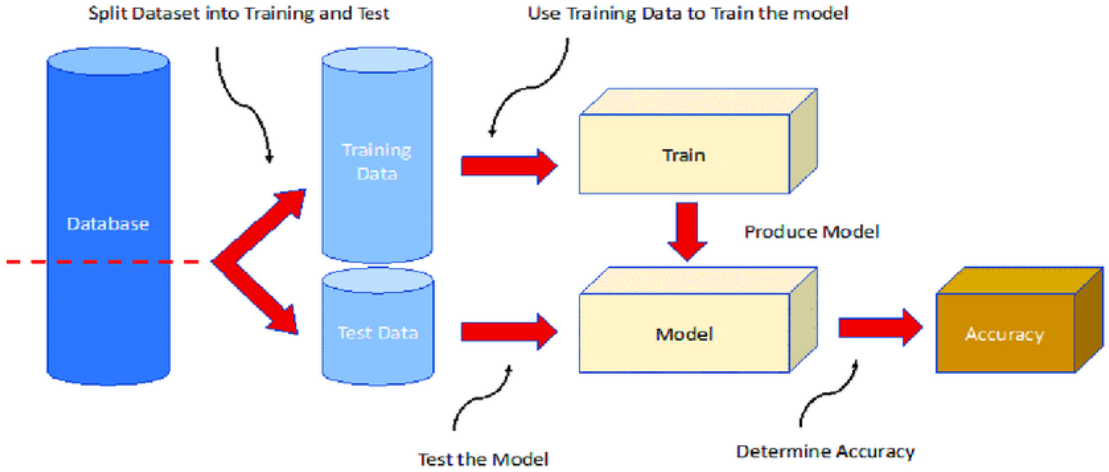


Supervised Learning

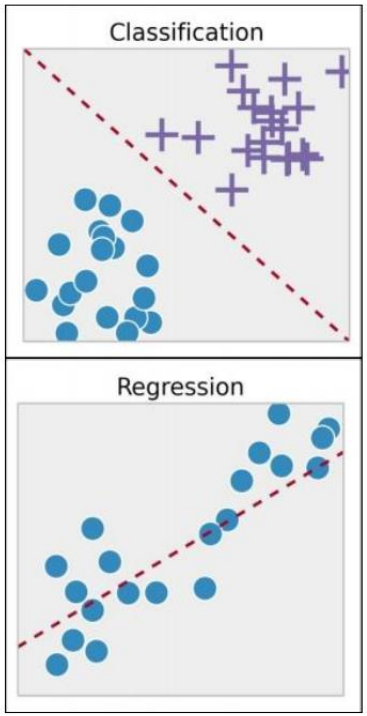


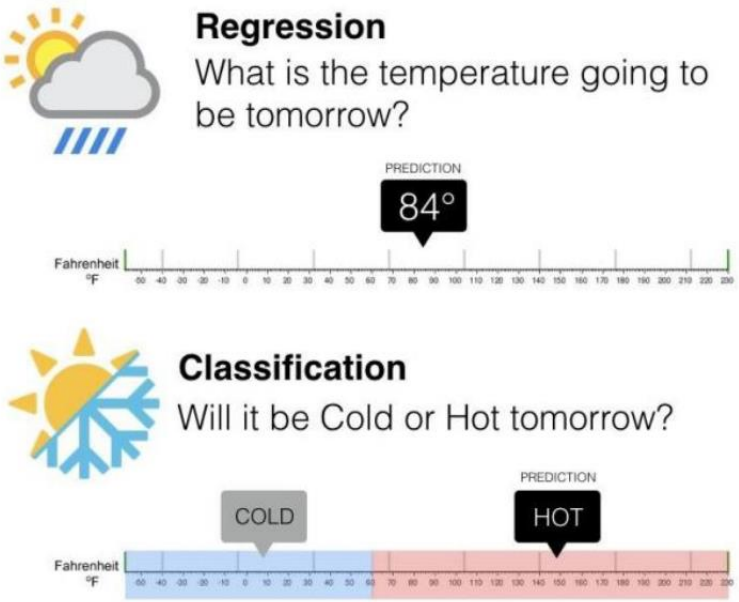
Supervised Learning



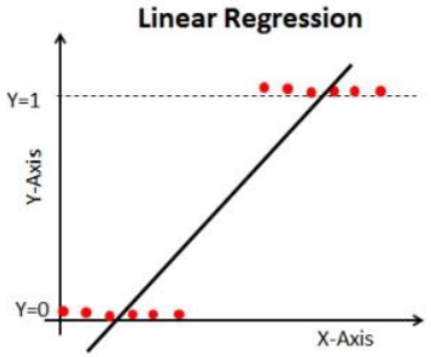


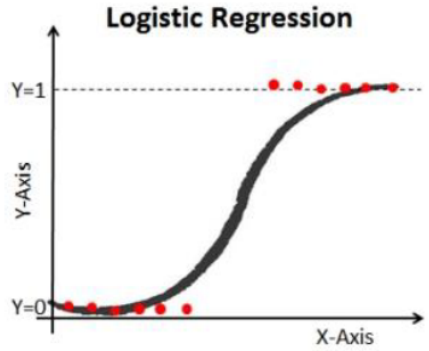
Supervised Learning





Supervised Learning - Regression





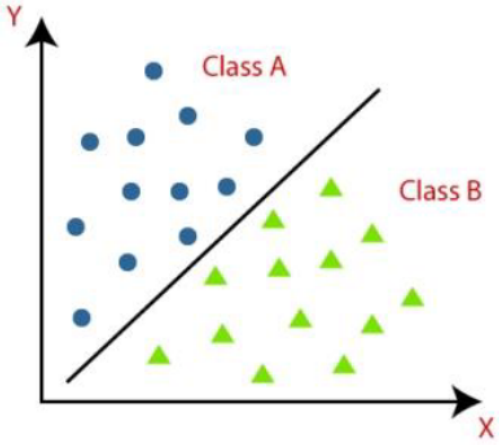
- In regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function.

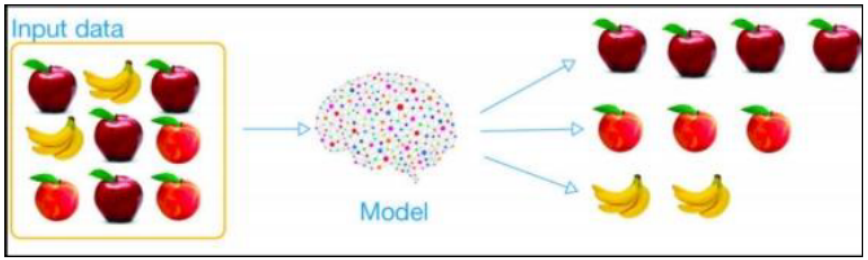
- Example 1 : given data about the size of houses on the real estate market, try to predict their price.

- Example 2: given a picture of a person, we have to predict their age or gender.

Linear Regression	Logistic Regression
Target is an interval variable.	Target is a discrete (binary or ordinal) variable.

Supervised Learning - Classification



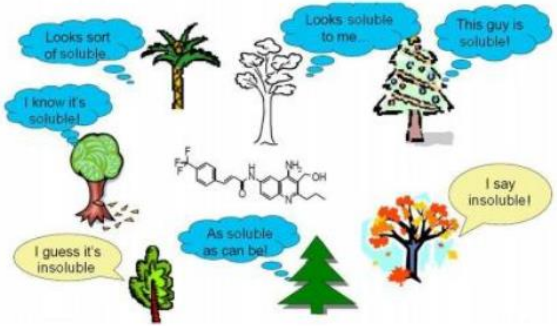


- Finding the category of the input variable, or in more academic terms, mapping input variables into discrete categories.
- like , yes or no, 0 or 1, true or false.
- Example 1: from the example of house price given above, if we change the output to “Sells for more or less than asking price,” then it is a classification problem. (Binary classification)
- Example 2: given a patient with tumour , we have to predict whether the tumour is malignant or benign . (Binary classification)
- Example 3: is this patient in cancer stage 1, 2, 3 or 4? (multi class classification)

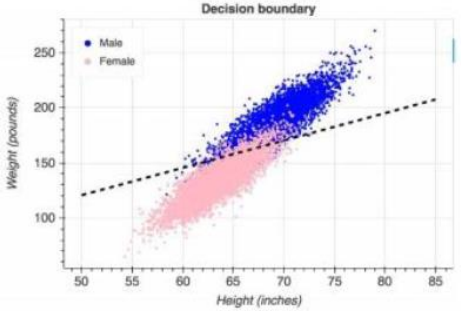
Supervised Learning Algorithms

Decision tree

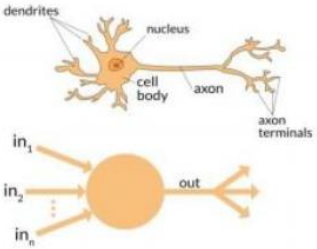
Random forest



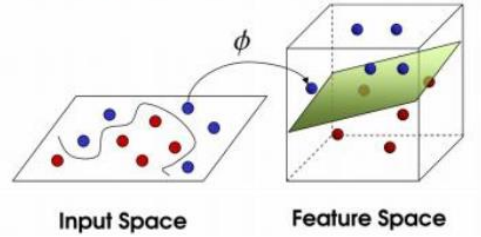
Logistic regression



Artificial neural networks



Support vector machine



Q&A

Schedule

Time	Topics
5:00 - 6:30	Lesson 1: Introduction to machine learning
6:30 - 6:45	Break
6:45 - 7:45	Lesson 2: Linear Regression
7:45 - 8:00	Wrap up and Q&A

Key concepts - Linear Regression

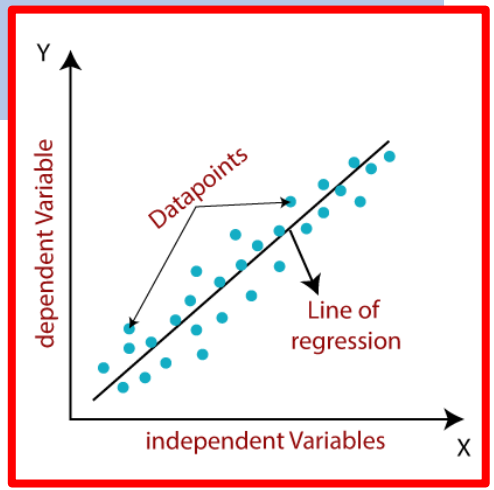
Definition:
Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables.

Purpose:
Predict the dependent variable based on the values of the independent variable(s) using a linear equation.

Dependent Variable (Y):
The variable we want to predict.

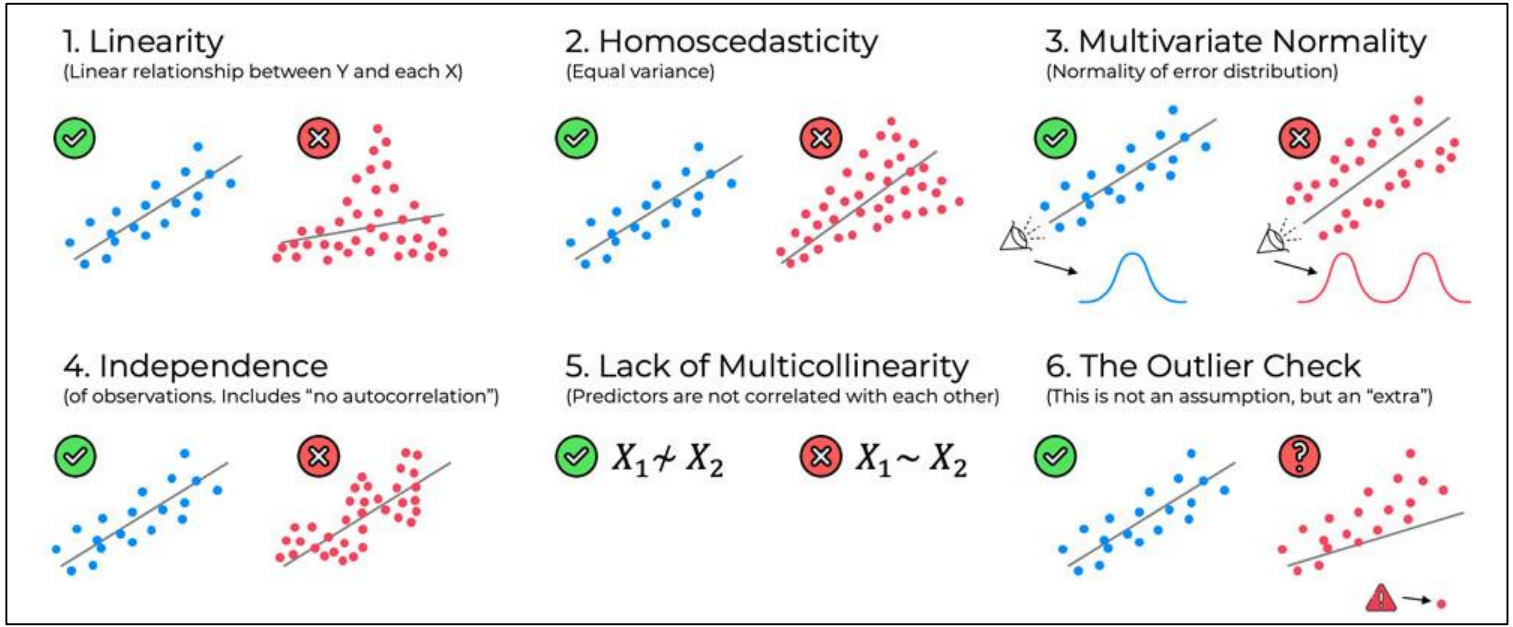
Independent Variable(s) (X):
The variable(s) used to predict the dependent variable.

Linear Equation:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$



Assumptions of Linear Regression

- ✓ ***Linearity:*** The relationship between the variables is linear.
- ✓ ***Independence:*** The residuals (ε) are independent of each other.
- ✓ ***Homoscedasticity:*** The variance of the residuals is constant.
- ✓ ***Normality:*** The residuals follow a normal distribution.



Simple Linear Regression

The task of simple linear regression is to exactly determine the straight line which best describes the linear relationship between the dependent and independent variable. In linear regression analysis, a straight line is drawn in the scatter plot. To determine this straight line, linear regression uses the method of least squares.

The regression line can be described by the following equation:

$$\hat{y} = b \cdot x + a$$

Estimated dependent variable

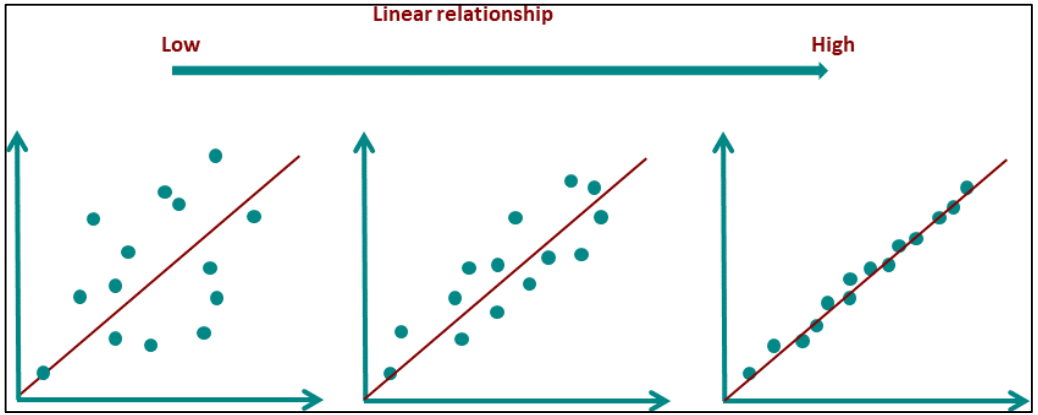
Slope

Independent variable

y intercept

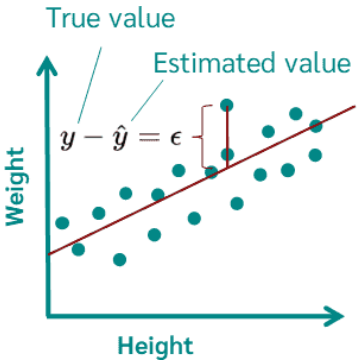
- ***a*** : point of intersection with the *y*-axis
- ***b*** : gradient of the straight line

$\hat{y}$ is the respective estimate of the *y*-value. This means that for each *x*-value the corresponding *y*-value is estimated. In our example, this means that the height of people is used to estimate their weight.



Simple Linear Regression

If all points (measured values) were exactly on one straight line, the estimate would be perfect. However, this is almost never the case and therefore, in most cases a straight line must be found, which is as close as possible to the individual data points. The attempt is thus made to keep the error in the estimation as small as possible so that the distance between the estimated value and the true value is as small as possible. This distance or error is called the "residual", is abbreviated as "e" (error) and can be represented by the greek letter epsilon (ϵ).
When calculating the regression line, an attempt is made to determine the regression coefficients (a and b) so that the sum of the squared residuals is minimal. (OLS- "Ordinary Least Squares").



The regression coefficient b can now have different signs, which can be interpreted as follows:

Error epsilon

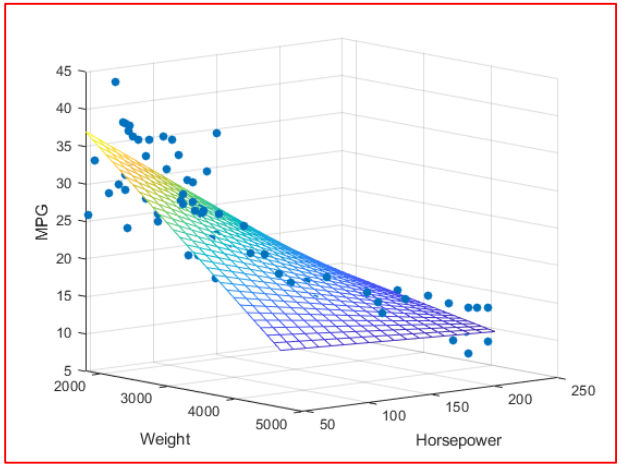
$$y = b \cdot x + a + \epsilon$$

b > 0: there is a positive correlation between x and y (the greater x, the greater y)
b < 0: there is a negative correlation between x and y (the greater x, the smaller y)
b = 0: there is no correlation between x and y

Multiple Linear Regression

Simple Linear Regression	**Multiple** Linear Regression
$\hat{y} = b \cdot x + a$	$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$

The coefficients can now be interpreted similarly to the linear regression equation. If all independent variables are 0, the resulting value is a. If an independent variable changes by one unit, the associated coefficient indicates by how much the dependent variable changes. So if the independent variable xi increases by one unit, the dependent variable y increases by bi.



Multiple Linear Regression

Marketing example:
For a video streaming service you should predict how many times a month a person streams videos. For this you get a record of user's data (age, income, gender, ...).

Medical example:
You want to find out which factors have an influence on the cholesterol level of patients. For this purpose, you analyze a patient data set with cholesterol level, age, hours of sport per week and so on.

Q&A