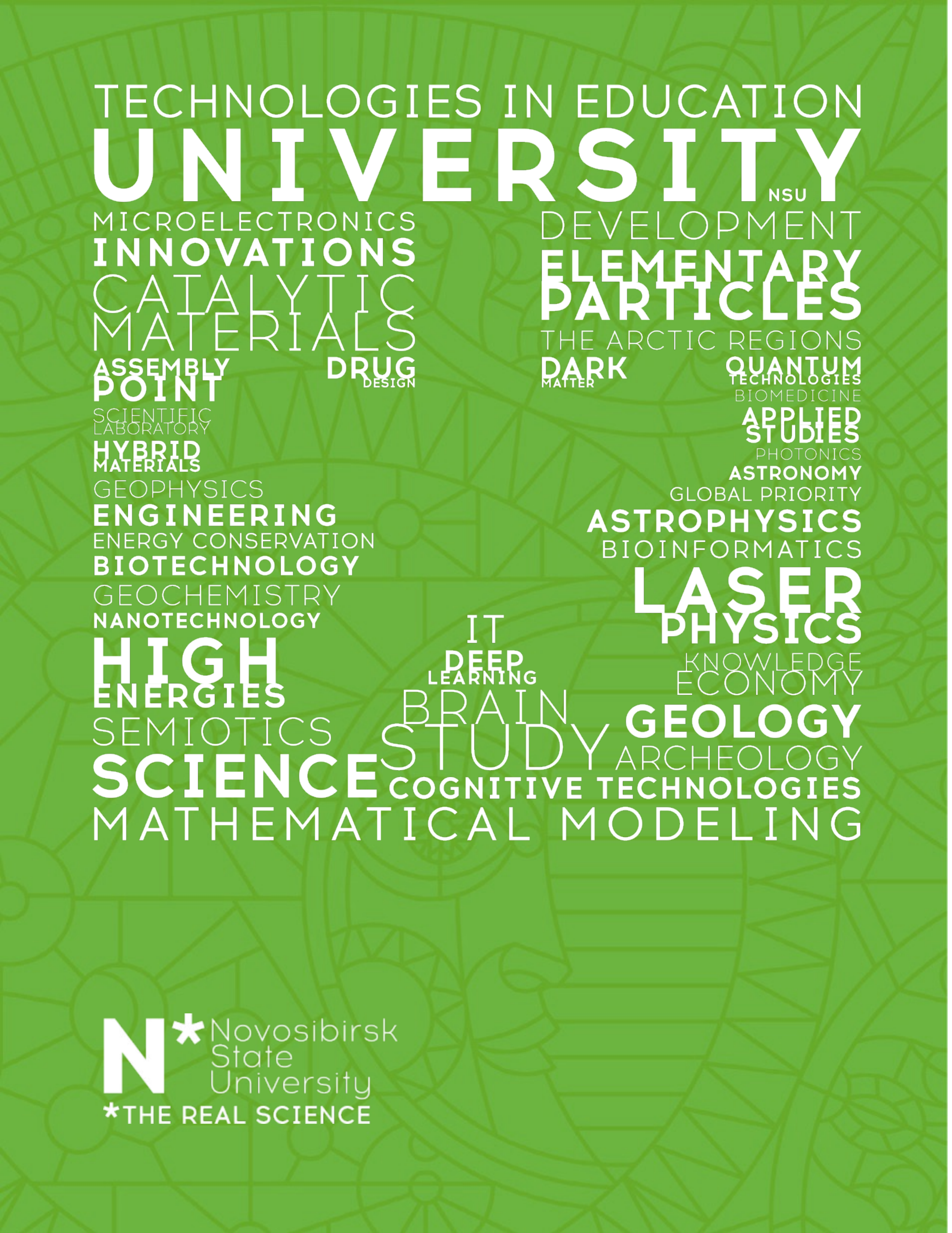


Regression

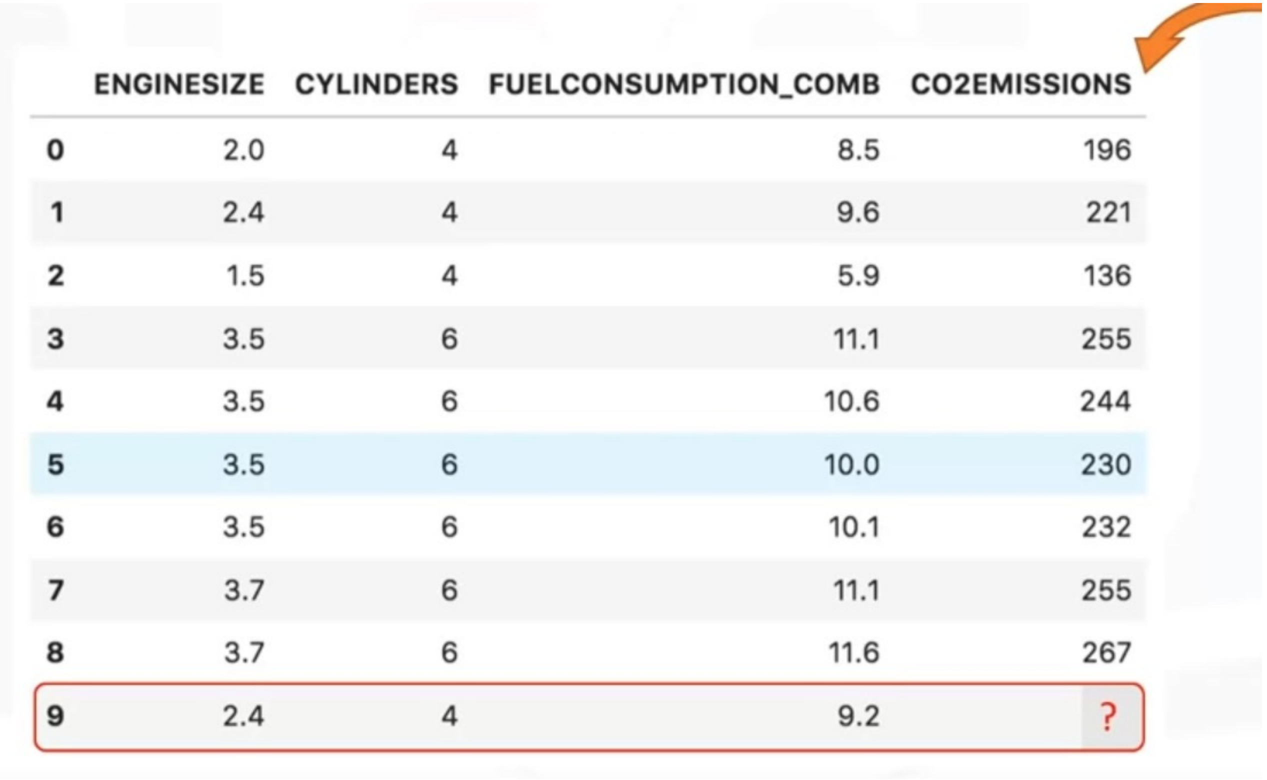
Luu Minh Sao Khue

Department of Mathematics and Mechanics

Overview

- Introduction to Regression
- Simple Linear Regression
- Practical Implementation
- Model Evaluation and Diagnostics
- Advanced Topics, Applications, and Q&A

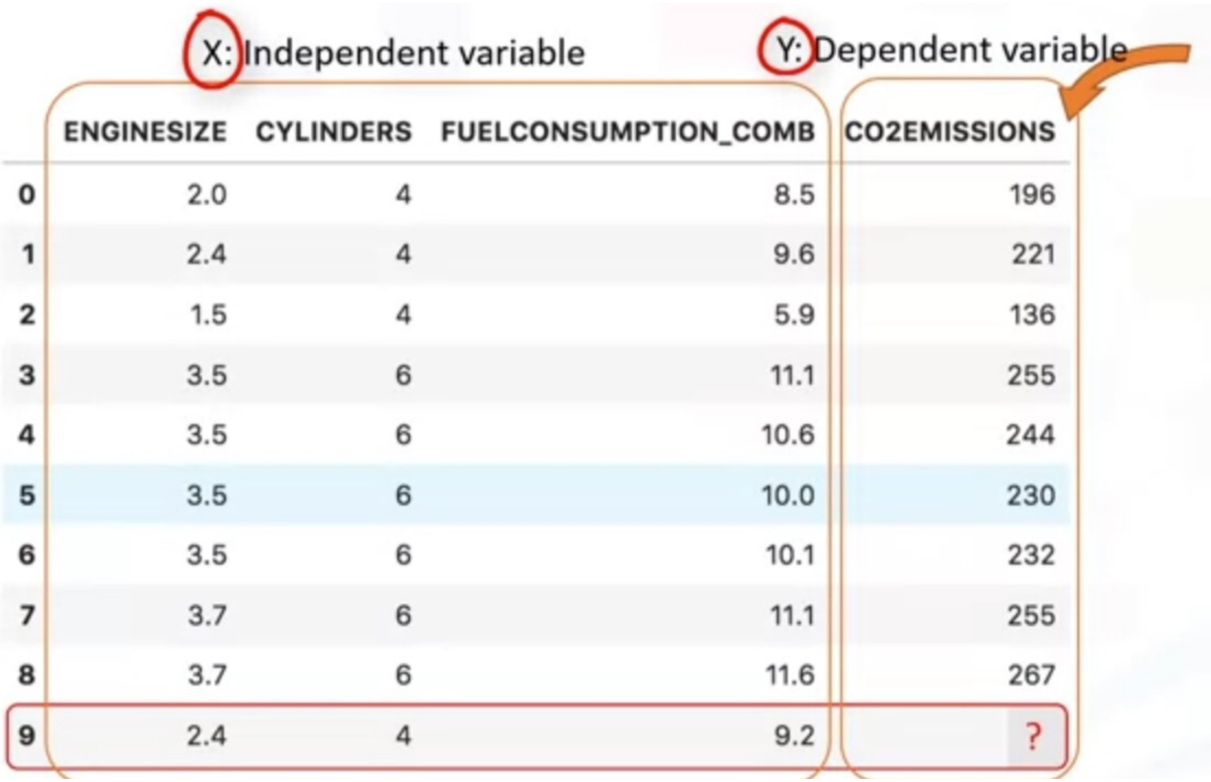
Regression



	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	?

- Regression is a process of predicting a continuous value

Regression



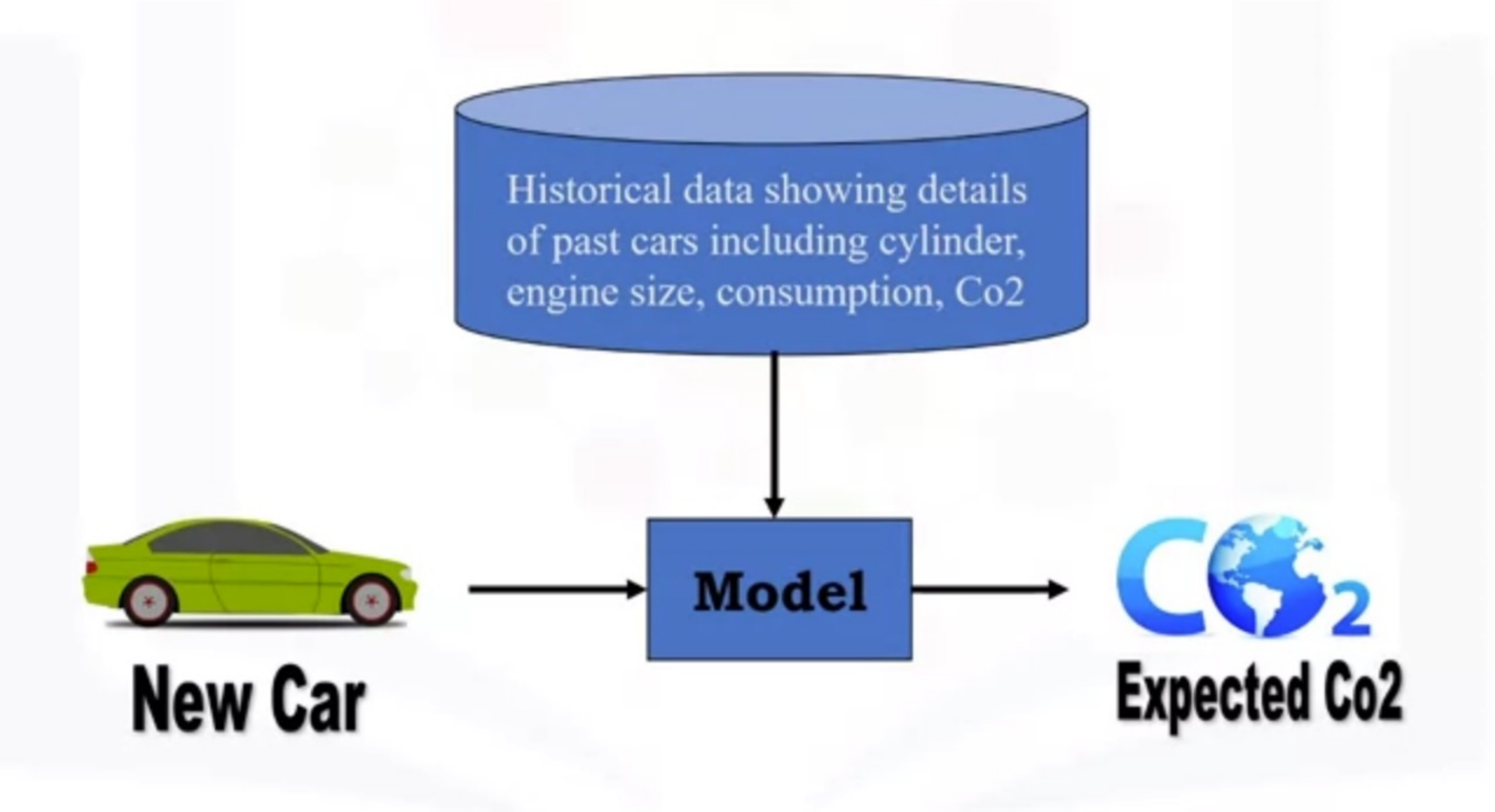
X: Independent variable

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	?

- Dependent variable: the value we want to predict
- Independent variables: the values that explain or cause the value of the dependent variable

Regression



Historical data showing details of past cars including cylinder, engine size, consumption, Co2

Types of regression

- Simple Regression
  - Using one independent variable to predict the dependent variable
  - Ex: predict CO2 emssion using engine size.
  - Simple Linear Regression, Simple Non-Linear Regresison
- Multiple Regression
  - Using more than one independent variables to predict the dependent variable
  - Ex: predict CO2 emssion using engine size and number of cylinders.
  - Mulitple Linear Regression, Multiple Non-Linear Regresison

Applications of regression

- Sales forcasting
  - Predict a yearly sale of a person based on Age, Years of Experience, etc.
- Satisfaction analysis
  - Detemine individual satisfaction based on demographic and psychological factors.
- Price estimation
  - Predict a price of a house based on its size, number of rooms, etc.
- Employment income
  - to predict employment income for independent variables such as hours of work, education, occupation, sex, age, years of experience

Quiz

- Which one is a sample application of regression?
  - Predicting whether a patient has cancer or not.
  - Grouping of similar houses in an area.
  - Forecasting rainfall amount for next day.
  - Predicting if a team will win or not.

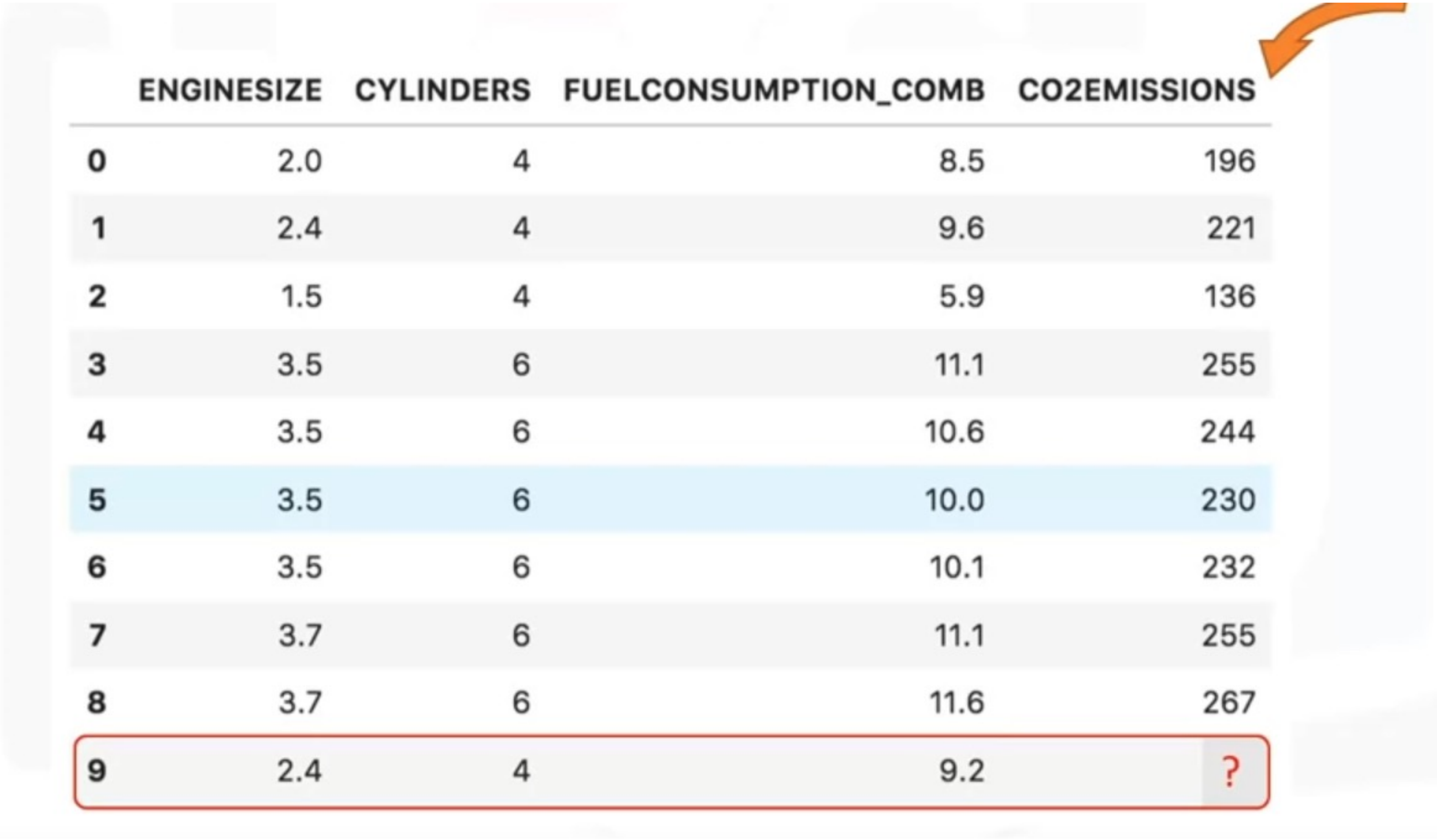
Some regresison algorithms

- Ordinal Regression
- Poisson Regression
- **Linear Regression**
- Polinomial Regression
- Lasso Regression
- Ridge Regression
- Decision Forest Regression
- Boosted Decision Tree Regression

Simple Linear Regresison

Simple Linear Regresison

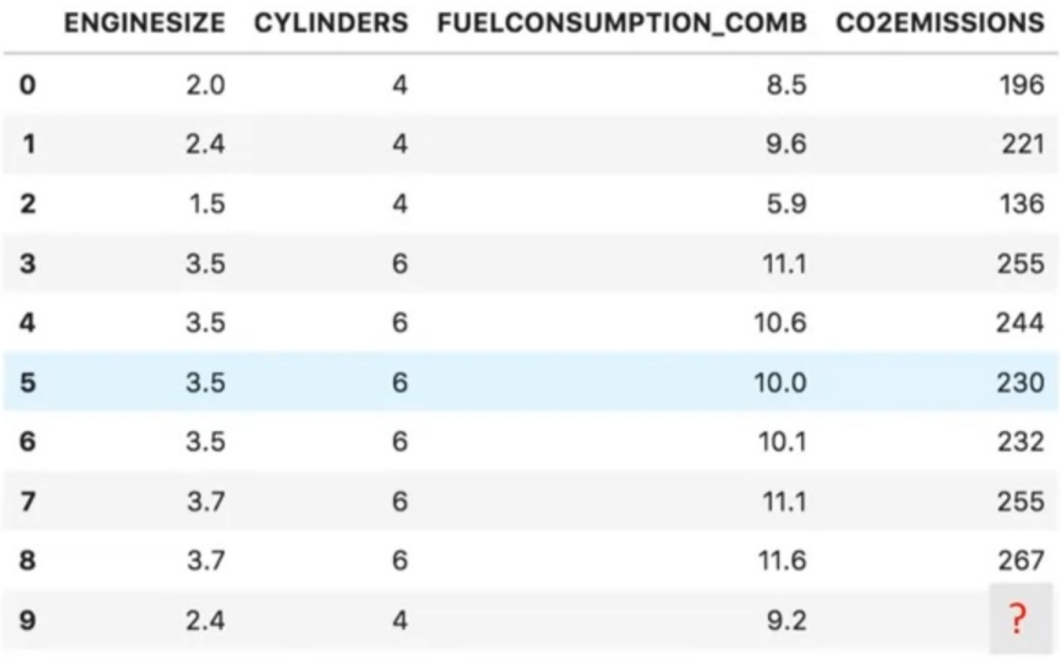
- Linear regression is the approximation of a linear model used to describe the relationship between two or more variables



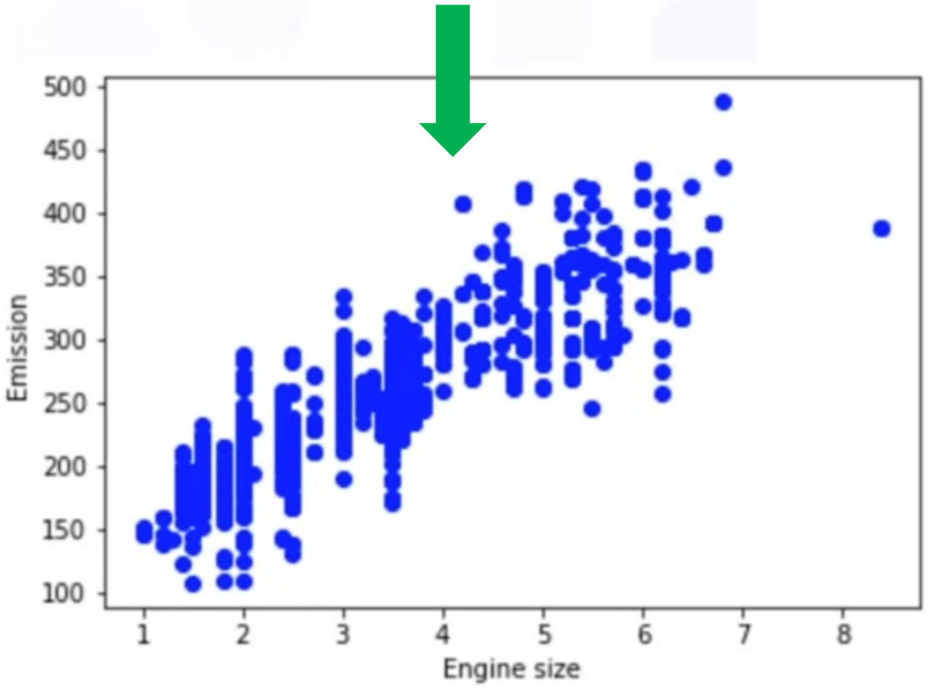
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Simple Linear Regression

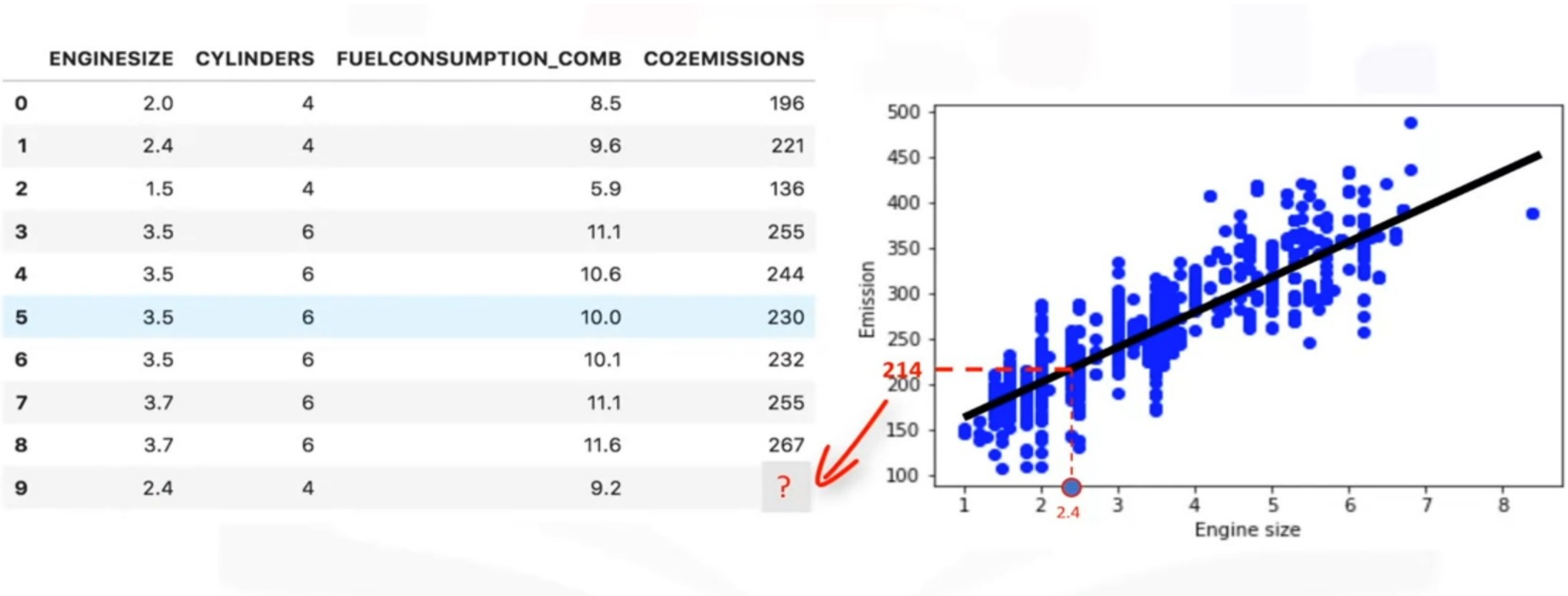
Changes in 1 variable explain changes in the other



	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	?



Simple 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	?

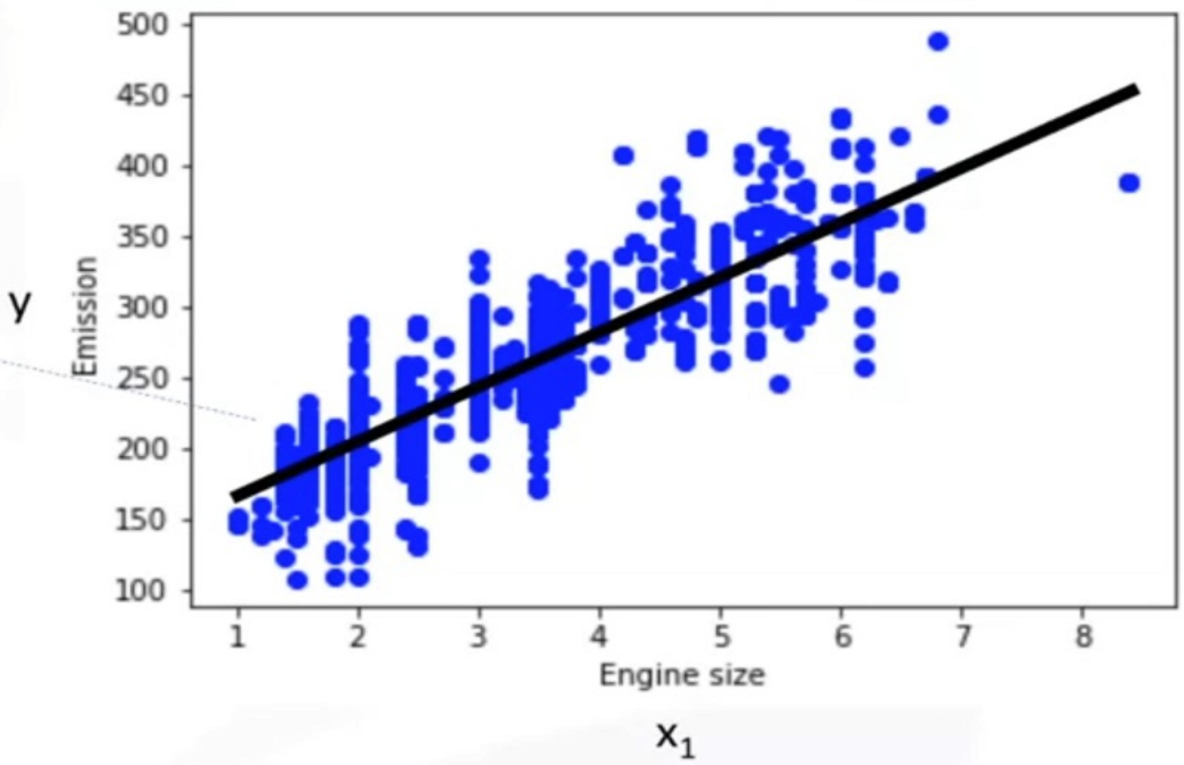
Simple Linear Regression

Parameters we need to find

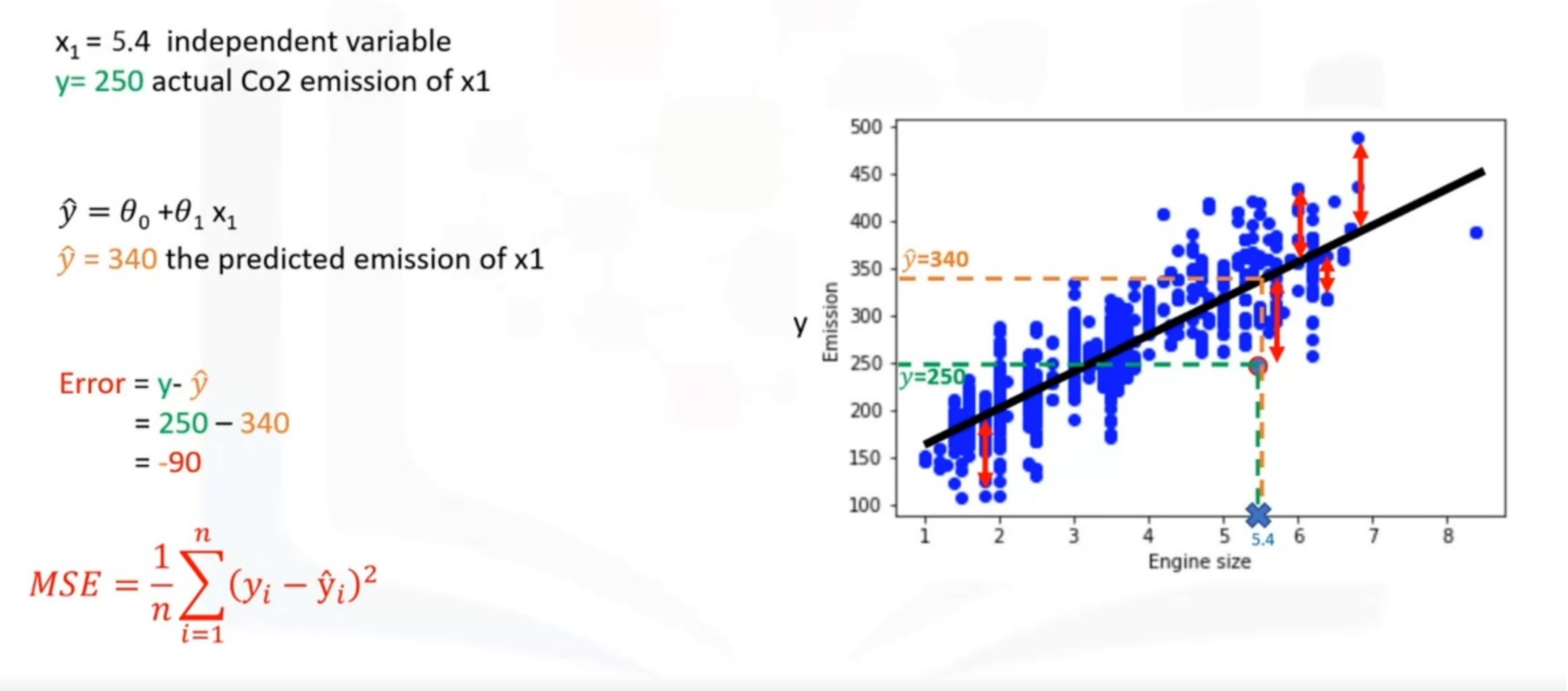
$$\hat{y} = \theta_0 + \theta_1 x_1$$

Independent variable

Independent variable



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

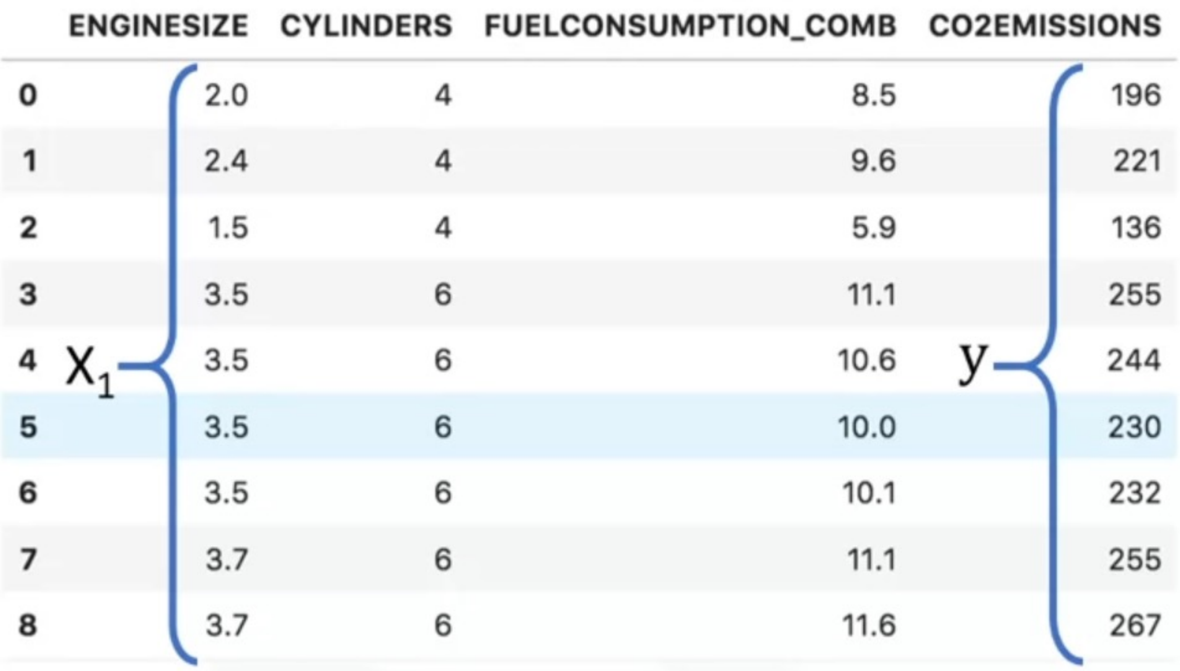
Error = y- $\hat{y}$

= 250 – 340

= -90

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

How to find the best fit? (Mathematic approach)



	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

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

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

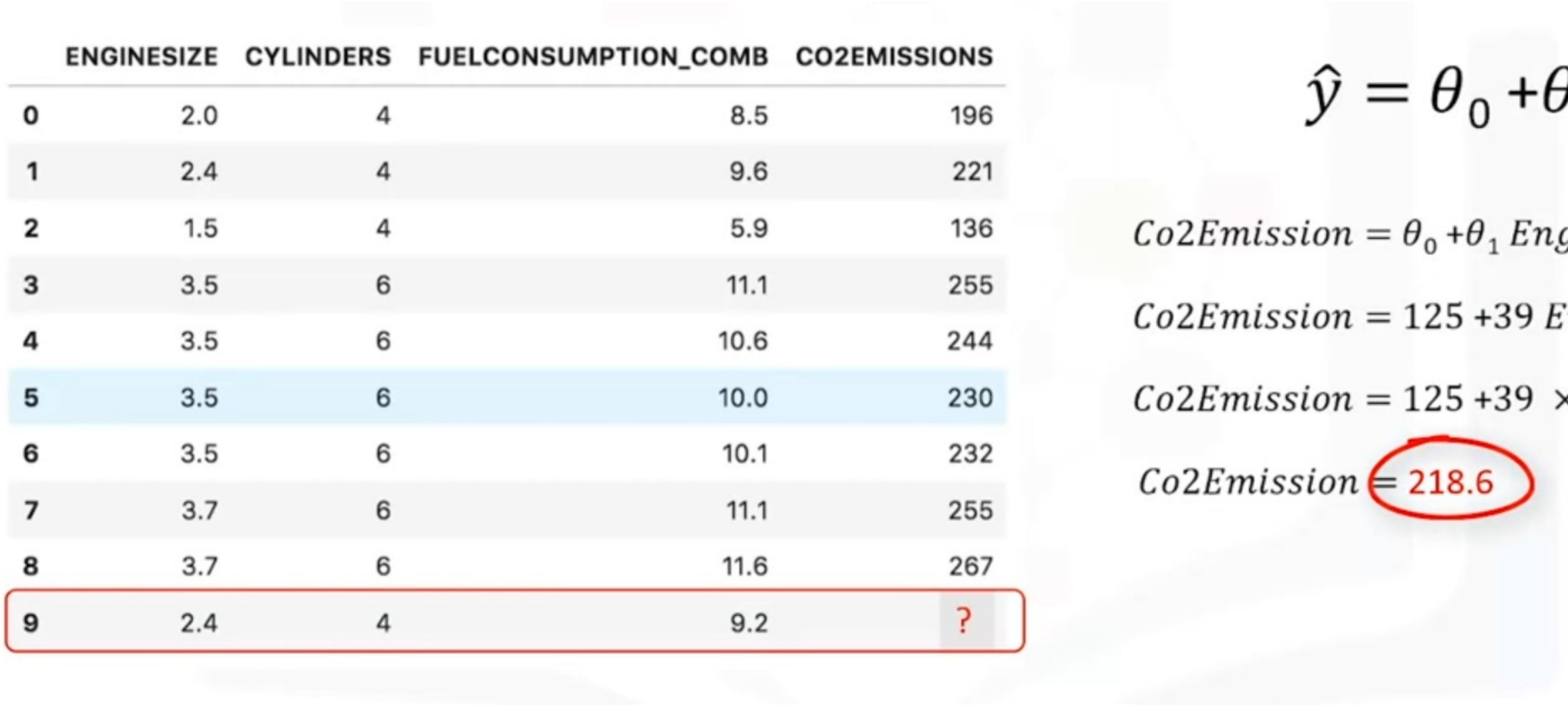
$\bar{x} = (2.0 + 2.4 + 1.5 + \ldots)/9 = 3.03$

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

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

Minimizing the sum of squared errors (SSE)

Prediction with simple 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	?

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

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

$$Co2Emission = 125 + 39 \, EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission \boxed{= 218.6}$$

Advantages

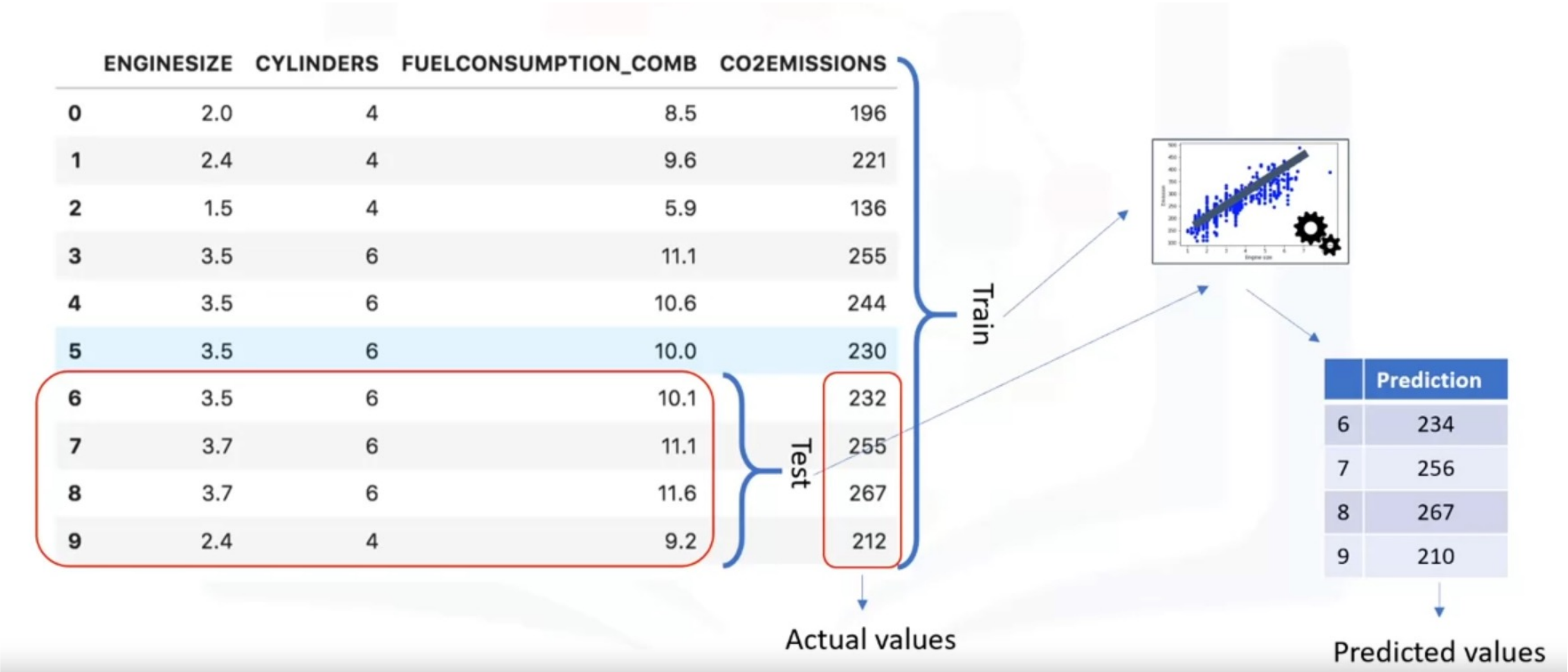
- Very fast
- No hyperparameters tuning
- Intepretable

Disadvantages

- Sensitive to outliers
- Cannot handle non-linear relationships

How to test the accuracy of the model

Train all data

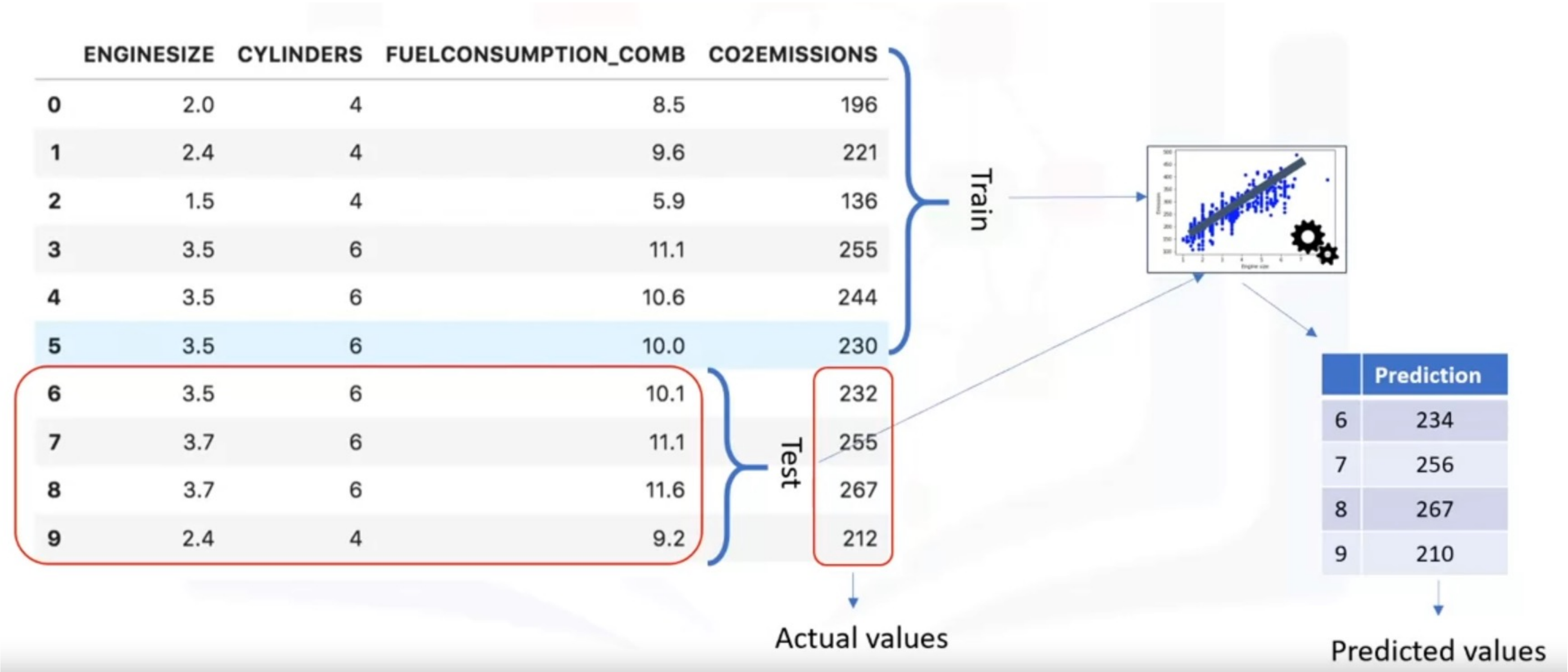


	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

	Prediction
6	234
7	256
8	267
9	210

How to test the accuracy of the model

Train/Test split

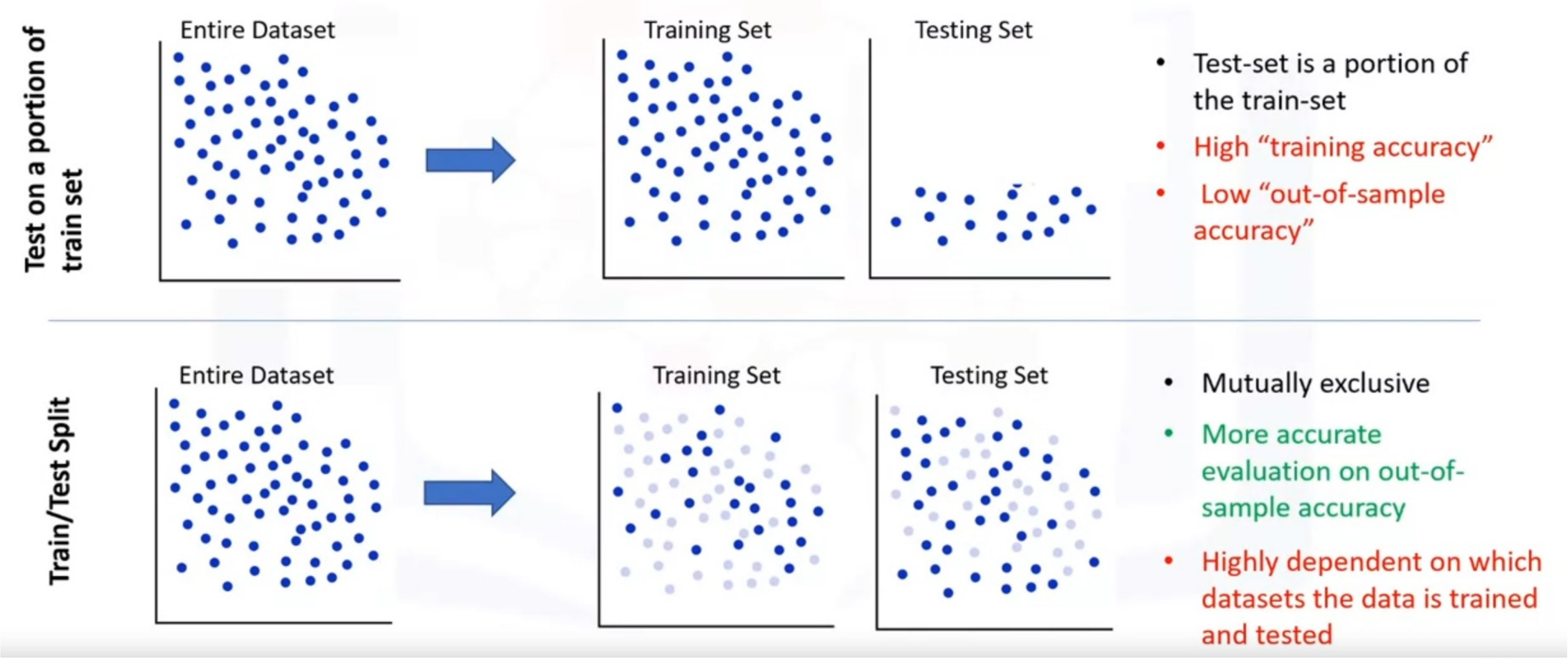


	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

	Prediction
6	234
7	256
8	267
9	210

How to test the accuracy of the model

Train/Test split

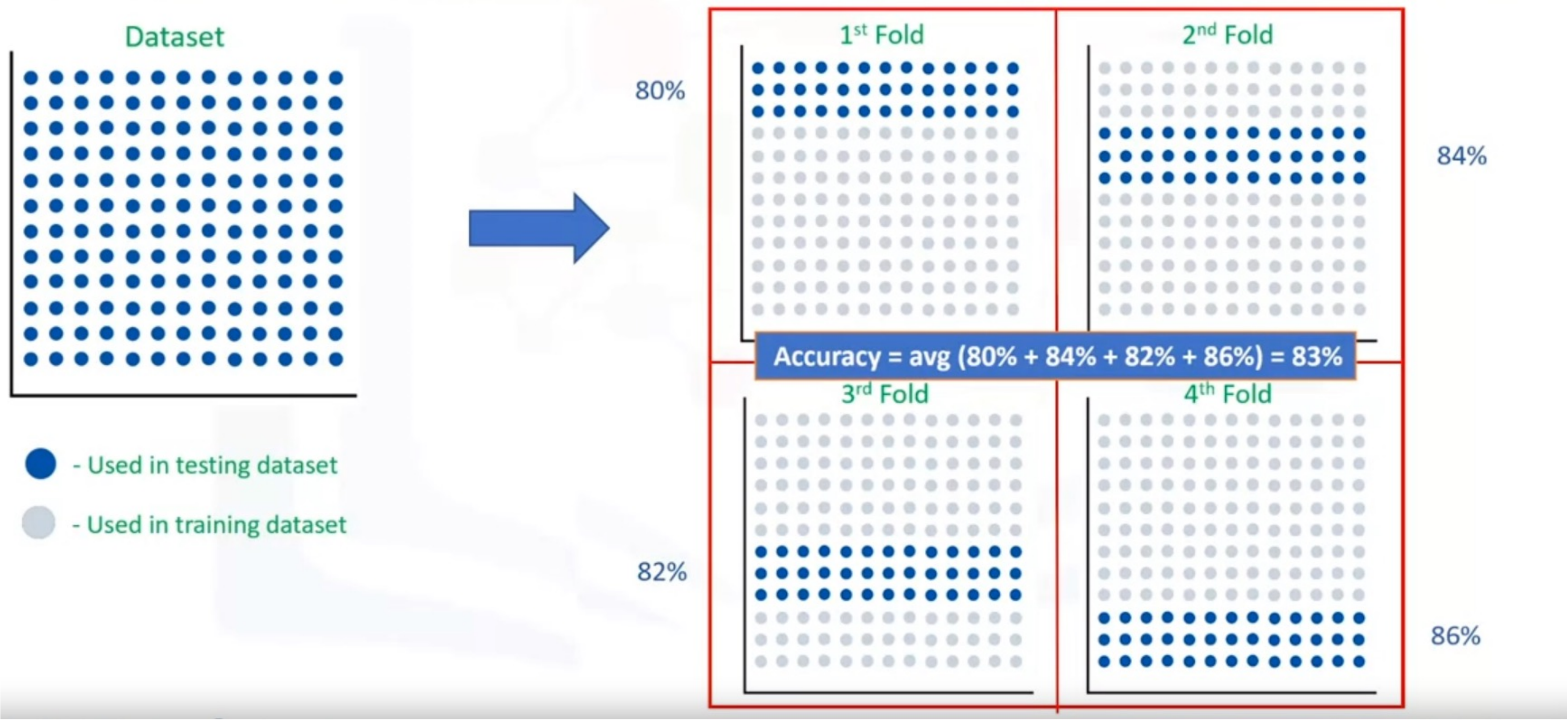


- Test-set is a portion of the train-set
- High “training accuracy”
- Low “out-of-sample accuracy”

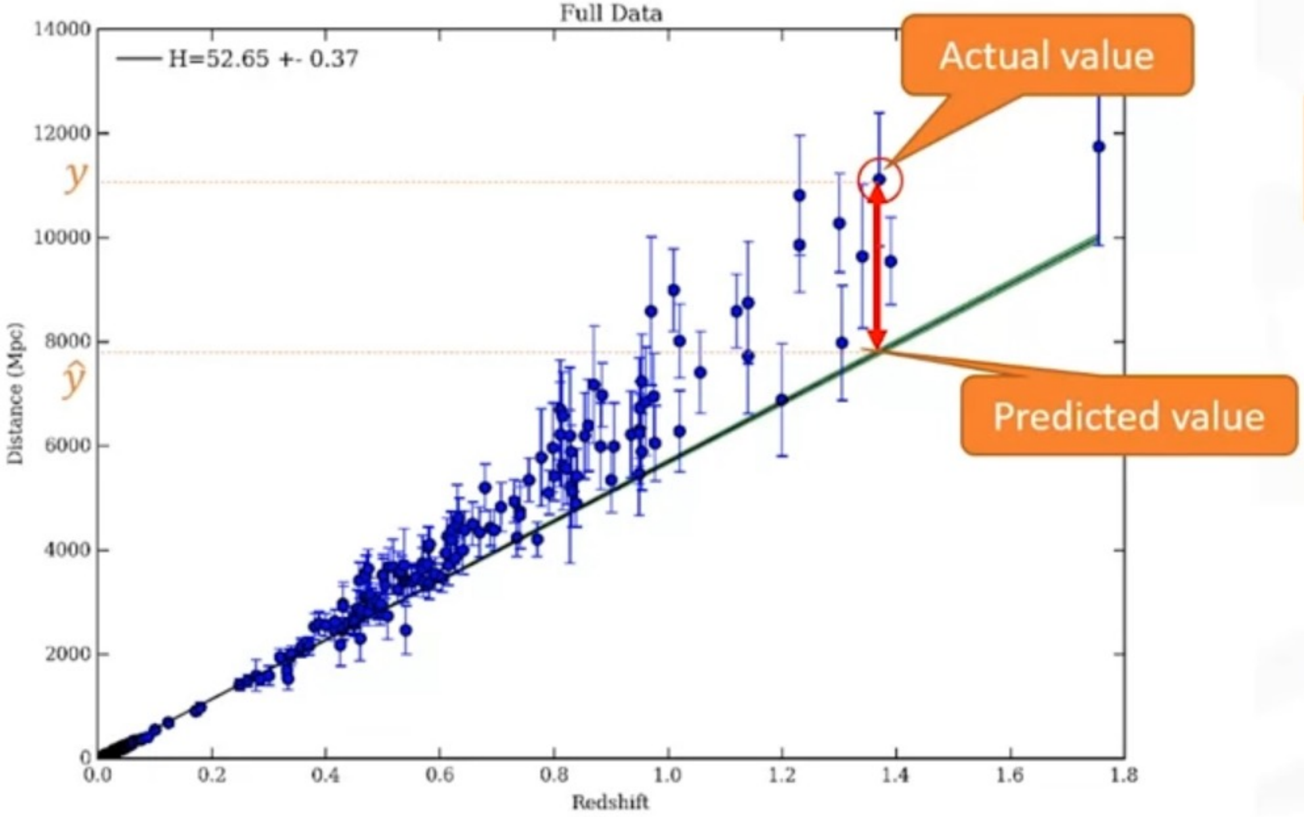
- Mutually exclusive
- More accurate evaluation on out-of-sample accuracy
- Highly dependent on which datasets the data is trained and tested

How to test the accuracy of the model

K-fold Cross Validation


Dataset
● - Used in testing dataset
● - Used in training dataset
1st Fold
80%
2nd Fold
84%
Accuracy = avg (80% + 84% + 82% + 86%) = 83%
3rd Fold
82%
4th Fold
86%

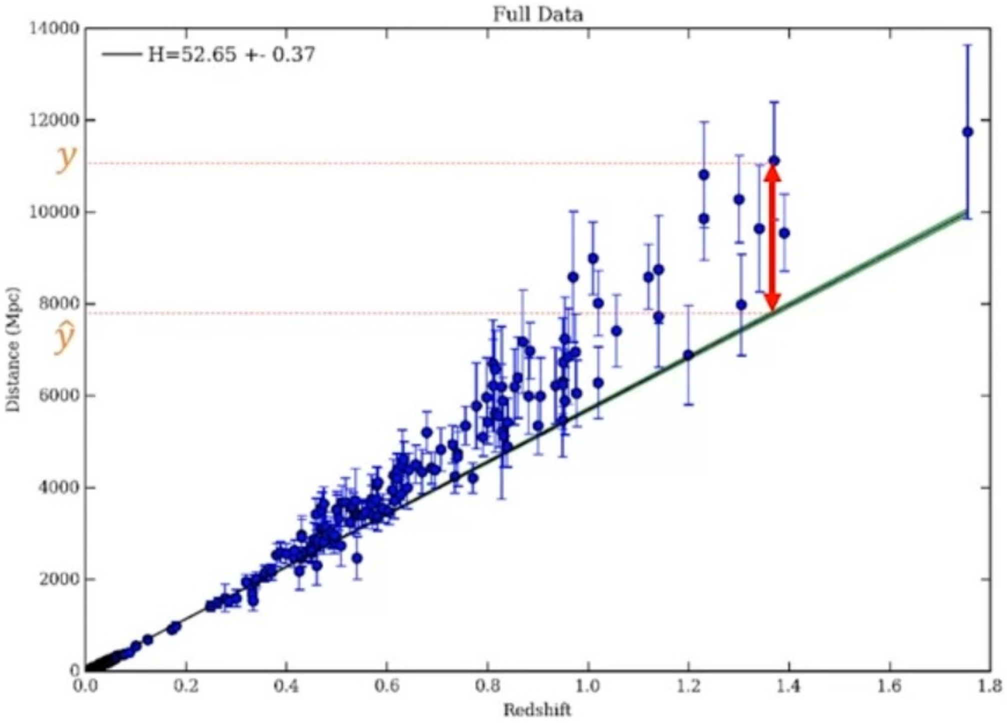
Evaluation Metrics in Regression Models



Error: measure of how far the data is from the fitted regression line.

Error is the difference between the data points and the trend line generated by the algorithm

Evaluation Metrics in Regression Models



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

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

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

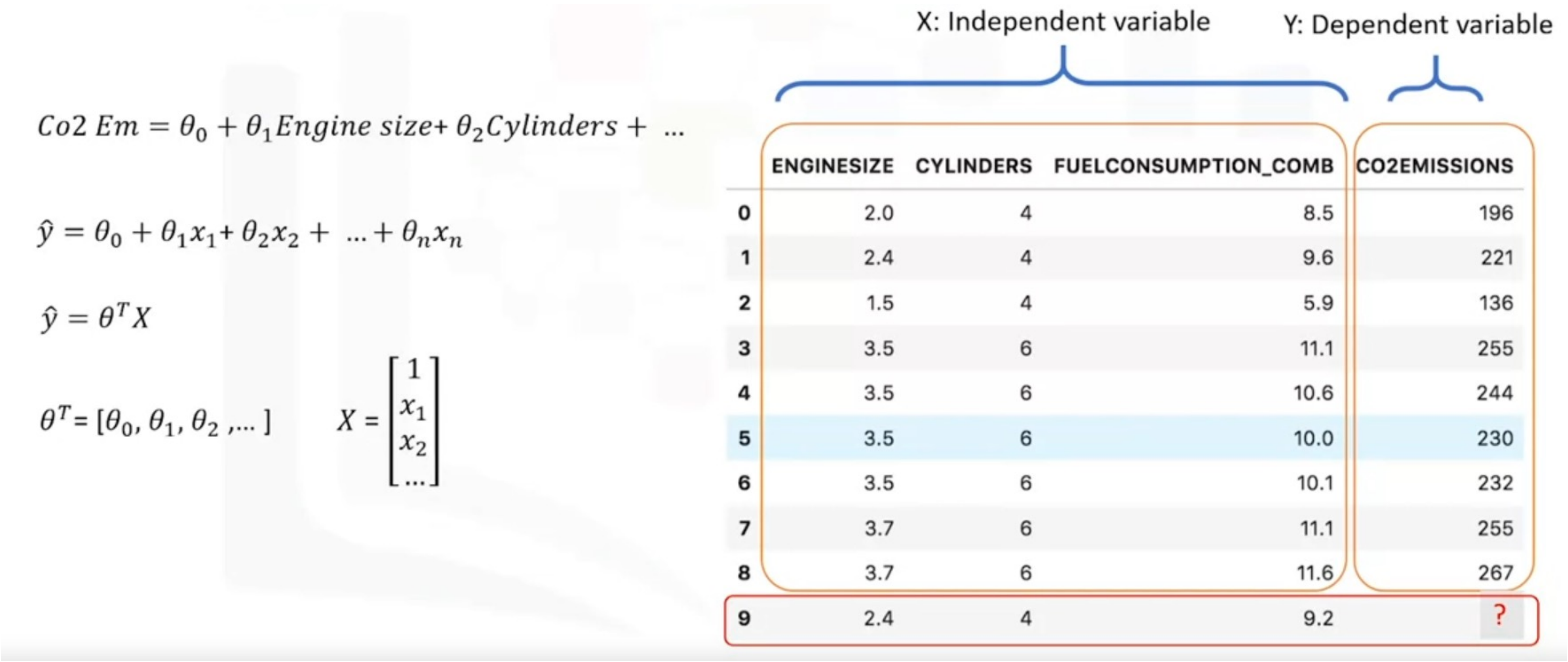
$$RAE = \frac{\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|}{\sum_{j=1}^{n}\left|y_j - \bar{y}\right|}$$

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

$$R^2 = 1 - RSE$$

Multiple Linear Regression

Multiple Linear Regression



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

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

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

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

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	?

How to find best parameters for the Multiple Linear Regression?

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

$\hat{y}_i = 140$ the predicted emission of x_i

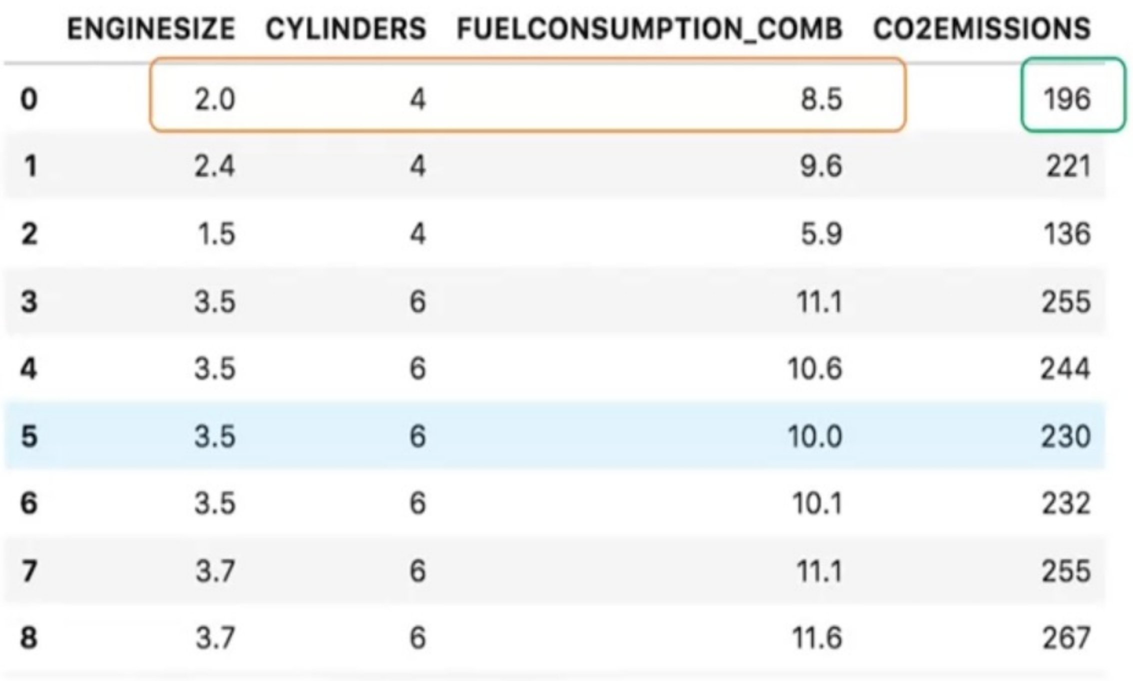
$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

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

↑

Minimize



	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

How to find best parameters for the Multiple Linear Regression?

- Mathematical approach
  - Linear Algebra operation
  - For small dataset
- Optimization approach
  - Gradient Descent
  - For large dataset

Graded Assignment

Using Linear Regression to predict house price for this dataset:

https://www.kaggle.com/datasets/prokshitha/home-value-insights

- Deadline: **Monday, 25.11.2024**

References

- https://en.wikipedia.org/wiki/Linear_regression
- https://machinelearningmastery.com/regression-metrics-for-machine-learning/
- https://www.coursera.org/learn/machine-learning-with-python?specialization=ibm-data-science