# **1. Setting the Stage: Introduction to Machine Learning**

## **Can we Learn from the Data?**

* Absolutely Yes. We can decipher patterns that can unravel secrets or help us understand things better.

|  |  |
| --- | --- |
| * We can find patterns in nature. * A fingerprint is unique pattern in human nature, that can be recognised and used to identify the person. * Similarly, we can find patterns in data to make predictions. | Figure 1 A Human Fingerprint |

**We want to use the data to learn about the real world.**

One of the biggest **assumptions** about learning from data is that “***Past is a good representation of future”.***

A diagram of a line and a line

Description automatically generated

Figure 2 Timeline, Past & Future

*Future is not the same as past, but it is similar to past.*

Past (data) is made up of 2 Components:

1. Information
2. Noise

A diagram of a line and a line

Description automatically generated

Figure 3 Data = Information + Noise

Our task is to look into the past (data) and capture **Information** and leave out **noise**. We achieve this by creating a **model (mathematical)** that summarizes or generalizes the information.

Such mathematical summaries or generalizations can be used for future predictions.

## **So, the question is What is a Model?**

A Model is a 3-dimensional representation of a person or thing or of a proposed structure, typically on smaller scale than original.



Figure 4 Model of a Housing Society

In General, Data Science includes everything that we do learn about the real world. Its starts with precise observations (which is data) and ends at mathematical summary (a model).

An example could be a bunch of observations in the real world pertaining to Grade Point Average (GPA) of students and their respective salaries.

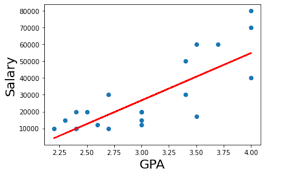


Figure 5 GPA vs Salary

Here our model is a Straight line, and we can model this line as a mathematical equation in the slope intercept format.

|  |  |
| --- | --- |
| Figure 6 Slope, Intercept Format | Figure 7 Can you calculate the Y-Point |

Model could be a straight line or some other form. What important is our ability to express it mathematically.

For example, if you're predicting someone's salary based on their years of experience, your linear regression model might look like this:

***Salary = (Slope x Experience) + Intercept***

* The intercept is the predicted salary when a person has zero years of experience.
* The slope tells you how much salary increases for each additional year of experience.

So, the intercept represents the base value or starting point of your prediction, even before considering the effect of the input variable (like experience). It is like saying, “If I have no experience at all, this is what my salary would likely be.”

## **What Is a Machine Learning Model?**

* A Machine Learning Model can be a mathematical representation of a real-world process.
* An ML model takes the training data as input and with help of an ML algorithm forms a mathematical expression that gives us an output.
* So, an example could be House Price prediction model – that will take the area of the house, the number of rooms, lawn, pool etc as input parameters and give the price as output.

Model Building is a 3 Step Process:

* Take the data as input.
* Find patterns in the data (in the GPA vs Salary case we found a positive relationship or pattern)
* Summarize the pattern in a mathematically precise way. (we can find the parameters for the GPA vs Salary line in the slope intercept form)
* Make predictions using the Mathematical model. (So, in the GPA vs salary case, for any give GPA we can predict the salary.)

Machine Learning automates this entire model building process.

**What is so challenging about Machine learning?**

* Data unfortunately contains noise and our mathematical summary should model only information.
* Think of **DATA = Information + Noise** The Challenge is to identify the **information** content and distil away the noise. A diagram of a graph

  Description automatically generated

Figure 8 Misclassification (challenges ML faces)

* In the above visual, the line is the model.
* It does a fairly good job in predicting the squares and the circles.
* However, there are 3 instances of noise in the data which makes the model performance suffer.
* In **Statistical Modelling**, we assume that the model we are making is a linear model or a quadratic model or some other nonlinear model. We make this assumption upfront.
* With **Machine Learning** we don’t make such assumptions, with machine learning we try out different models and then evaluate the performance of each model and subsequently go with the winner.
* To do this kind of evaluation we split the data into test and train set.

## **Train and Test Split**

A diagram of a graph

Description automatically generated with medium confidence

Figure 9 Train & Test Split

* We divide the dataset into 2 subsets: One for **Training** the model and the other for **testing** its performance.
* This splitting is important to assess how well the model generalizes to *new, unseen data*.
* Its crucial that to randomize the data before splitting to ensure that both sets are representative of the overall distribution of dataset.
* The size of the training and test sets depends on the available data. Common split is 70-30 or 80-20.

|  |
| --- |
| Figure 10 Problem of complicated model -Overfit |
| * The most complicated model will fit the data points well but, in the process, it will end up modelling both information and noise. * Such a model cannot generalize or predict future well. While it will explain all the points in training data set, it fails miserably with the test set. |

## **Model Overfit and Underfit**

* If the Model that we created ends up:
  + Modelling the noise as well as the information, we call it **“Over-fit”** – this is bad for future predictions.
  + Not modelling all the information, we call it **“Under-fit”** - this again bad for future predictions.

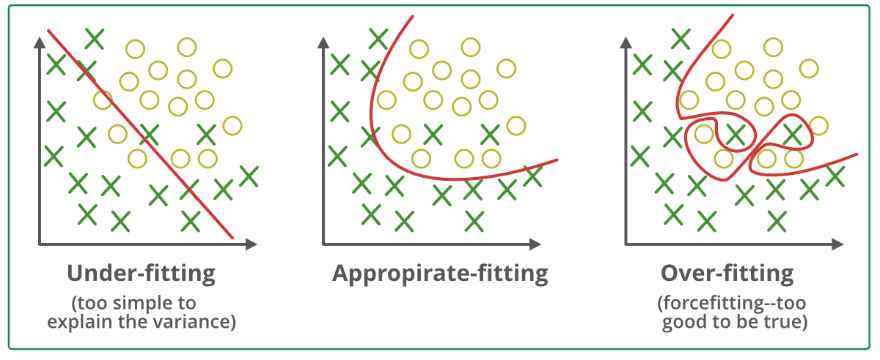


Figure 11 Under-fit, Appropriate fit and Over-fit models

In the visual above:

* The first model is simple line but it’s under fit as it misclassifies a couple of X as O and vice versa.
* The Third Model is super complicated, it does not misclassify any of the x’s and o’s but it difficult to rely on its future predictions. This is an overfit mode.
* The second model (a curve of some sort) does the best job, it wrongly classifies 2 x’s and get all the o’s correct. We can rely on this model for future predictions.

## **Variables – Dependent and Independent**

* Variables are attributes or characteristics in the data that can vary and take on different values.
* Variables are essential components of analysing data.
* **Definition** - Variables are measurable characteristics that can vary or change in the research or statistical analysis of data.
* **Example** - Age, Height, test scores, temperature, time, and income are all examples of variables.

**Types of Variables**

A comparison of different types of plants

Description automatically generated

Figure 12 Types of Variables - Dependent & Independent

**Dependent Variable –**

* **Definition-** The dependent variable is the outcome or result that researchers are interested in predicting, explaining, or understanding.
* **Example -** In examining the effect of a new drug on patients, the dependent variable would be the patients’ health improvement.

**Independent Variable –**

* **Definition-** The independent variable is the variable that is manipulated or controlled in an experiment.
* **Example-** In the same drug study, the independent variable would be the administration of the new drug. Researchers would manipulate or vary the independent variable to observe its effects on the dependent variable.

# **Machine Learning Tasks**

There are 2 types of Machine Learning tasks:

1. Supervised Learning task
2. Un-Supervised Learning task.

A screenshot of a graph

Description automatically generated

Figure 13 Supervised vs Un-Supervised Learning

**Supervised Learning** is building a mathematical model using data that contains both the inputs and the desired output (labels.)

* **Example 1** - Determining if the image has a horse. The data would include images with and without the horse (the input), and for each image we would have a label (the output) indicating if there is a horse in that image.
* **Example 2** - Determining if a client might default a loan.
* **Example 3** - Determining if the employee is likely to quit.

**Unsupervised Learning** is building a mathematical model using data that contains only inputs and no desired output (labels).

This is used to find structure in the data, like grouping or clustering of data points. Help us discover groups or patterns or categories within the input data.

* **Example** - an advertising platform segments the population into smaller groups with similar demographics and purchasing habits. Helping advertisers reach their target market with relevant ads.

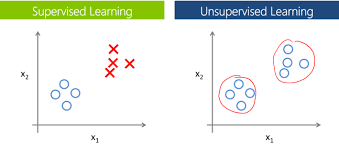


Figure 14 Evaluating the performance of Unsupervised could be a challenge.

In un-supervised learning, since we don’t have a label or desired output, it becomes a challenge to evaluate the performance of the model.

# **Machine learning Tools and Techniques**

A diagram of machine learning

Description automatically generated

Figure 15 Machine Learning tools & techniques

**Supervised Learning**

* **Regression:** Desired output is a continuous number
* **Classification:** Desired output is a category.

**Unsupervised Learning**

* **Clustering:** Grouping data (K-means).
* **Dimensionality reduction**: Compressing data.
* **Association rule learning:** If `x` then `y`

# **2. Unveiling the Mysteries of Linear Regression**

Greeting fellow learners, now we embark on our quest to understand the elegant dance of data. Let us delve into the foundational realms of linear regression.

**The Essence of Linear Regression**

* Linear regression, in its simplest guise, is a narrative of relationships told through the language of mathematics.
* It is the harmonious symphony that seeks to uncover the linear association between independent and dependent variables.

**Dual Goals:**

* *We want to define the relationship between the 2 variables.*
* *We want to make predictions using this relationship.*

Before, getting to the insightful Linear model lets understand an instance of a Dumb (un-intelligent) model.

A dumb model is one where we don’t have an independent variable that is providing us with some intelligence.

For example, let’s say we have the mpg (mileages) of 15 different cars, and we also have the weight (in Kg) of these cars.

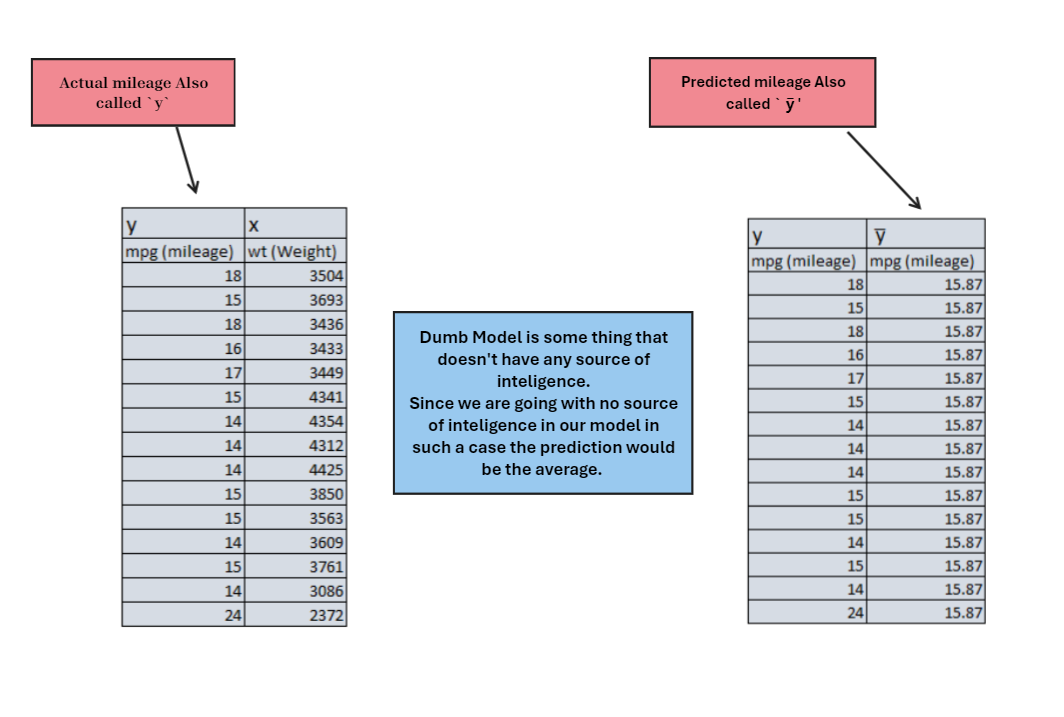


Figure 16 A Dumb (no source of Intelligence) Model

* Ideally, we should first establish the relation between the mpg (dependent variable) to Weight (independent variable) and subsequently use this relationship for further predictions.
* So, the source of intelligence in our model would be the independent variable Weight.
* In a Dumb model we don’t have any source of intelligence, we don’t have the independent variable (weight in this case). What could be the dumb model prediction?
* Since the model doesn’t have any source of intelligence it will go with the **average** as the prediction.
* For a dumb model all the prediction would be the same and that would be the average. In the above case the average of all the 15 mpg observations is 15.87 miles per gallon (mpg).
* We further need to evaluate how good or bad the prediction of this dumb model is.

A screenshot of a computer

Description automatically generated

Figure 17 Dumb Model Performance check

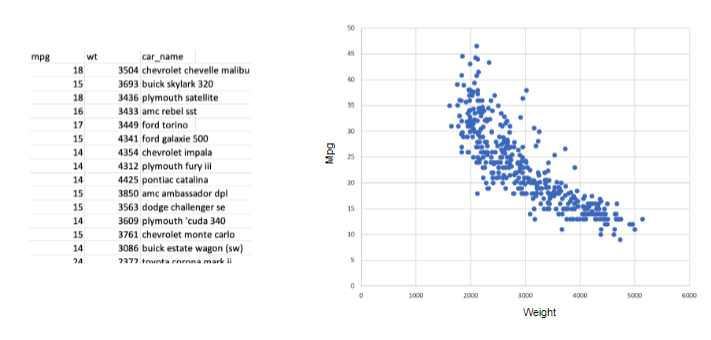
* Error is the difference between the actual and the predicted.
* We took a sum of the errors, and to our great surprise, even though the prediction were wrong on each of the 15 instances, the sum of errors is a remarkable small number. Why is this happening?
* Turns out there is directionality to the errors (some are positive some are negative) and hence they cancel out.
* Error in itself is not good because of the directionality; hence we square the errors. We can always normalize this later by taking a square root.
* The final take away is that the Sum of Square Errors of any intelligent model should be less than the dumb model. Else the average as a model is better than any other model.

## **Now can we answer a simple question?**

***What sort of relationship is there between weight and mileage?***

***Do heavier cars have lower mileage?***

***Can we use data to better understand the relationship between the two variables: Weight and mpg?***



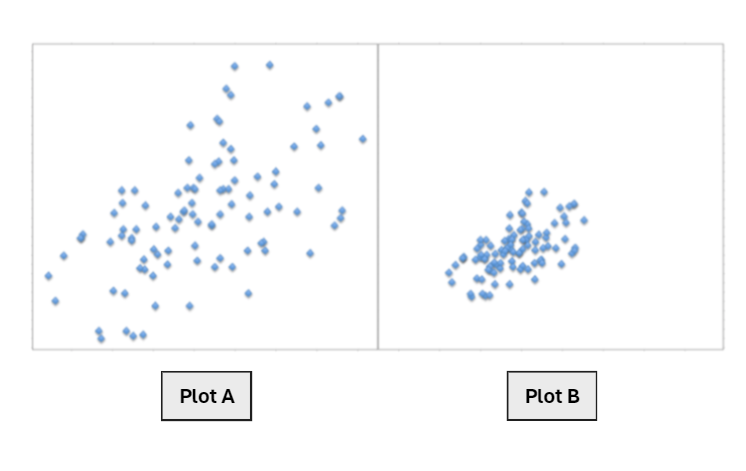
* Easiest thing to do is to plot the data in a `x` and `y` axis.
* Weight which is the independent variable takes on the `x` axis and mpg takes on the `y` axis.
* What does the pattern look like? Here it’s a negative relationship. As the weight increase, the mpg decreases.
* ***So, we can conclude that the heavier cars have lower mileage.***
* What could be the other kind of relationships:

A diagram of negative correlation

Description automatically generated with medium confidence

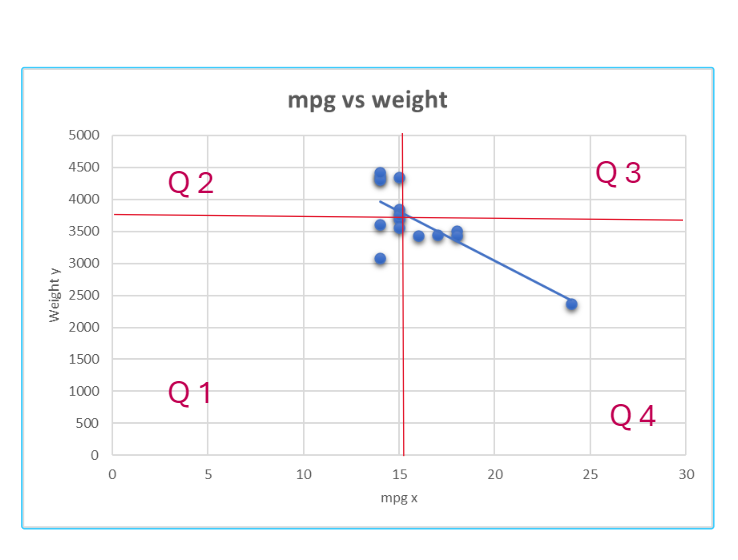
* So, we could have positive, negative or no relationship scenarios.
* Keep in mind we are only looking at linear relationships, nonlinear relationships will discuss elsewhere.
* We now know the kind of relationship that exists between weight and mpg. But we have not qualified it and hence we still can’t make predictions.

## **Which visual below has stronger relationship between x and y?**



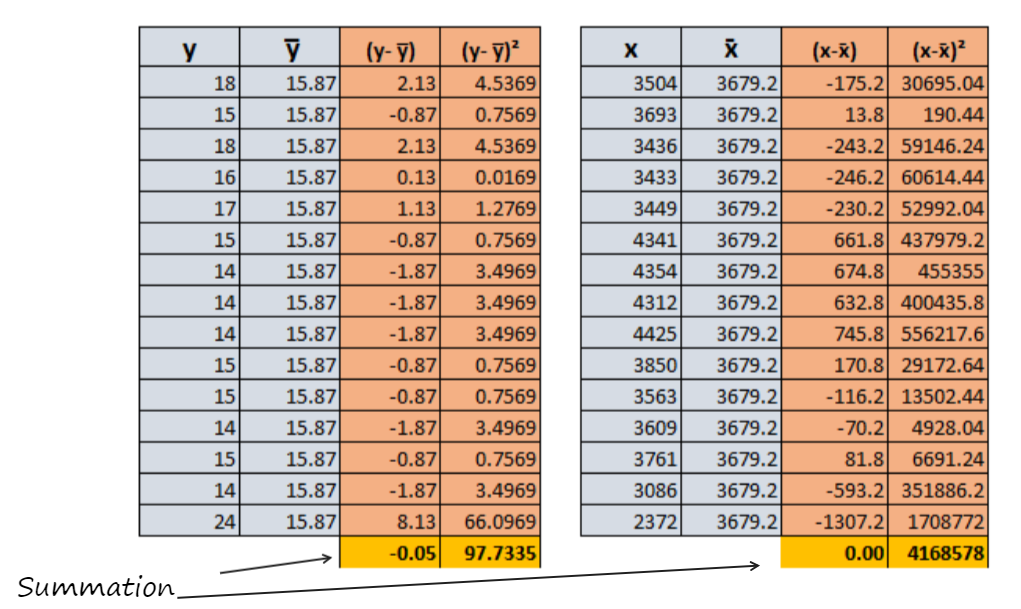
* Plot A is a zoomed-out version of Plot B – this is a visual paradox. We cannot rely on visual inspection to quantify the strength of relationships.
* We need a measure to quantify the strength of relationship. Correlation is the most common measure of linear association. (Remember, we are still dealing only with linear relationships)
* By association we mean the strength (and direction) of a linear relationship between two numerical variables.
* The Relationship is strong if the points in a scatterplot cluster tightly around some straight line.
* If this line rises from left to right, then the relationship is positive.
* If it falls from left to right, then the relationship is negative.

**Calculating these Measures:**



* For each point on the scatter cloud, we are going to keep a score.
* If the points are on the 1st and the 3rd quadrant then it contributes positively to the score.
* If the points are on the 2nd and the 4th quadrant then it contributes negatively to the score.
* We keep doing this and then average it out.

**Calculating Variance and Standard Deviation**



**Variance:**

* **Definition:** Variance is a measure of how much each number in a set differs from the mean (average) of the set.

**Simple Explanation:** Imagine you have a set of numbers. Variance tells you how spread out or clustered those numbers are around the average. If the numbers are close to the average, the variance is low. If they are spread out, the variance is high.

**Standard Deviation:**

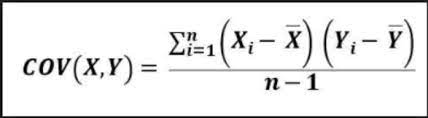
* **Definition:** Standard deviation is the square root of the variance. It provides a more interpretable measure of the spread of numbers in a dataset.
* **Simple Explanation:** Standard deviation is like a "typical" or "average" distance of numbers from the mean. If the standard deviation is small, it means the numbers are close to the average. If it's large, the numbers are more spread out.

A diagram of a mathematical equation

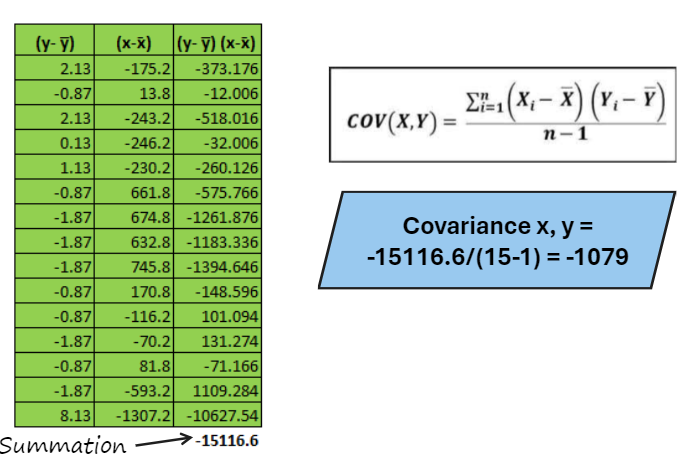
Description automatically generated

**Covariance**

* Covariance is a statistical measure that describes the degree to which two variables change together.
* It indicates whether an increase in one variable is associated with an increase or decrease in another variable.
* Covariance can take positive or negative values, and its magnitude represents the strength of the relationship between the variables.



**In Our mpg vs weight example, the covariance can be calculated as under:**



**Covariance can be interpreted as follows:**

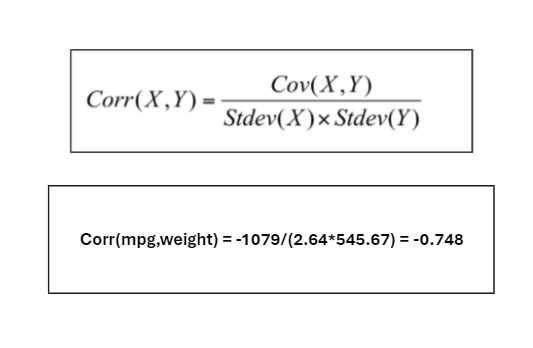
* Positive covariance indicates that the two variables tend to increase or decrease together.
* Negative covariance indicates that as one variable increases, the other tends to decrease.
* Zero covariance suggests no linear relationship between the variables.

We now have covariance (which is -1079). The larger this number i.e. more positive the association is that much stronger. The smaller this number (i.e. to the negative side) the weaker is the association. In mpg, weight case we have negative relationship.

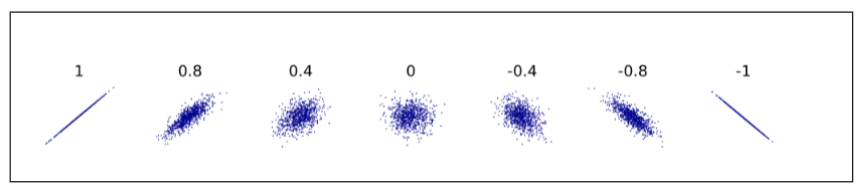
While covariance provides insights into the direction of the relationship, its value is not standardized. It a challenge to compare covariance from two different datasets and make any meaningful comparison.

**To overcome this limitation of covariance we use another measure called the correlation coefficient.**

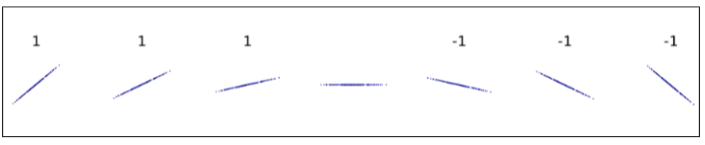
**Correlation** is unit free. It ranges between -1 and 1. We don’t need a scale or units to interpret. Its by far the most favourite measure of association. Keep in mind that correlation is a measure of linear relationship.



* So, in case of mileage and weight of the car, we can see a strong negative relationship.
* However, we can’t interpret strong correlation as causation. Meaning we can’t say for sure that the high mpg of car is because of weight. All we can say is as per data if mpg is high the weight could be low.
* Both correlation and Covariance are measures of Linear association.



* If Correlation is zero, it only means that there is no linear association.



* The correlation is 1 irrespective on the slope.
* If the correlation of 0 only means, there is no linear relationship. Correlation can be misleading when the association is non-linear.

A close-up of a smiley face

Description automatically generated

* In the above example, the correlation in each of the case is o, but we can clear see a pattern.
* However, this pattern cannot be modelled using a Linear modelling assumption.

Till this point

* We have found that the mpg and weight are negatively related.
* We also know that the relationship which -0.74 is reasonably strong.

However, this information is not enough to make predictions.

**Let’s say we are asked if the car’s weight is 4000 kilograms, what would we expect the mpg to be?**

We need a model to answer this question. Remember Model is a mathematical representation of the pattern. Here our model would be line. We need to find the line that best fits the data.

# **3. Linear Regression – Finding the line of best fit**

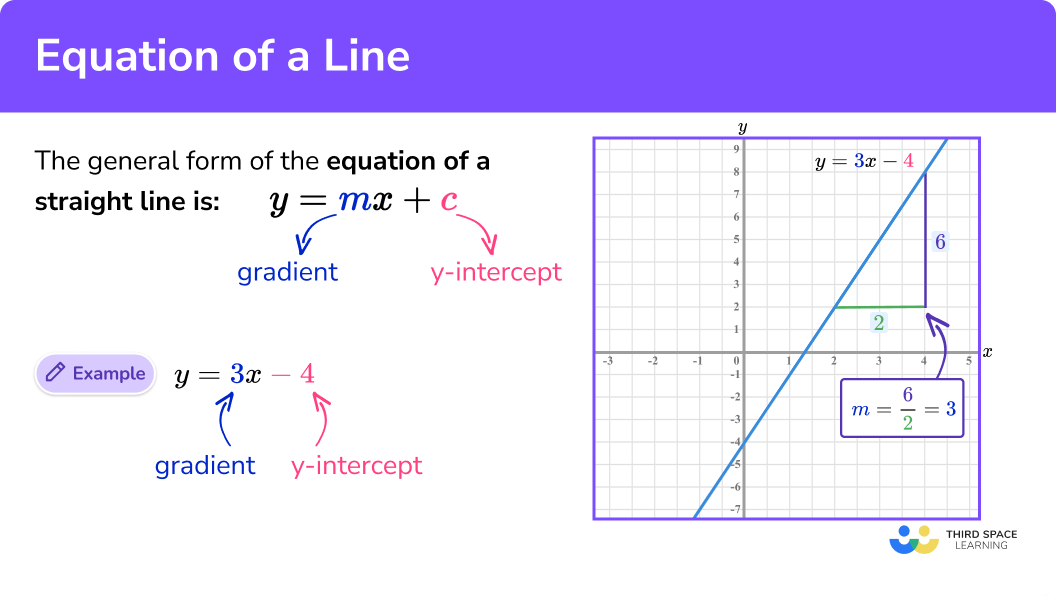
Correlation tells us about the strength of relationship between the dependent variable and the independent variable, we still need a model that will help us predict the dependent variable given the independent variable.

A model is a equation. It is an abstraction of reality.

A graph with colored lines and numbers

Description automatically generated

* Curves have more parameters than a line.
* In fact, the most popular curve is itself a line, it popular because its simple.
* Straight line will always have an equation like: y *= ax + b*



* The Procedure of fitting a line to a scatter of points is called as **Linear Regression**.
* We go with a line and not a curve because sometime coming with all the parameters of curve is error prone. Parameter estimation of line is more reliable to work with.

## **Now, how do we figure out the line that best fits the data.**

|  |  |
| --- | --- |
|  | * There could be infinite number of lines that can be drawn this 2-dimensional mathematical space or Euclidean space or cartesian plane. * Because each line can be defined as * y = mx + c. Each line has its own conclusion for the predictions. * Question is what is the best line? |

A diagram of a graph

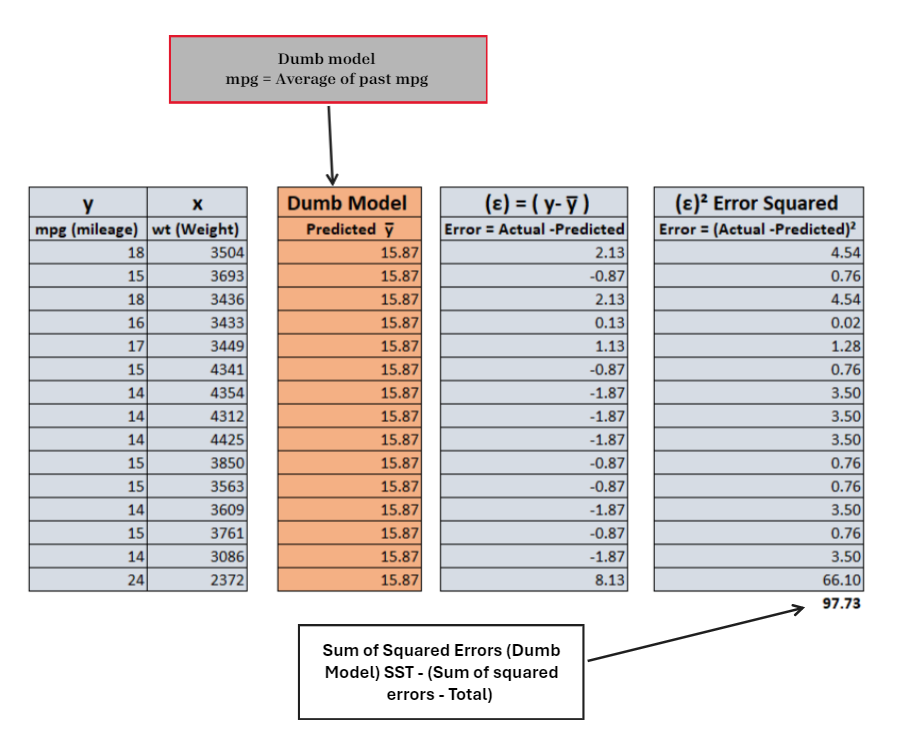
Description automatically generated

* The best line is the line with smallest summation error squared or the least sum of squared errors. It’s also represented by (ε)².
* Choosing a line that (in some sense) minimizes the vertical distance from the point to the line.
* We also choose to minimize the sum of squares of this vertical distance, as it is mathematically convenient.
* The method is called the “Least Squared estimation”.
* Looks like finding the best line is laborious, iterative process and that’s why computer does it for us.
* Computer does not really write each line to find the best fit. It actually writes the Sum of Squared error (ε)² in terms of y = mx + c
* We differentiate Sum of Squared error (ε)² and equate it to 0 by finding the best values of `m` and `c` in the line y = mx + c.

|  |  |
| --- | --- |
| * Turns out our best line looks like:   **Mpg = 0.0036 \* weight + 29.20**   * Now we can predict what would be the mileage for a car weighing 4000 kg. * Turns out it is 14.8 miles per gallon. | A graph with blue dots and numbers  Description automatically generated |

### **Measures of Regression fit – How good is our line as model over our Dumb Model**

* Now we have discussed 2 Models.
  + Average as a model (dumb – or no source of intelligence) model.
  + Linear Regression model.
* By definition dumb (un-intelligent) model should underperform our Regression model (which derives its intelligence from the variable weight of car).
* How do we compare performance of both the models – Answer is to look at Sum of squared errors in each case.

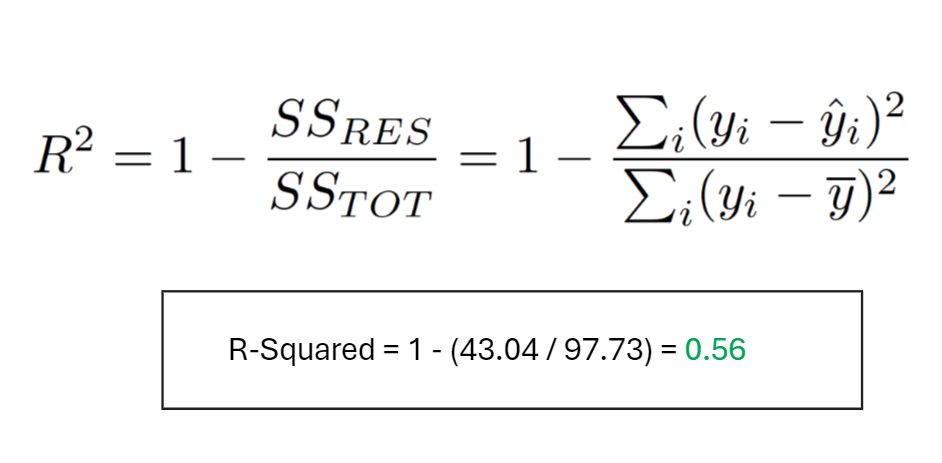


* The Sum of Square Errors (also called Sum of Square Error Total) SST for the dumb model is at 97.73.

A graph of a graph of a model

Description automatically generated with medium confidence

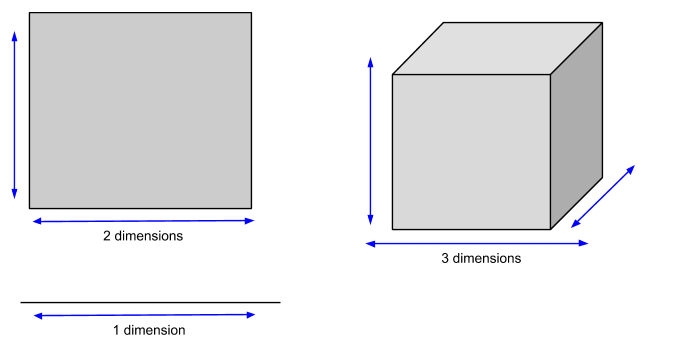
* The Sum of Square Error (also called Sum of Squared Errors Regression) SSR is at 43.04.
* With the regression model the error has reduced. It still has errors, but its less than the dumb model.
* However, we still need a neat measure of regression fit. And that measure is called a Coefficient of Determination R2 – R Squared.



* R-squared lends itself a really nice interpretation.
* It is the percentage of variation of the dependent variable (mpg) explained by the regression.
* R-squared has no units and lies between 0 and 1.
* R-squared basically, is the lift you got from your regression model over your dumb model.
* It also turns out the R-squared is the same as the Correlation squared. (check this we both the numbers now – correlation coefficient which was -0.74 and R-Squared which is 0.56)

## **Multiple Regression**

* In the case of Simple Linear Regression, we had one independent variable (weight in our previous example),
* However, we can have more than one independent variable.
* Because we were having one dependent variable and one independent variable, we were looking at 2 dimensions.
* In 2 dimensions the equation is a straight line, in 3 dimensions it’s a plane and in more than 4 dimensions it’s a hyperplane (we cannot visualize this).



* The specifications of Multiple linear regression would look like

A math equations on a white background

Description automatically generated