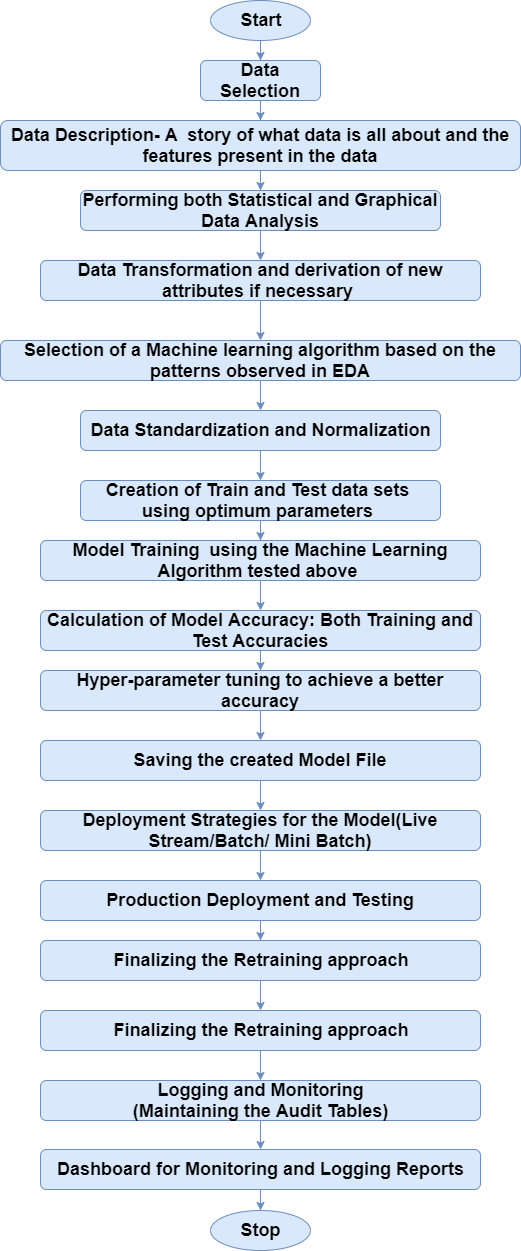
**Linear Regression**

Linear Regression is one of the most fundamental algorithms in the Machine Learning world. It is the door to the magical world ahead. But before proceeding with the algorithm, let’s first discuss the lifecycle of any machine learning model. This diagram explains the creation of a Machine Learning model from scratch and then taking the same model further with hyperparameter tuning to increase its accuracy, deciding the deployment strategies for that model and once deployed setting up the logging and monitoring frameworks to generate reports and dashboards based on the client requirements. A typical lifecycle diagram for a machine learning model looks like:



## What is Regression Analysis?

Regression in statistics is the process of predicting a Label(or Dependent Variable) based on the features(Independent Variables) at hand. Regression is used for time series modelling and finding the causal effect relationship between the variables and forecasting. For example, the relationship between the stock prices of the company and various factors like customer reputation and company annual performance etc. can be studied using regression.

Regression analysis is an important tool for analysing and modelling data. Here, we fit a curve/line to the data points, in such a manner that the differences between the distance of the actual data points from the plotted curve/line is minimum. The topic will be explained in detail in the coming sections.

## The use of Regression

Regression analyses the relationship between two or more features. Let’s take an example:

Let’s suppose we want to make an application which predicts the chances of admission a student to a foreign university. In that case, the

The benefits of using Regression analysis are as follows:

* It shows the significant relationships between the Lable (dependent variable) and the features(independent variable).
* It shows the extent of the impact of multiple independent variables on the dependent variable.
* It can also measure these effects even if the variables are on a different scale.

These features enable the data scientists to find the best set of independent variables for predictions.

## Linear Regression

Linear Regression is one of the most fundamental and widely known Machine Learning Algorithms which people start with. Building blocks of a Linear Regression Model are:

* Discreet/continuous independent variables
* A best-fit regression line
* Continuous dependent variable. i.e., A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

Y=a+b\*X + e

Where,

* a is the intercept,
* b is the slope of the line,
* and e is the error term. The equation above is used to predict the value of the target variable based on the given predictor variable(s).

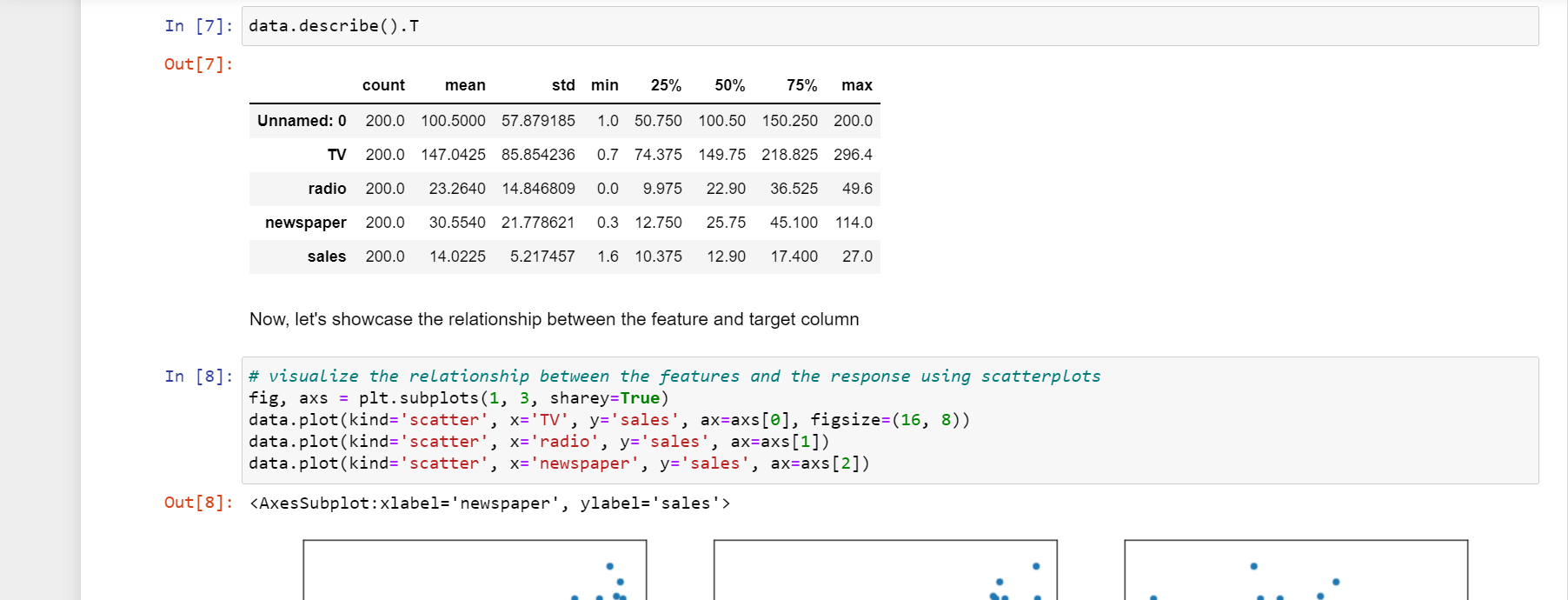
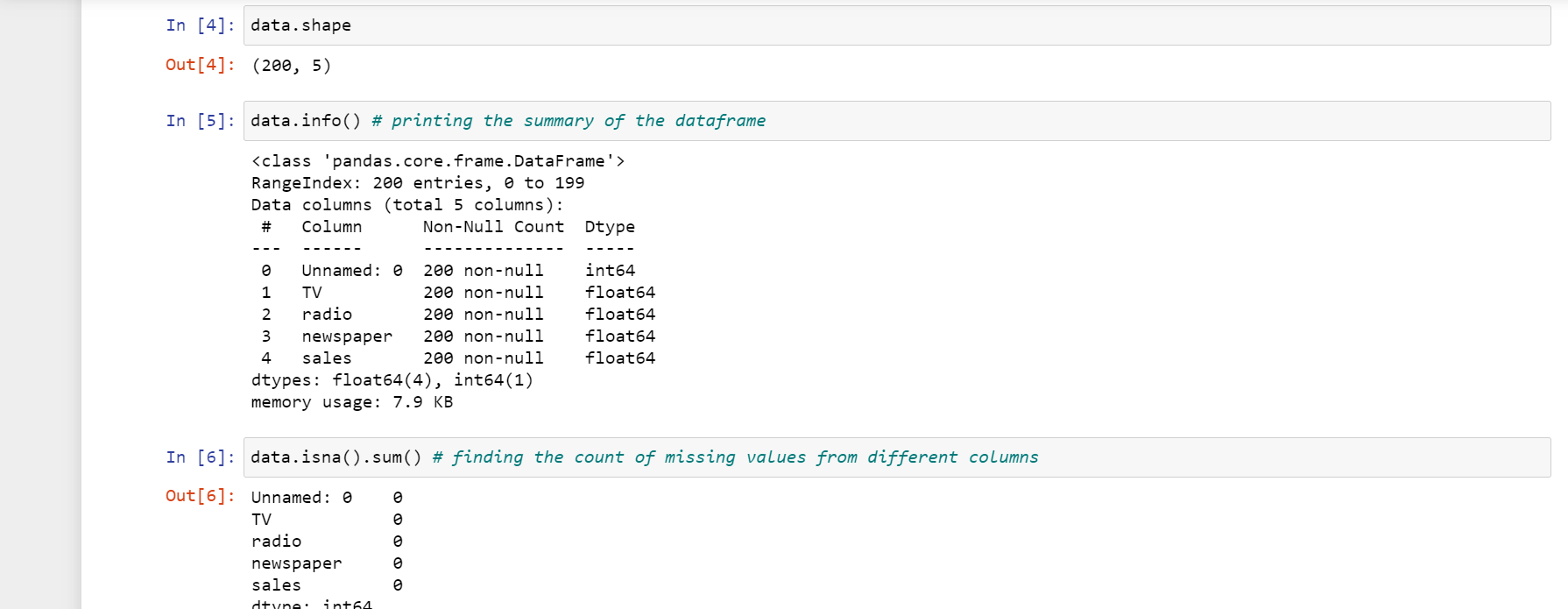


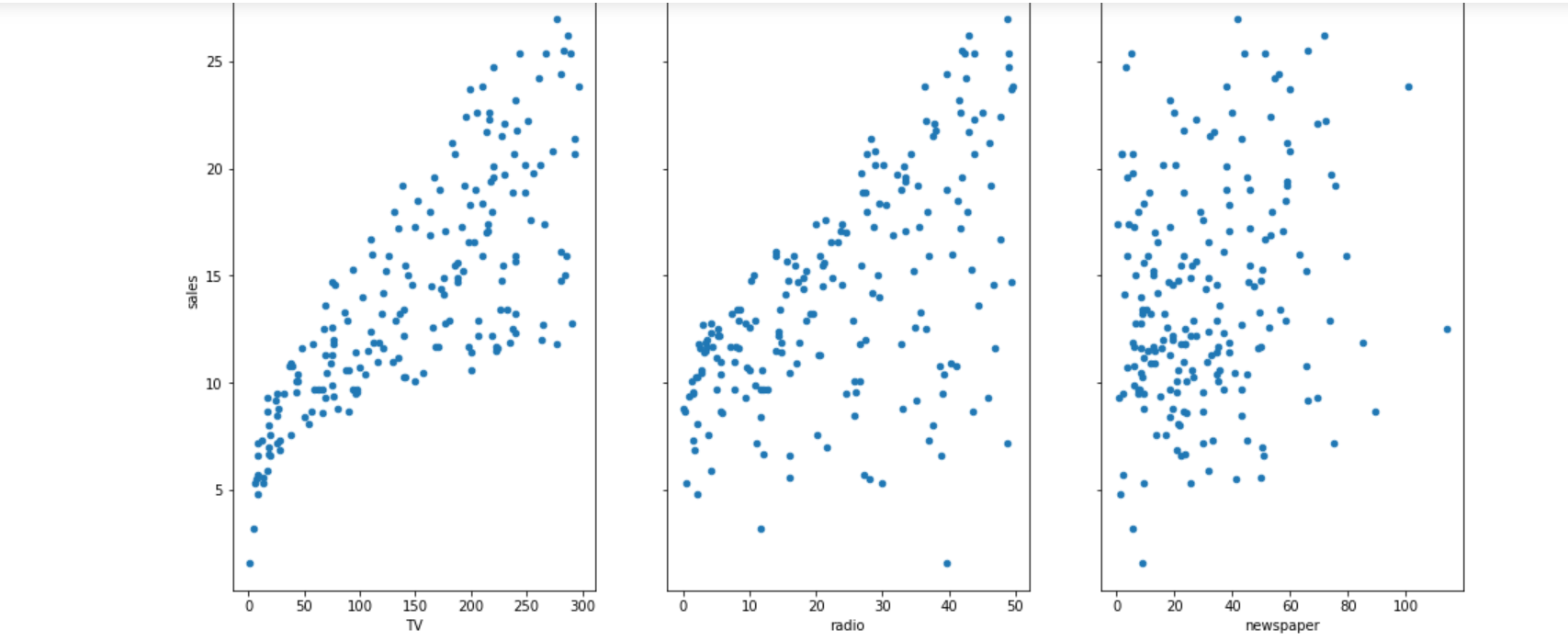
What are the **features**?

* TV: Advertising dollars spent on TV for a single product in a given market (in thousands of dollars)
* Radio: Advertising dollars spent on Radio
* Newspaper: Advertising dollars spent on Newspaper

What is the **response**?

* Sales: sales of a single product in a given market (in thousands of widgets)





## Questions about the data

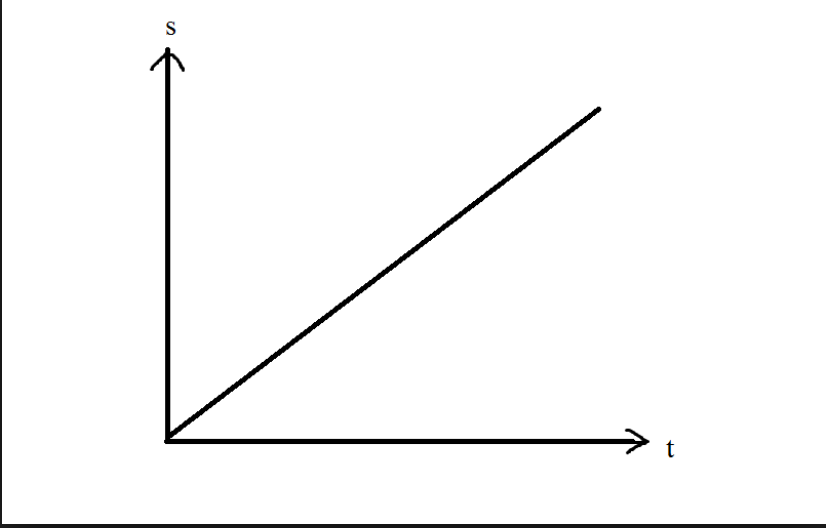
A generic question shall be: How the company should optimise the spends on advertising to maximise the sales?

These general questions might lead you to more specific questions:

1. What’s the relationship between ads and sales?
2. How prominent is that relationship?
3. Which ad types contribute to sales?
4. How each ad contributes to sales?
5. Can sales be predicted based on the expense of the advertisement?

We will explore these questions below!

From the relationship diagrams above, it can be observed that there seems to be a linear relationship between the features TV ad, Radio ad and the sales is almost a linear one. A linear relationship typically looks like:



## Simple Linear Regression

Simple Linear regression is a method for predicting a **quantitative response** using a **single feature** ("input variable"). The mathematical equation is:

𝑦=𝛽0+𝛽1𝑥y=β0+β1x

What do terms represent?

* 𝑦y is the response or the target variable
* 𝑥x is the feature
* 𝛽1β1 is the coefficient of x
* 𝛽0β0 is the intercept

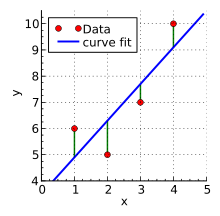
𝛽0β0 and 𝛽1β1 are the **model coefficients**. To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the Sales!

#### Estimating ("Learning") Model Coefficients

The coefficients are estimated using the **least-squares criterion**, i.e., the best fit line has to be calculated that minimizes the **sum of squared residuals** (or "sum of squared errors").

### The mathematics involved

Take a quick look at the plot created. Now consider each point, and know that each of them has a coordinate in the form (X, Y). Now draw an imaginary line between each point and the current "best-fit" line. We'll call the distance between each point and the current best-fit line as D. To get a quick image of what we're trying to visualize, take a look at the picture below:



What elements are present in the diagram?

* The red points are the **observed values** of x and y.
* The blue line is the **least squares line**.
* The green lines are the **residuals**, which is the distance between the observed values and the least squares line.

Before, we're labelling each green line as having a distance D, and each red point as having a coordinate of (X, Y). Then we can define our best fit line as the line having the property were:

𝐷21+𝐷22+𝐷23+𝐷24+....+𝐷2𝑁D12+D22+D32+D42+....+DN2

So how do we find this line? The least-square line approximating the set of points:

(𝑋,𝑌)1,(𝑋,𝑌)2,(𝑋,𝑌)3,(𝑋,𝑌)4,(𝑋,𝑌)5,(X,Y)1,(X,Y)2,(X,Y)3,(X,Y)4,(X,Y)5,

has the equation:

𝑌=𝑎0+𝑎1𝑋Y=a0+a1X

this is basically just a rewritten form of the standard equation for a line:

𝑌=𝑚𝑥+𝑏Y=mx+b

We can solve for these constants a0 and a1 by simultaneously solving these equations:

Σ𝑌=𝑎0𝑁+𝑎1Σ𝑋ΣY=a0N+a1ΣX

Σ𝑋𝑌=𝑎0Σ𝑋+𝑎1Σ𝑋2ΣXY=a0ΣX+a1ΣX2

These are called the normal equations for the least-squares line. There are further steps that can be taken in rearranging these equations to solve for y, but we'll let scikit-learn do the rest of the heavy lifting here.

Let’s see the underlying assumptions: -

* The regression model is linear in terms of coefficients and error term.
* The mean of the residuals is zero.
* The error terms are not correlated with each other, i.e. given an error value; we cannot predict the next error value.
* The independent variables(x) are uncorrelated with the residual term, also termed as **exogeneity**. This, in layman term, generalises that in no way should the error term be predicted given the value of independent variables.
* The error terms have a constant variance, i.e. **homoscedasticity**.
* No Multicollinearity, i.e. no independent variables should be correlated with each other or affect one another. If there is multicollinearity, the precision of prediction by the OLS model decreases.
* The error terms are normally distributed.

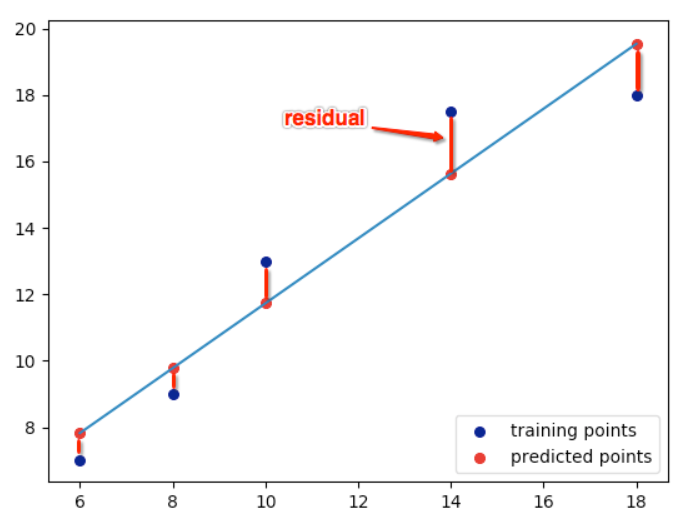
The general equation of a straight line is:

𝑦=𝑚𝑥+𝑏𝑦

It means that if we have the value of m and b, we can predict all the values of y for corresponding x. During construction of a Linear Regression Model, the computer tries to calculate the values of m and b to get a straight line. But the question is:

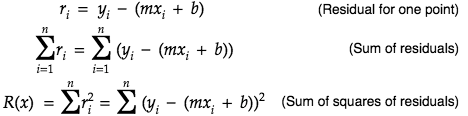
###### *How Do you Know this is the best fit line?*

The best fit line is obtained by minimizing the residual. Residual is the distance between the actual Y and the predicted Y, as shown below:

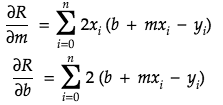


Mathematically, Residual is: r = y-(mx+b)

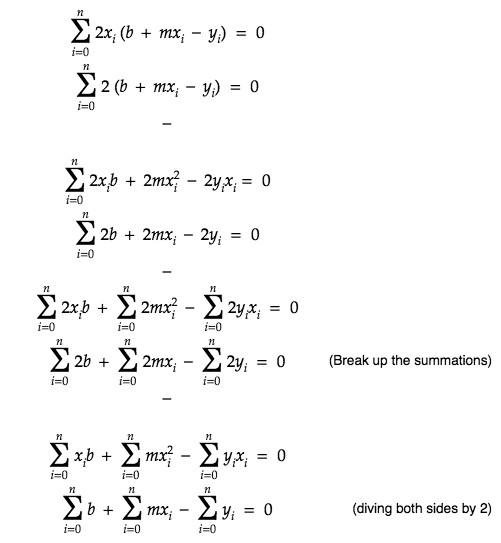
Hence, the sum of the square of residuals is:



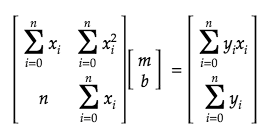
As we can see that the residual is both a function of m and b, so differentiating partially with respect to m and b will give us:



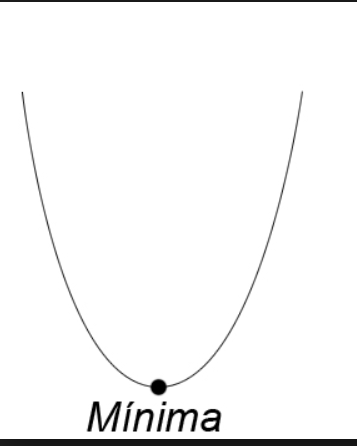
For getting the best fit line, residual should be minimum. The minima of a function occurs where the derivative=0. So, equating our corresponding derivatives to 0, we get:



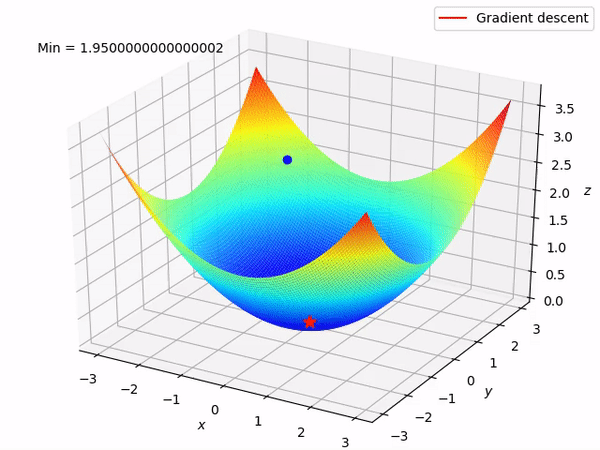
This same equation can be written in matrix form as:



Ideally, if we'd have an equation of one dependent and one independent variable the minima will look as follows:

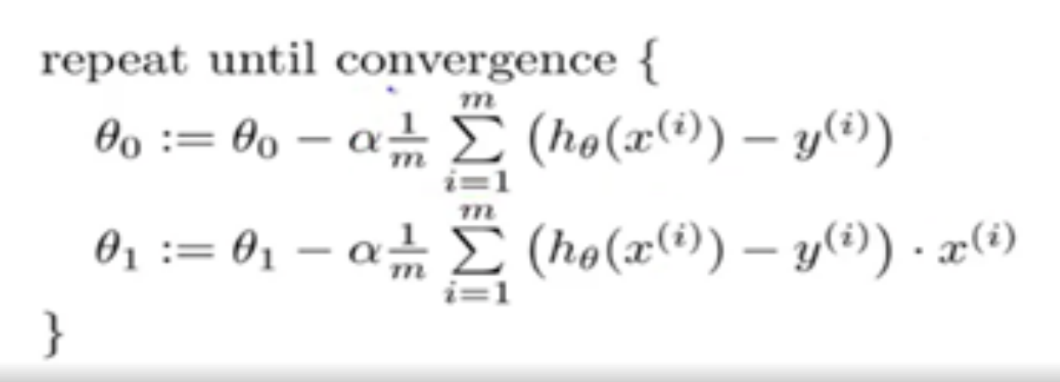


But as the residual's minima is dependent on two variables m and b, it becomes a \_Paraboloid\_ and the appropriate m and b are calculated using \_\*Gradient Descent\*\_ as shown below:



Now, let’s understand how to check, how well the model fits our data.

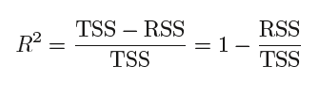
The new values for 'slope' and 'intercept' are caluclated as follows:



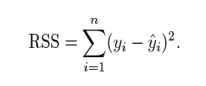
where, 𝜃0θ0 is 'intercept' , 𝜃1θ1 is the slope, 𝛼α is the learning rate, m is the total number of observations and the term after the ∑∑ sign is the loss. Google Tensor board recommends a Learning rate between 0.00001 and 10. Generally a smaller learning rate is recommended to avoid overshooting while creating a model.

### **R2 statistics**

The R-squared statistic provides a measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1. In simple words, it represents how much of our data is being explained by our model. For example, R2 statistic = 0.75, it says that our model fits 75 % of the total data set. Similarly, if it is 0, it means none of the data points is being explained and a value of 1 represents 100% data explanation. Mathematically R2 statistic is calculated as :



Where RSS: is the Residual Sum of squares and is given as :



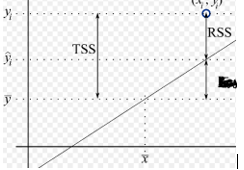
RSS is the residual(error) term we have been talking about so far.

And, TSS: is the Total sum of squares and given as :



TSS is calculated when we consider the line passing through the mean value of y, to be the best fit line.

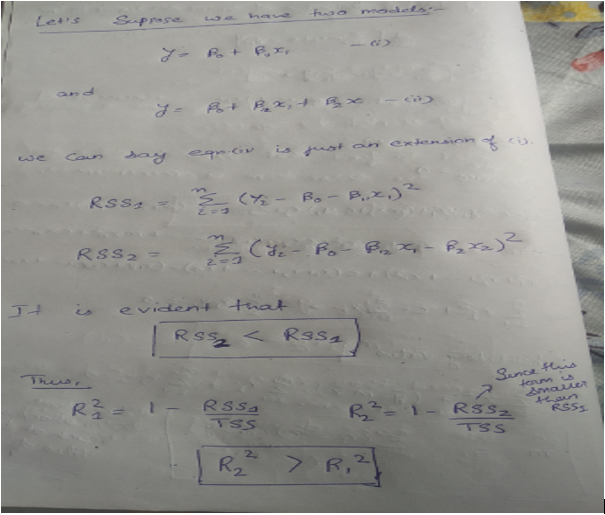
Just like RSS, we calculate the error term when the best fit line is the line passing through the mean value of y and we get the value of TSS.



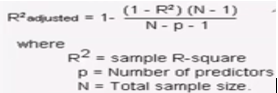
The closer the value of R2 is to 1 the better the model fits our data. If R2 comes below 0(which is a possibility) that means the model is so bad that it is performing even worse than the average best fit line.

### Adjusted **R2** statistics

As we increase the number of independent variables in our equation, the R2 increases as well. But that doesn’t mean that the new independent variables have any correlation with the output variable. In other words, even with the addition of new features in our model, it is not necessary that our model will yield better results but R2 value will increase. To rectify this problem, we use Adjusted R2 value which penalises excessive use of such features which do not correlate with the output data. Let’s understand this with an example:



We can see that R2 always increases with an increase in the number of independent variables. Thus, it doesn’t give a better picture and so we need Adjusted R2 value to keep this in check. Mathematically, it is calculated as:



In the equation above, when p = 0, we can see that adjusted R2 becomes equal to R2. Thus, adjusted R2 will always be less than or equal to R2, and it penalises the excess of independent variables which do not affect the dependent variable.

