## **Supervised Learning:** (Labelled Data)

- 1. **Classification:** is the process of predicting discrete class labels or categories.
- 2. **Regression:** is the process of predicting continuous values.

## **Un-supervised Learning:** (Unlabelled Data)

- 1. **Clustering:** is a grouping of data points or objects that are somehow similar by:
  - Discovering structure
  - Summarization
  - Anomaly detection

Supervised Learning	Un-supervised Learning		
Classification: Classifies labelled data.	Clustering: Finds patters and groupings from unlabelled data.		
Regression: Predicts trends using previous labelled data.	Has fewer evaluation methods than supervised learning.		
Has more evaluation methods than supervised learning and has more controlled environment.	Less controlled environment.		

## **Introduction to Regression:**

**Regression** is the process of predicting a continuous value.

## **Types of Regression Models:**

- 1. Simple Regression: (1 independent variable)
  - a. Simple Linear Regression
  - b. Simple Non-linear Regression
- 2. Multiple Regression: (more than 1 independent variable)
  - a. Multiple Linear Regression
  - b. Multiple Non-linear Regression

## **Applications of regression:**

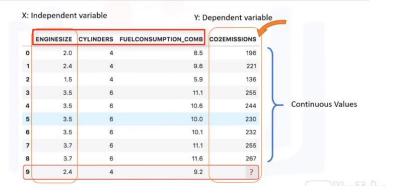
- 1. Sales forecasting
- 2. Satisfaction analysis
- 3. Price estimation
- 4. Employment income

## **Regression algorithms:**

- 1. Ordinal regression
- 2. Poisson regression
- 3. Fast forest quantile regression
- 4. Linear, Polynomial, Lasso, Stepwise, Ridge regression
- 5. Bayesian linear regression
- 6. Neural network regression
- 7. Decision forest regression
- 8. Boosted decision tree regression
- 9. KNN (K-nearest neighbours)

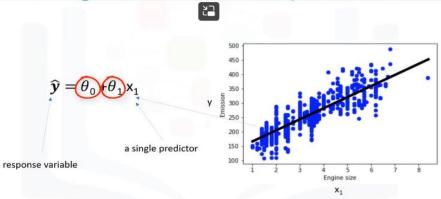
## **Simple Linear Regression:**

# Using linear regression to predict continuous values

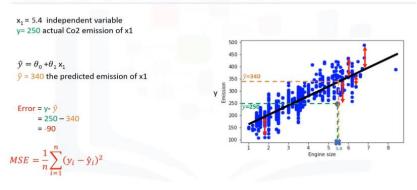


The intercept and the slope.

# Linear regression model representation



# How to find the best fit?



# Estimating the parameters

ENG	INESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	(2.0	4	8.5	( 196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4 X <sub>1</sub> -	3.5	6	10.6	y 244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
В	3.7	6	11.6	267

$$\widehat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{S} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{S} (x_i - \overline{x})^2}$$

$$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.03$$

$$\bar{y} = (196 + 221 + 136 + \dots)/9 = 226.22$$

$$\theta_1 = \frac{(2.0 - 3.03)(196 - 226.22) + (2.4 - 3.03)(221 - 226.22) + \dots}{(2.0 - 3.03)^2 + (2.4 - 3.03)^2 + \dots}$$

$$\theta_1 = 39$$

$$\theta_0 = \overline{y} - \theta_1 \overline{x}$$

$$\theta_0 = 226.22 - 39 * 3.03$$

$$\theta_0 = 125.74$$

$$\widehat{y} = 125.74 + 39x_1$$

# Predictions with linear regression

<b>^ 0 .0</b>	D2EMISSIONS	UELCONSUMPTION_COMB	CYLINDERS	ENGINESIZE	
$\hat{y} = \theta_0 + \theta_1 x_1$	196	8.5	4	2.0	0
	221	9.6	4	2.4	1
$Co2Emission = \theta_0 + \theta_1 EngineSize$	136	5.9	4	1.5	2
C-DP-1-1-1-1-1-105 - 20 P1 01-	255	11.1	6	3.5	3
Co2Emission = 125 + 39 EngineSize	244	10.6	6	3.5	4
$Co2Emission = 125 + 39 \times 2.4$	230	10.0	6	3.5	5
Co2Emission = 218.6	232	10.1	6	3.5	6
COZEmission = 218.6	255	11.1	6	3.7	7
	267	11.6	6	3.7	8
	?	9.2	4	2,4	9

## Pros of linear regression:

- 1. Very fast
- 2. No parameter tuning
- 3. Easy to understand, and highly interpretable.

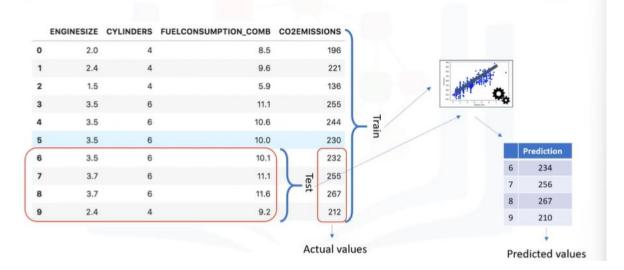
# **Model Evaluation in Regression Model:**

# Model evaluation approaches:

- 1. Train and test on the same dataset
- 2. Train/Test split.

# Train and test on the same dataset:

# Best approach for most accurate results?



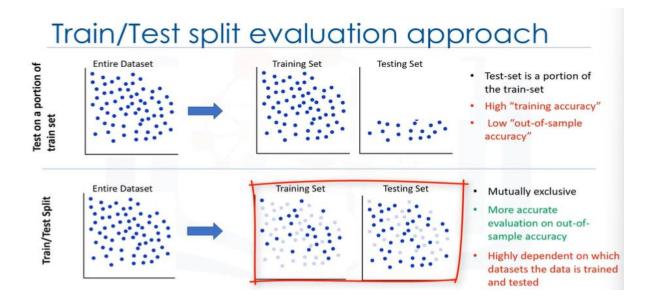


0	2.0	4	8.5	196		$Error = \frac{1}{2}$	,	4	55 – 256) -
1	2.4	4	9.6	221				4	
2	1.5	4	5.9	136		1 2			
3	3.5	6	11.1	255		$Error = \frac{1}{n} \sum_{n}$	$ y_j - \hat{y}_j $		
4	3.5	6	10.6	244		j=1		┛	
5	3.5	6	10.0	230					Table State of the
6	3.5	6	10.1	232					Prediction
7	3.7	6	11.1	च 255				6	234
В	3.7	6	11.6	Fig. 255	y		9	7	256
					8			8	267
9	2.4	4	9.2	212	J			9	210

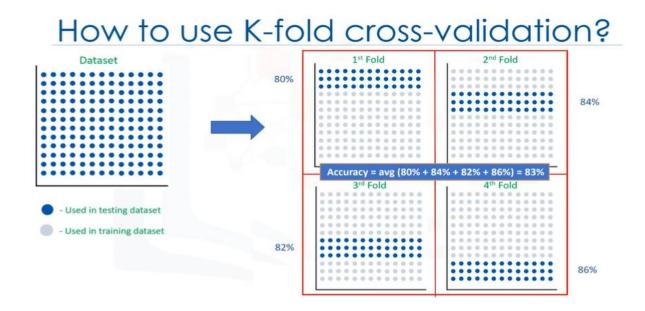
# What is training and out-of-sample accuracy?

- Training Accuracy.
  - High training accuracy isn't necessarily a good things.
  - Result of over-fitting
    - Over-fit: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model.
- Out-of-Sample Accuracy.
  - It's important that our models have a high, out-of-sample accuracy.
  - How can we improve out-of-sample accuracy?

# Train/Test Split evaluation approach:

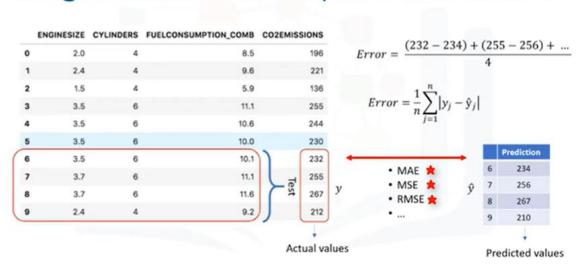


How to use K-fold cross-validation?



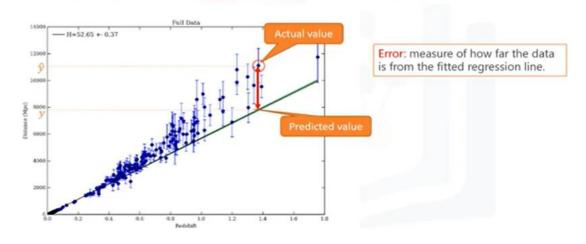
**Evaluation Metrics in Regression Models:** 

# Regression accuracy



## What is an ERROR?

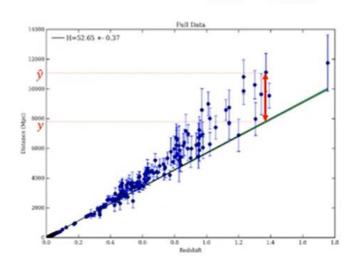
# What is an error of the model?



MAE: Mean Absolute Error
MSE: Mean Squared Error
RMSE: Root Mean Square Error
RAE: Relative Absolute Error
RSE: Relative Squared Error

Higher the **R(Square)** the better your predictions fit the data.

# What is an error of the model?



$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_{j} - \hat{y}_{j}|$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_{i} - \hat{y}_{i})^{2}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (y_{j} - \hat{y}_{j})^{2}}$$

$$RAE = \frac{\sum_{j=1}^{n} |y_{j} - \hat{y}_{j}|}{\sum_{j=1}^{n} |y_{j} - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n} (y_{j} - \hat{y}_{j})^{2}}{\sum_{j=1}^{n} (y_{j} - \bar{y})^{2}}$$

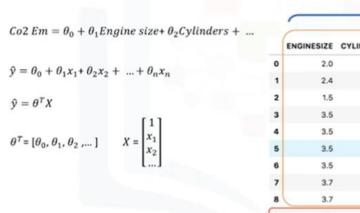
$$R^{2} = 1 - RSE$$

## **Multiple Linear Regression:**

## **Example of multiple linear regression:**

- Independent variables effectiveness on prediction
  - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?
- Predicting impacts of changes
  - How much does blood pressure go up (or down) for every unit increase (or decrease) in the BMI of a patient?

Predicting continuous values with multiple linear regression



	X: Independent variable Y: Dependent variab					
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	COZEMISSIONS		
)	2.0	4	8.5	196		
1	2.4	4	9.6	221		
2	1.5	4	5.9	136		
3	3.5	6	11.1	255		
1	3.5	6	10.6	244		
5	3.5	6	10.0	230		
ı	3.5	6	10.1	232		
,	3.7	6	11.1	255		
3	3.7	6	11.6	267		
9	2.4	4	9.2	3		

# Estimating multiple linear regression parameters

- How to estimate  $\theta$ ?
  - · Ordinary Least Squares
    - Linear algebra operations
    - Takes a long time for large datasets (10K+ rows)
  - · An optimization algorithm
    - Gradient Descent
    - Proper approach if you have a very large dataset

# Making predictions with multiple linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	COZEMISSIONS	$\hat{y} = \theta^T X$
0	2.0	4	8.5	196	$\theta^T = [125, 6.2, 14, \dots]$
1	2.4	4	9.6	221	$\hat{y} = 125 + 6.2x_1 + 14x_2 +$
2	1.5	4	5.9	136	y 120 · 0.2.1 · 1.1.2
3	3.5	6	11.1	255	Co2Em = 125 + 6.2EngSize + 14 Cylinders +
4	3.5	6	10.6	244	
5	3.5	6	10.0	230	$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 +$
6	3.5	6	10.1	232	
7	3.7	6	11.1	255	Co2Em = 214.1
8	3.7	6	11.6	267	C02EM = 214.1
9	2.4	4	9.2	?	

## Q&A - on multiple linear regression

- How to determine whether to use simple or multiple linear regression?
  - Requirements
- How many independent variables should you use?
  - Look for overfit modelling.
- Should the independent variable be continuous?
  - Should be meld.
- What are the linear relationships between the dependent variable and the independent variables?
  - This can be checked by plotting a graph and then visualizing whether there is any linearity in the data.

## **Non-linear Regression:**

# What is polynomial regression?

- Some curvy data can be modeled by a polynomial regression
- · For example:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

 A polynomial regression model can be transformed into linear regression model.

$$x_1 = x$$

$$x_2 = x^2$$

$$x_3 = x^3$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

Multiple linear regression

Least Squares

Minimizing the sum of the squares of the differences between y and  $\hat{y}$ 

# What is non-linear regression?

- To model non-linear relationship between the dependent variable and a set of independent variables
- $\hat{y}$  must be a non-linear function of the parameters  $\theta$ , not necessarily the features x

$$\hat{y} = \theta_0 + \theta_2^2 x$$

$$\hat{y} = \theta_0 + \theta_1 \theta_2^x$$

$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$

$$\hat{y} = \frac{\theta_0}{1 + \theta_2^{(x - \theta_2)}}$$



#### Linear vs non-linear regression

- How can I know if a problem is linear or non-linear in an easy way?
  - o Inspect visually
  - Based on accuracy
- How should I model my data, if it displays non-linear on a scatter plot?
  - o Polynomial regression
  - Non-linear regression model
  - Transform your data