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

X: Independent variable

Y: Dependent variable

	ENGINESIZE	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	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values

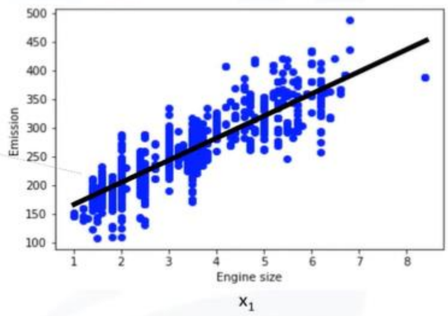
The intercept and the slope.

Linear regression model representation

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

response variable

a single predictor



How to find the best fit?

$x_1 = 5.4$ independent variable

$y = 250$ actual Co2 emission of x1

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

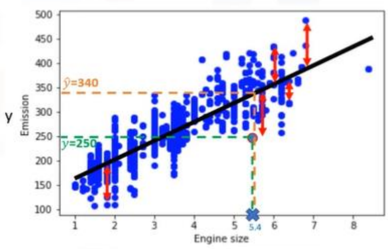
$\hat{y} = 340$ the predicted emission of x1

Error $= y - \hat{y}$

$= 250 - 340$

$= -90$

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



Estimating the parameters

	ENGINESIZE	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	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 = ENGINESIZE, y = CO2EMISSIONS

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

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

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

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

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

$\theta_1 = 39$

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

$\theta_0 = 226.22 - 39 * 3.03$

$\theta_0 = 125.74$

$$\boxed{\hat{y} = 125.74 + 39 x_1}$$

Predictions with linear regression

	ENGINESIZE	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	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

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

$Co2Emission = \theta_0 + \theta_1 \, EngineSize$

$Co2Emission = 125 + 39 \, EngineSize$

$Co2Emission = 125 + 39 \times 2.4$

$Co2Emission = 218.6$

Pros of linear regression:

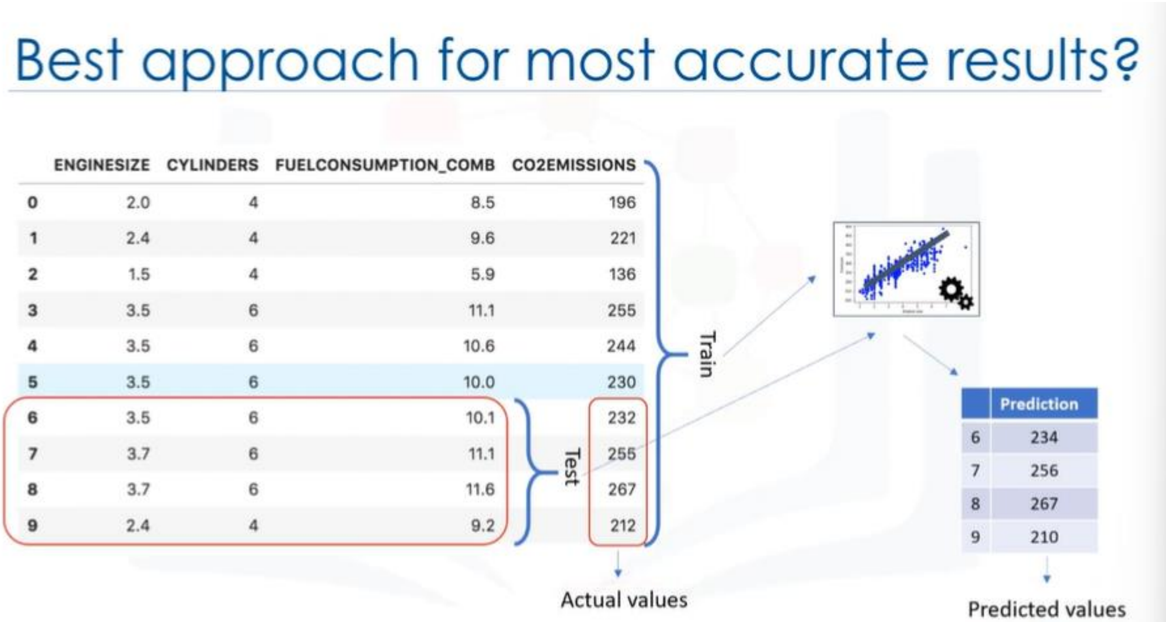
1. Very fast
2. No parameter tuning
3. Easy to understand, and highly interpretable.

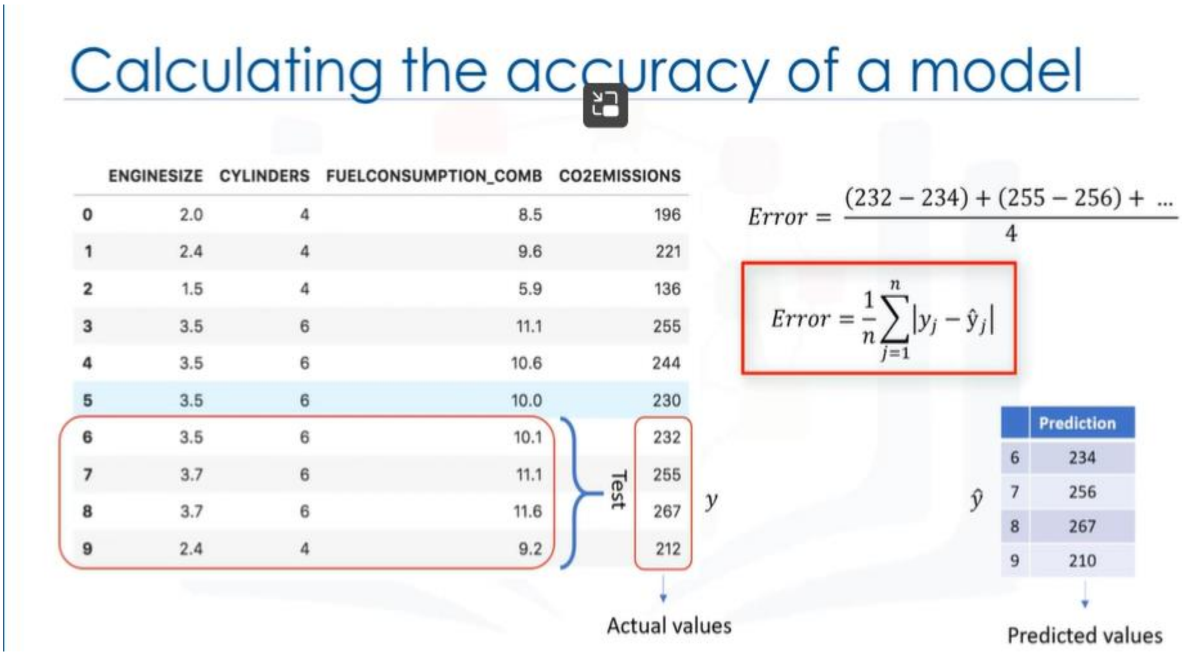
Model Evaluation in Regression Model:

Model evaluation approaches:

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

Train and test on the same dataset:

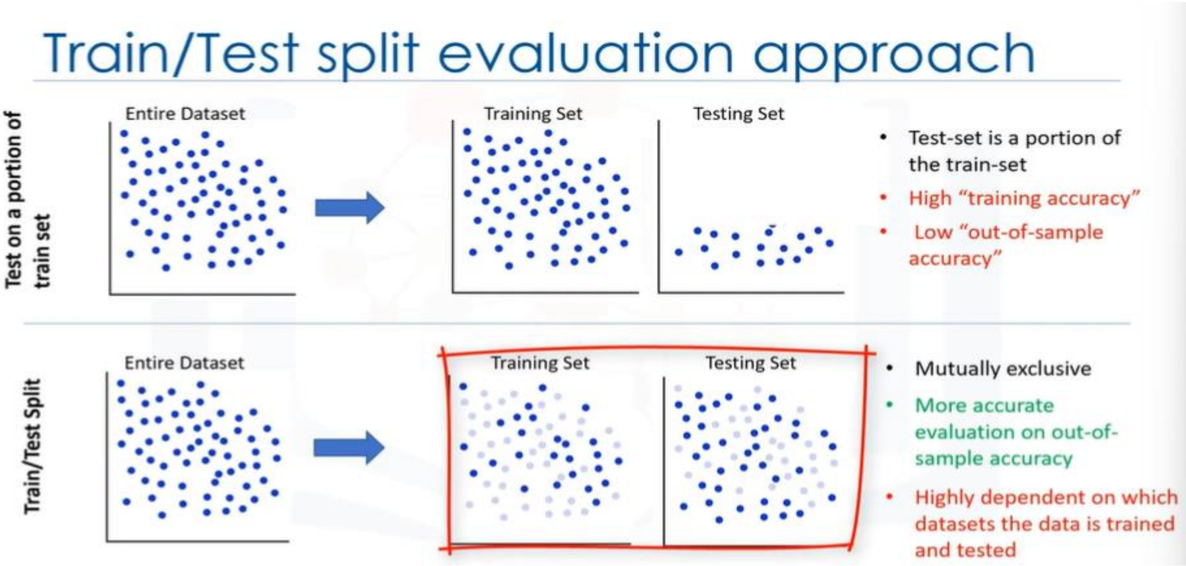




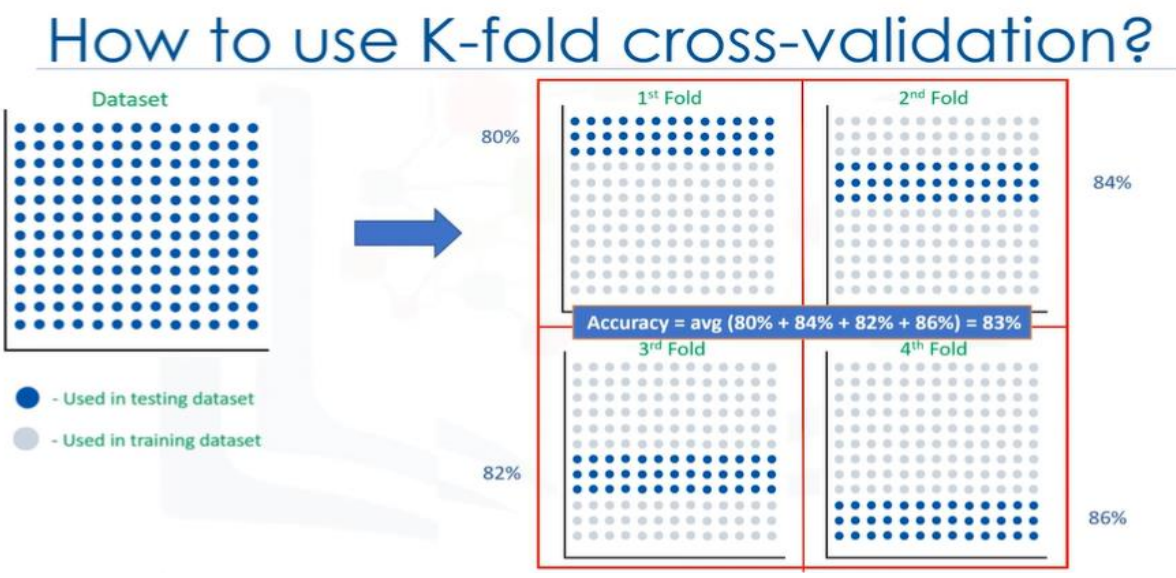
What is training and out-of-sample accuracy?

- **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?

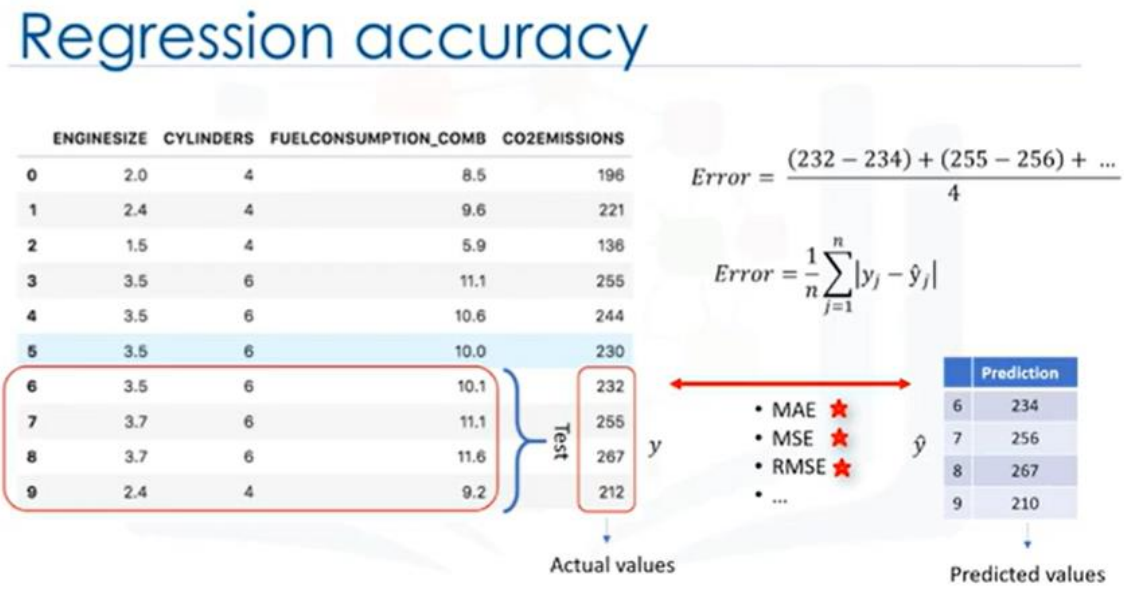
Train/Test Split evaluation approach:



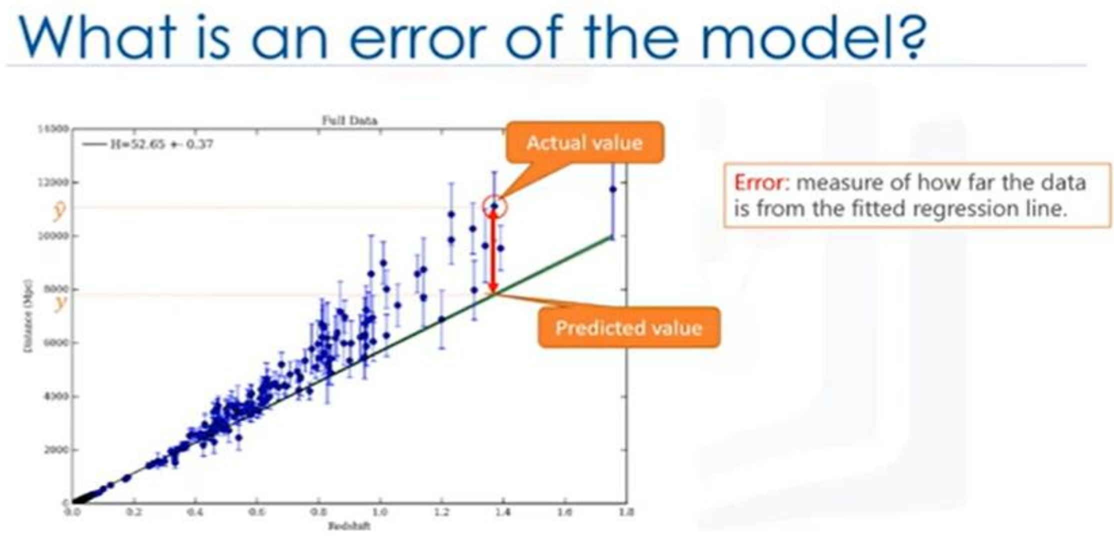
How to use K-fold cross-validation?



Evaluation Metrics in Regression Models:



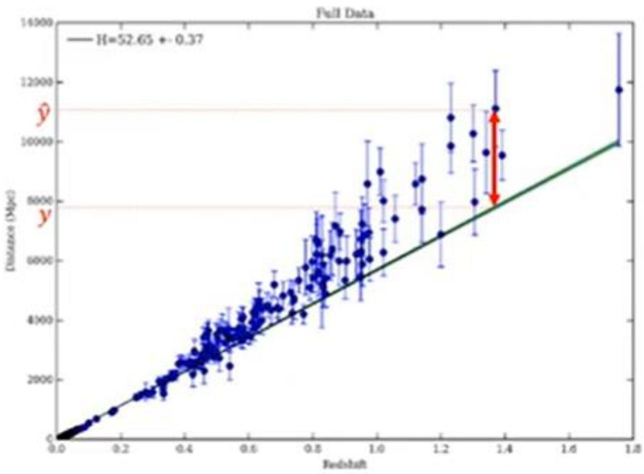
What is an ERROR?



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?



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

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

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

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

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

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

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

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

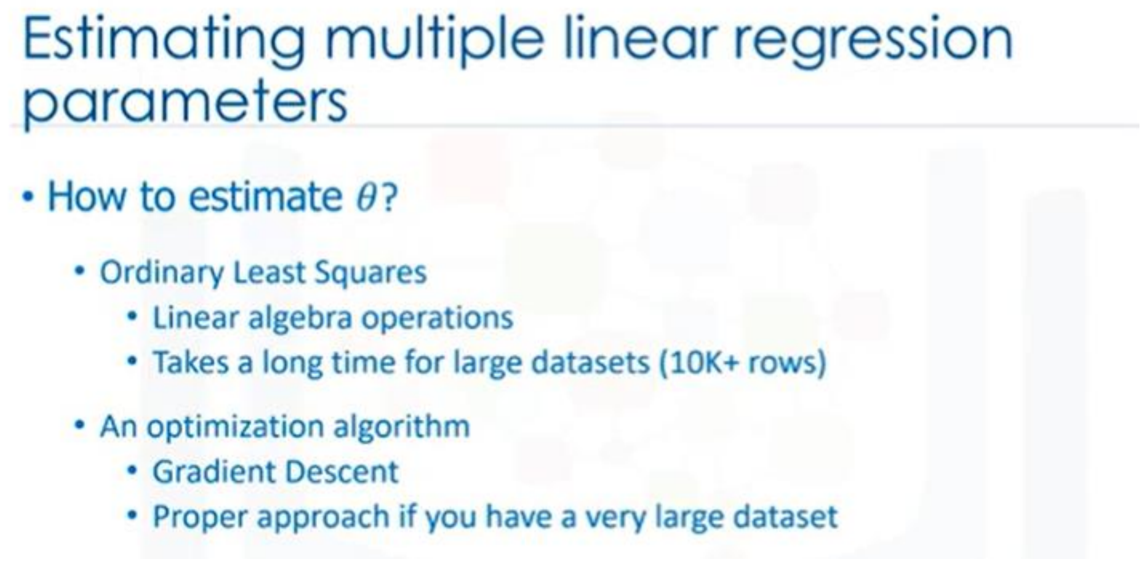
$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

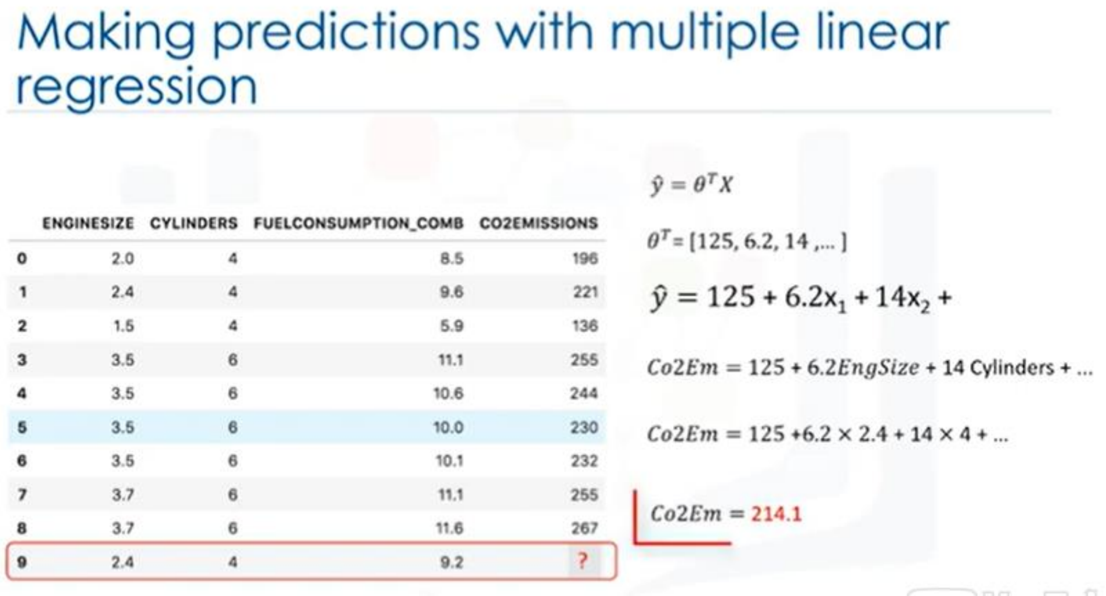
$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

X: Independent variable — Y: Dependent variable

	ENGINESIZE	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	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

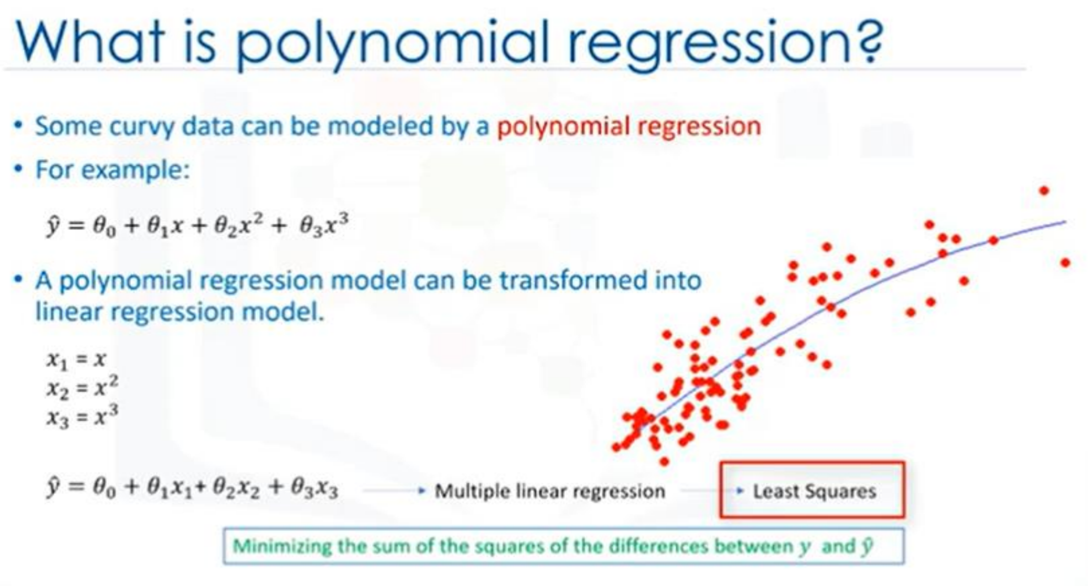


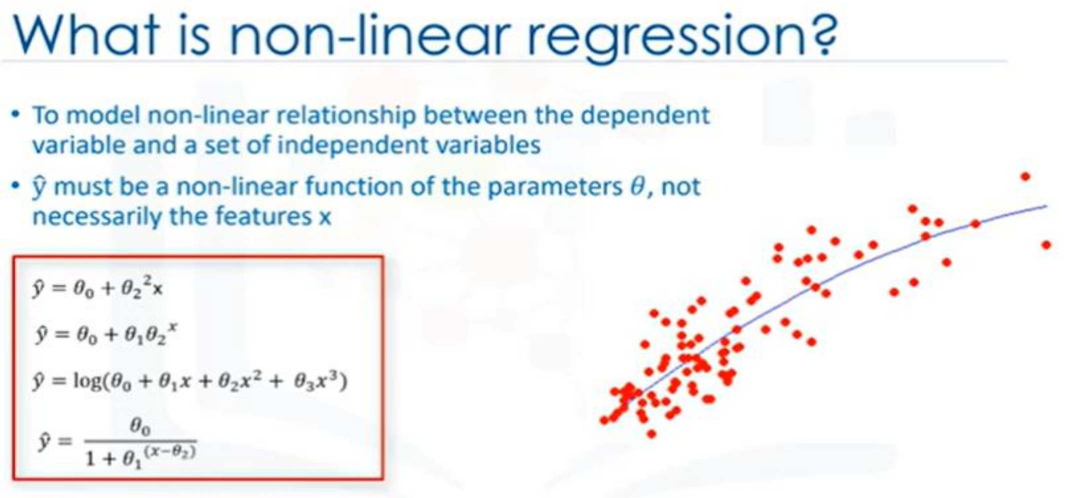


Q&A - on multiple linear regression

- 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:





Linear vs non-linear regression

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