# Unit 3
# Linear Regression

Linear Regression

# Objective

- Introduction to Linear Regression
- Regression use case
- Types of regression models
- Regression modelling
- Parameter estimation

# Introduction

- In 1800, a person named Francis Galton, was studying the relationship between parents and their children.
- He investigated the relationship between height of fathers and their sons.
- He discovered that a man's son tends to be roughly as tall as his father, **however, son's height tended to be closer to the overall average height of all people.**

**Galton call this phenomenon as "Regression" as "father's son height tends to regress (or drift towards) the mean (average) height of everyone else.**

# Linear Regression

- Regression is used to study the relationship between two variables.
- We can use simple regression if both the **dependent variable (DV)** and the **independent variable (IV)** are numerical.
- If the DV is numerical but the IV is categorical, it is best to use Logistic Regression.

# Linear Regression - Example

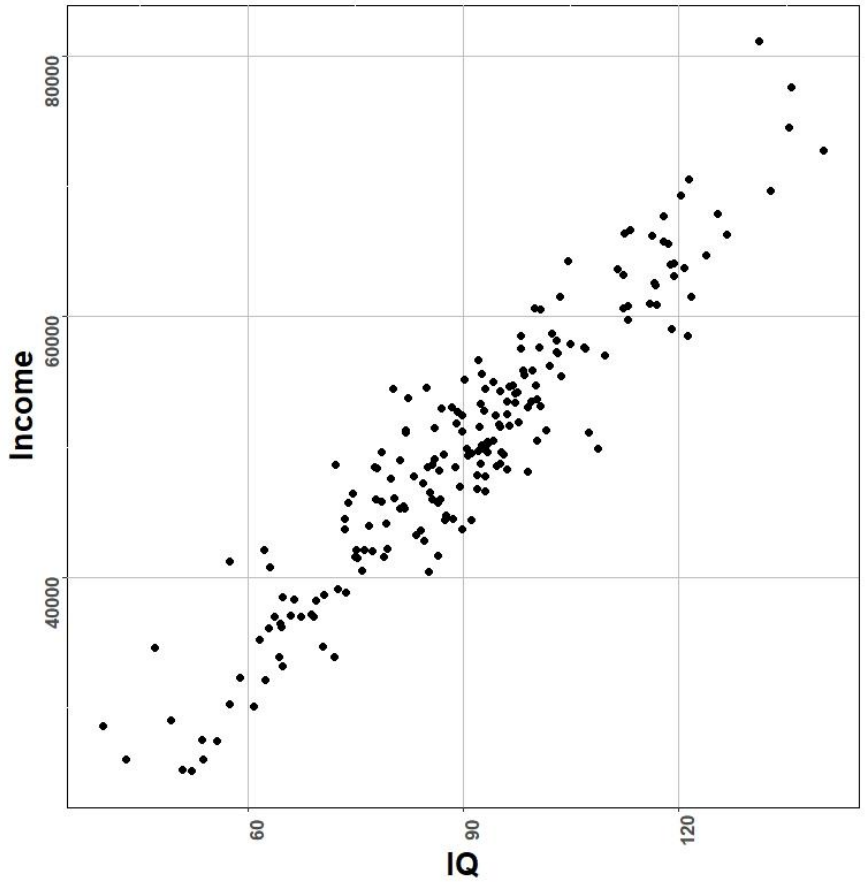
The following are situations where we can use regression:

1. Testing if IQ affects income (IQ is the IV and income is the DV).
2. Testing if hours of work affects hours of sleep (DV is hours of sleep and the hours of work is the IV).
3. Testing if the number of cigarettes smoked affects blood pressure (number of cigarettes smoked is the IV and blood pressure is the DV).
4. Chances of heart failure due to high body fat

# Displaying the data

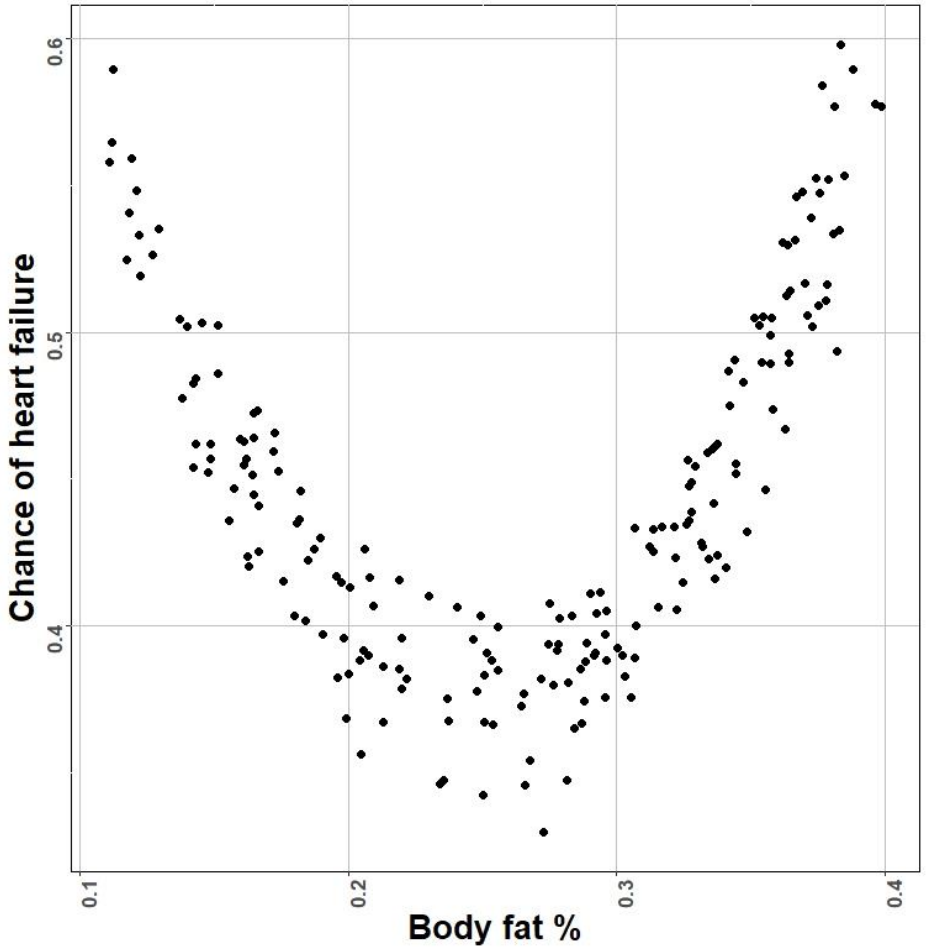
- Displaying data for Testing if IQ affects income (IQ is the IV and income is the DV).

- When both the DV and IV are numerical, we can represent data in the form of a scatterplot.



# Displaying the data

- Displaying data of Chances of heart failure due to high body fat
- It is important to perform a scatterplot because it helps us to see if the relationship is linear.



# Regression Case

- Dataset related to Co2 emissions from different cars.

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSION |
|---|---|---|---|---|
| 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 Case

- Looking to the existing data of different cars, can we estimate the approx. CO2 emission of a car, *which is yet not manufactured*, such as in row 9 ?

- We can use regression methods to predict a continuous value, such as CO2 Emission, using some other variables.

- In regression there are **two types of variables:**
  1. a dependent variable (DV, which we want to predict) and
  2. one or more independent variables (IV, existing features).

# Regression Essentials

- The key point in the regression is that our dependent value should be continuous and cannot be a discreet value.

- However, the independent variable or variables can be either a categorical or continuous.

- We use regression to build such a regression/estimation model which would be used to predict the expected Co2 emission for a new or unknown car.

# Types of Regression Model

1. **Simple regression** is when one independent variable is used to estimate a dependent variable.
    - It can be linear or non-linear.
    - Ex: predicting Co2 emission using the variable of Engine Size.
2. When more than one independent variable is present, the process is called **multiple linear regression**.
    - Ex: predicting Co2 emission using Engine Size and the number of Cylinders in any given car.

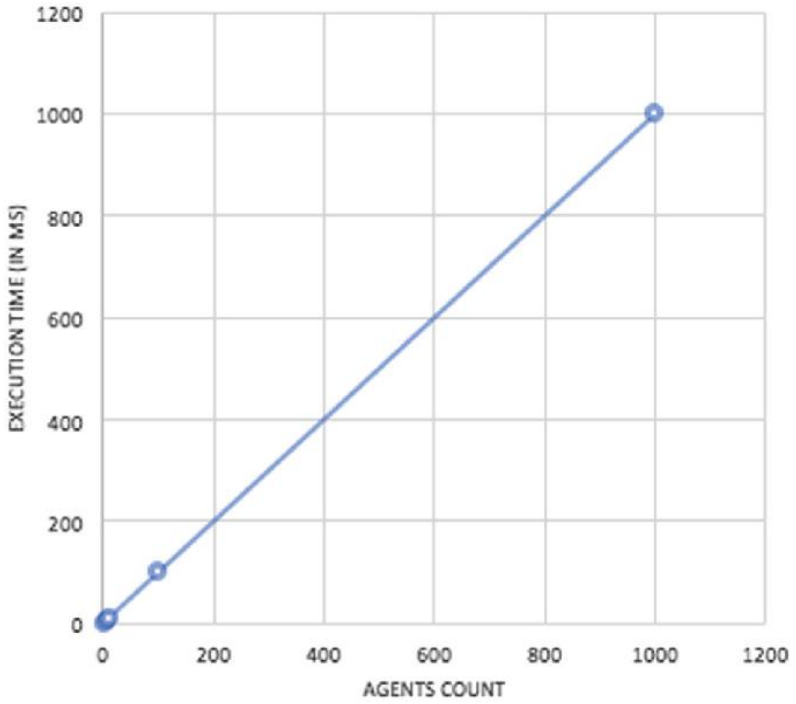
Linearity of regression depends on the relation between dependent and independent variables; it can be either linear or non-linear regression.

# Regression Application Areas

- Essentially, we use regression when we want to estimate a continuous value.
- You can try to **predict a salesperson's total yearly sales (sales forecast)** from independent variables such as age, education, and years of experience.
- We can use regression analysis to **predict the price of a house** in an area, based on its size, number of bedrooms, and so on.
- We can even use it to **predict employment income** for independent variables, such as hours of work, education, occupation, sex, age, years of experience, and so on.

# Simple Linear Regression

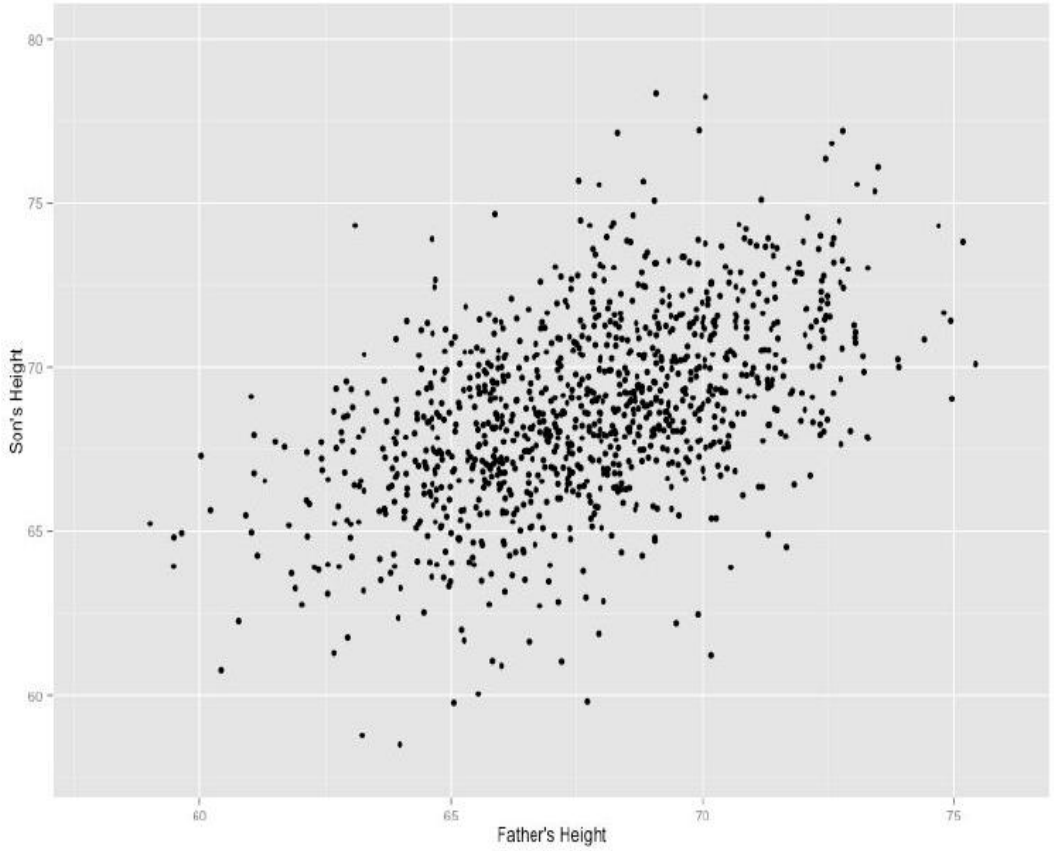
- How to calculate a regression with only 2 data points ?
- In linear regression, we calculate regression line by drawing a connecting line
- For classic linear regression or “Least Square Method”, you only measure the closeness in the “up and down” direction.
- Here we have perfectly fitted line because we have only 2 points.




Simple Linear Regression

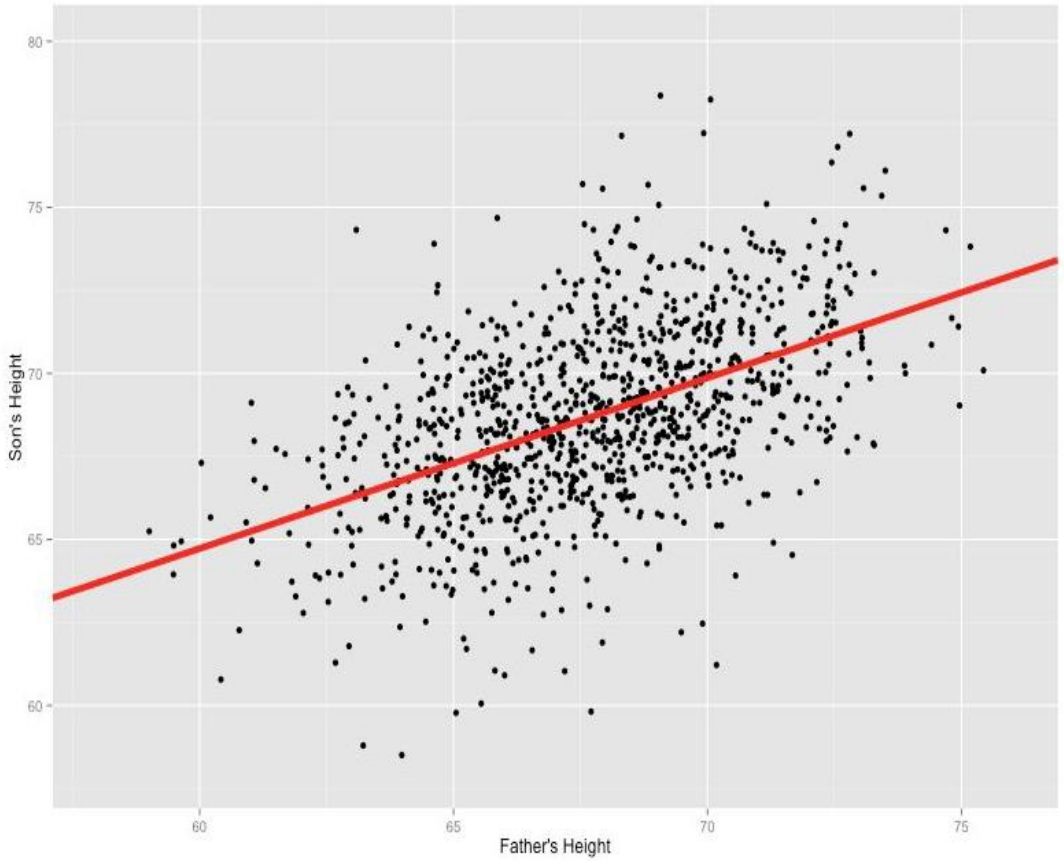
# Regression with more data points

- Now wouldn't it be great if we could apply this same concept to a graph with more than just two data points?
- By doing this, **we could take multiple men and their son's heights and do things like tell a man how tall we expect his son to be...before he even has a son!**
- This is the idea behind supervised learning!



# Regression Goal

- Our goal of linear regression is determining the best line by minimizing the vertical distance between all the data points and our line.
- There are lots of different ways to minimize this, (sum of squared errors, sum of absolute errors, etc), but all these methods have a general goal of minimizing this distance between your line and rest of data points.



# Case Study

- This dataset is related to the Co2 emission of different cars.
- The question is: Given this dataset, can we predict the Co2 emission of a car, using another field, such as Engine size?

Y: Dependent variable

X: Independent variable

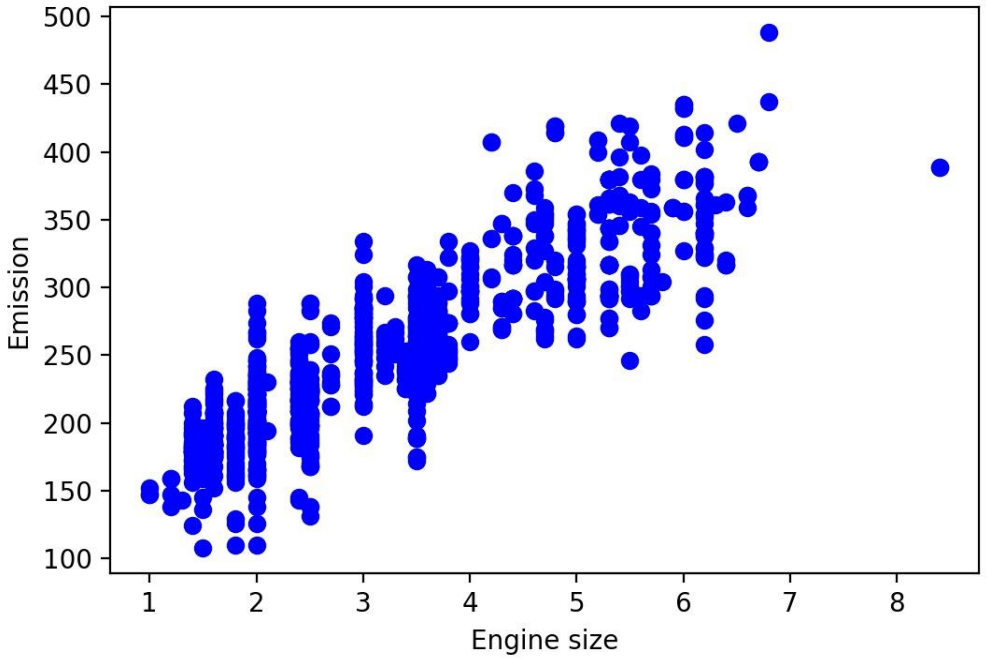
| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSION |
|---|---|---|---|---|
| 0 | 2.0 | 4 | 8.5 | 196 |
| 1 | 2.4 | 4 | 9.6 | 221 |
| 2 | 1.5 | 4 | 5.9 | 136 |
| 3 | 3.5 | 6 | 11.1 | 255 |
| 4 | 3.5 | 6 | 10.6 | 244 |
| 5 | 3.5 | 6 | 10.0 | 230 |
| 6 | 3.5 | 6 | 10.1 | 232 |
| 7 | 3.7 | 6 | 11.1 | 255 |
| 8 | 3.7 | 6 | 11.6 | 267 |
| 9 | 2.4 | 4 | 9.2 | ? |

Continuous values

Yes!

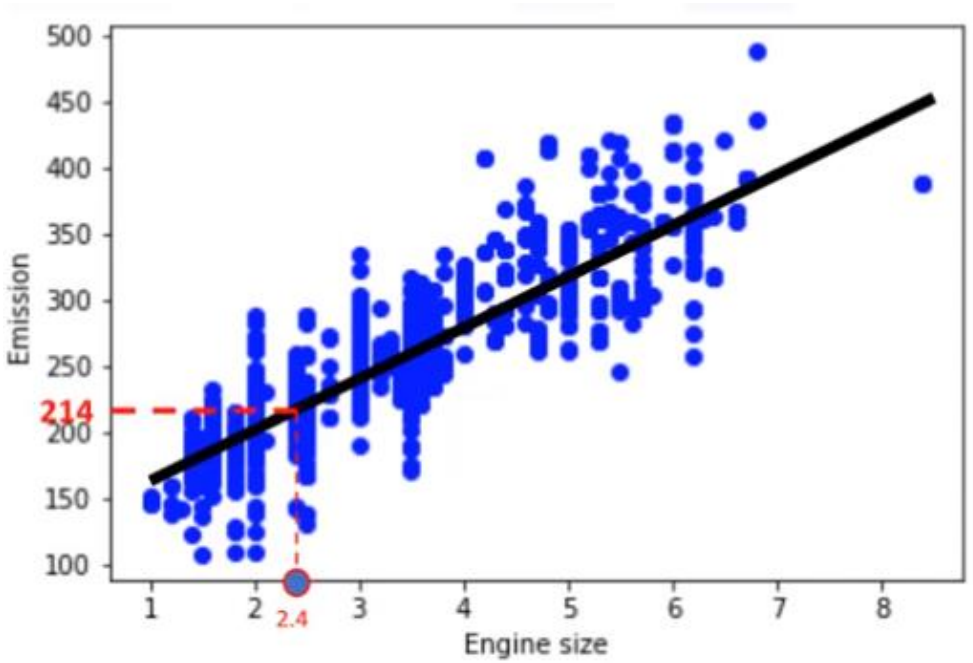
# Scatter Plot

- To understand linear regression, we can plot our variables here.
- **Engine size** -- independent variable, **Emission** – dependent/target value that we would like to predict.
- A scatterplot clearly shows the relation between variables where changes in one variable "explain" or possibly "cause" changes in the other variable.
- Also, it indicates that these variables are linearly related.



# Inference from Scatter Plot

- **As the Engine Size increases, so do the emissions.**
- How do we use this line for prediction now?
- Let us assume, for a moment, that the line is a good fit of data.
- We can use it to predict the emission of an unknown car.



# Regression Modeling – Fitting Line

- Fitting line help us to predict the target value, Y, using the independent variable 'Engine Size' represented on X axis
- The fit line is shown traditionally as a polynomial.
- In Simple regression Problem (single x), the form of the model would be

$$\hat{y} = \theta_1 + \theta_2 x_1$$

**$\theta_1$ = intercept $\theta_2$ = slope of the line**

- Where Y is the dependent variable, or the predicted value and X is the independent variable.
- $\theta_1$ and $\theta_2$ are coefficient of linear equation

# Regression Modeling

$$\hat{y} = \theta_1 + \theta_2 x_1$$

Now the questions are:

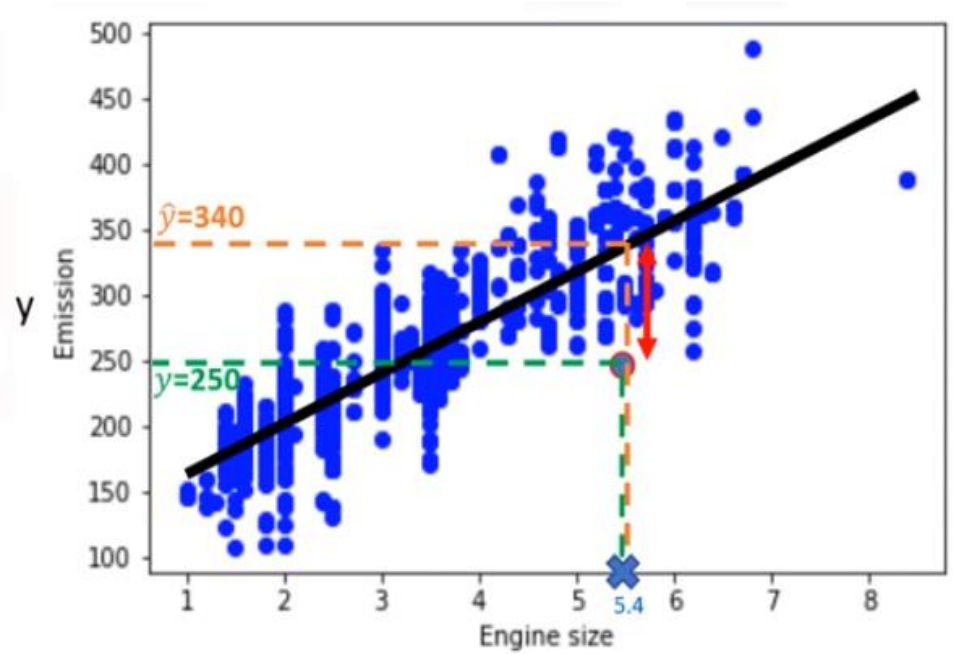
"How would you draw a line through the points?"

"How do you determine which line fits best?"

- **Linear regression estimates the coefficients of the line.**
- This means **we must calculate $\theta_0$ and $\theta_1$ to find the best line to 'fit' the data.**
- Let's see **how we can adjust the parameters to make the line the best fit for the data ?**
- Let's assume we have already found the 'best fit' line for our data.

# Model Error

- If we have, for instance, a car with engine size $x_1$ =5.4, and actual Co2=250,
- Its Co2 should be predicted very close to the actual value, which is y=250, **based on historical data**.
- But, if we use the fit line it will return
  ŷ =340.
- C**ompare the actual value with we predicted using our model, you will find out that we have a 90-unit error.**
- Prediction line is not accurate. This error is also called the **residual error.**




Error = ŷ – y = 340-250 = 90

linear-regression-machine-learning

# Mean Absolute Error

$$MAE = \frac{1}{n}\sum \left| y - \hat{y} \right|$$

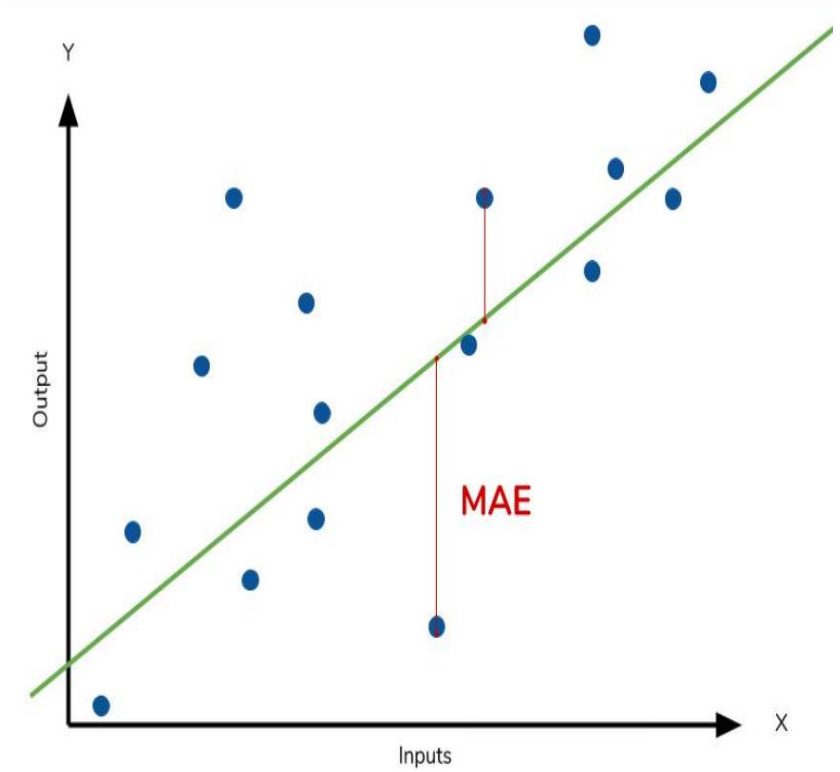
Divide by the total number of data points

Predicted output value

Actual output value

Sum of

The absolute value of the residual




In this, the residual for every data point, taking only the absolute value of each so that negative and positive residuals do not cancel out. Then take the average of all these residuals.

# Mean Squared Error

The mean square error (MSE) is just like the MAE but squares the difference before summing them all instead of using the absolute value. We can see this difference in the equation below.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

# $R^2$ Score

R-squared ($R^2$) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

So, if the $R^2$ of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

**Formula for R-Squared**

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

# Parameter Estimation

- **The objective of linear regression is to minimize this MSE equation, and to minimize it, we should find the best parameters, $\theta_0$ and $\theta_1$ .**

- How to find $\theta_0$ and $\theta_1$ in such a way that it minimizes this error?
- **We have two options here:**
  - Option 1 - We can use a mathematic approach Or,
  - Option 2 - We can use an optimization approach.

# Mathematical Approach

- $\theta_0$ and $\theta_1$ (intercept and slope of the line) are the coefficients of the fit line.
- Need to calculate the mean of the independent and dependent or target columns, from the dataset.

- Notice : All of the data must be available.
- It can be shown that the intercept and slope can be calculated using these equations.
- We can start off by estimating the value for $\theta_1$.

# Parameter Estimation

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

| | ENGINESIZE | CYLINDERS | FUELCONSUMPTION_COMB | CO2EMISSION |
|---|---|---|---|---|
| 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 | ? |

$$\boldsymbol{\Theta_1} = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x}$ = 3.34

$\bar{y}$ = 256

$$\Theta_1 = \frac{(2-3.34)(196-256)+(2.4-3.34)(221-256)+ \ldots\ldots\ldots}{(2.0-3.34)^2+(2.4-3.34)^2+ \ldots..}$$

$$\Theta_1 = 39$$

$$\Theta_0 = \bar{y} - \Theta_1 \bar{x} = 125.74$$

# Making Predictions

- We can write down the polynomial of the line.

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

- Making predictions is as simple as solving the equation for a specific set of inputs.
- Imagine we are predicting Co2 Emission(y) from EngineSize(x) for the Automobile in record number 9. So, looking to the dataset**, $x_1$ = 2.4**

- Implementing $x_1$ in above equation, **we can predict the CO2 emission of this specific car (row 9) with engine size 2.4 :**

$$\hat{y} = \mathbf{218.6}$$

# Hands On