**Packages to be known for Machine learning**

1. Numpy
2. Pandas
3. Scipy
4. Scikit-learn
5. Matplotlib
6. Seaborn

Scikit-learn :

* Lots of machine learning libraries.
* In built classification, Regression and clustering algorithms
* Build with flexibility to use Numpy and Scipy which helps easy to play with data.
* Most preprocessing tasks are inbuilt in scikit-learn such as Data preprocessing, Future Selection, extraction and train – test splitting, algorithm, prediction, evaluation and exploring the model.

**Supervised vs Unsupervised learning:**

**Supervised learning:**

Train the model with labelled dataset.

1. Classification – process of predicting discrete class label or category.

Eg. Predicting the class such as finding the cancer in cells of human, customer churn in telecom.

1. Regression – process of predicting a continuous value.

Eg. Predicting the house price, stock price.

**Unsupervised learning:**

We do not supervise the model, but the model trains on its own and discover information that are invisible to human eye.

1. Dimension reduction – reducing the redundant future to make classification easier.
2. Density estimation – explore data to find substructure
3. Market basket analysis or Association – based on theory of if a customer buys certain product, then he is likely to buy another product as well.
4. Clustering – grouping of data points that are similar somehow. Eg. Segment customer based on credit score in banking. Used for Discovering structure, summarization and Anomaly detection.

**Few more Machine learning techniques:**

Anomaly detection – discovering abnormal cases eg. Credit card fraud detection

Sequence mining – predict the next occurrence eg. Click stream

Recommender systems – Recommending items eg Netflix, youtube.

**Regression**

Process of predicting price for continuous value. It deals with two variables.

X: Independent variable which can be one or more. Explanatory variable which are the causes of the states.

Y: Dependent variable {State, target or final goal)

**Note: Dependent variable should be continuous and cannot be discreate value.**

**Independent variable can be numeric or categorical variable.**

1. **Simple Regression:**

One independent variable is used to estimate the dependent variable.

Eg. Car-size as independent variable and predicting Co2 emission for the new car make.

Depending on the relationship, it can be either linear or non-linear regression.

1. **Multiple Regression:**

More than one independent variable is used to predict the dependent variable.

Application of Regression:

1. Sales forecasting – total yearly sales prediction
2. Price estimation – house price
3. Employment income

**Simple Linear Regression:**

X: one Independent variable.

Y: Dependent variable {State, target or final goal)

**Y = Ɵ1 + Ɵ0 X**

**Ɵ1 , Ɵ0 =** parameters of the line.

**Ɵ1 =** Intercept of the line.

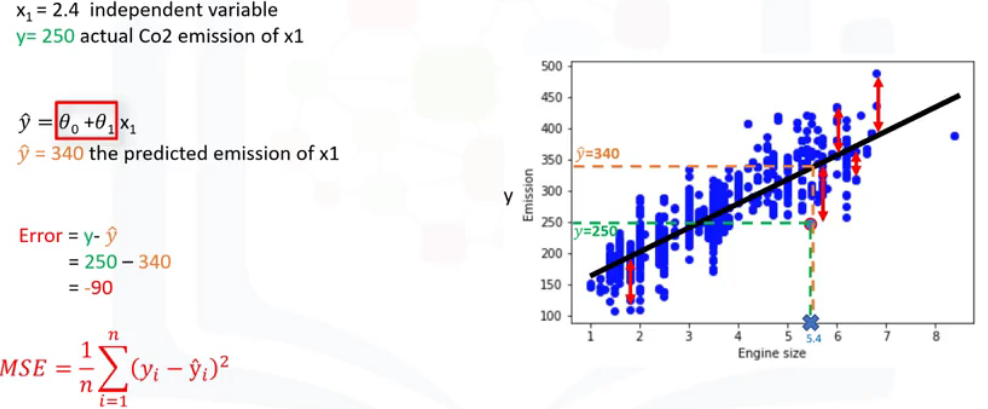
**Ɵ0 =slope** of the line.

Y^ = point after drawing linear line

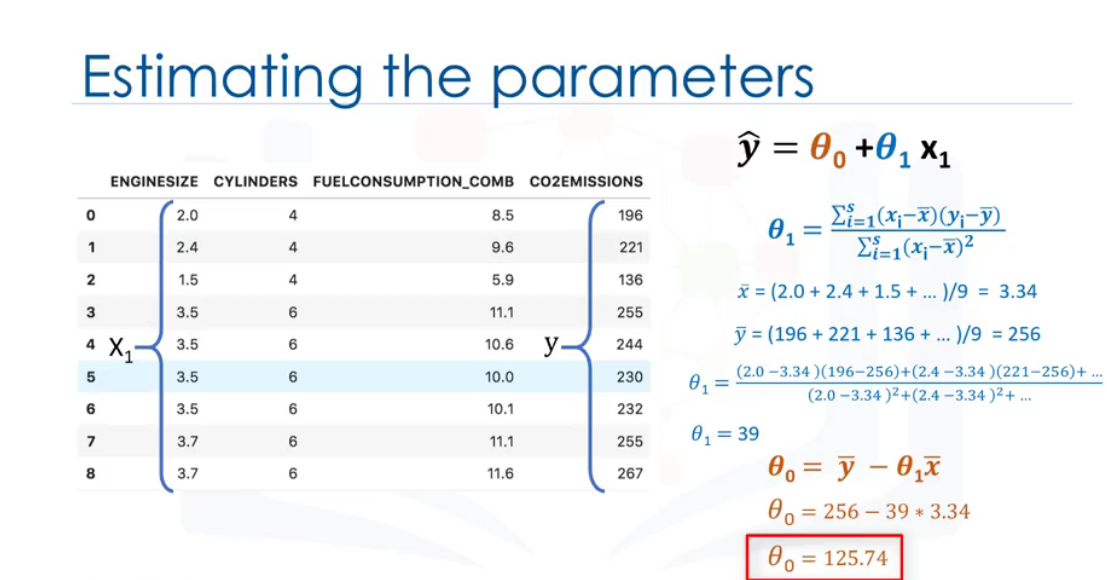
We can interpret as y is dependent of x.

We have to plot the x axis and y axis values in a scatterplot and try to fit a linear line to examine the accuracy of the linear line.

Error or residual= y – y^  
y^ is the predicted value



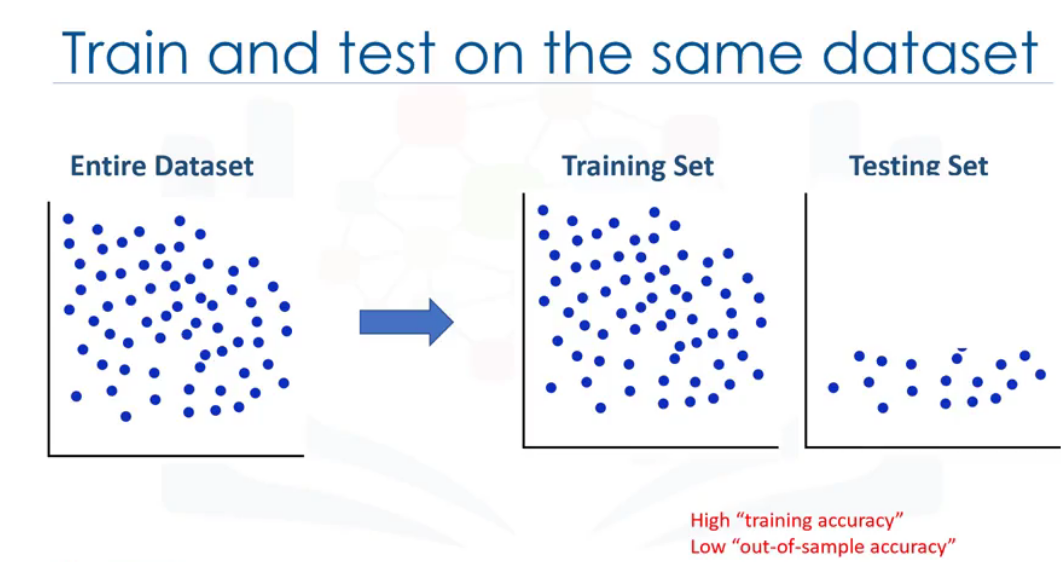
Objective of linear regression is to minimize the MSE (Mean square Error) by finding best parameters for **Ɵ1 , Ɵ0.**



**Pros:** Very fast, No parameter tuning, easy to understand and work with.

**Model Evaluation:**

1. **Train and test on same dataset**



This sample has high training accuracy and low out of sample accuracy.

Training accuracy:

High training accuracy isn’t a good which will result in overfitting.

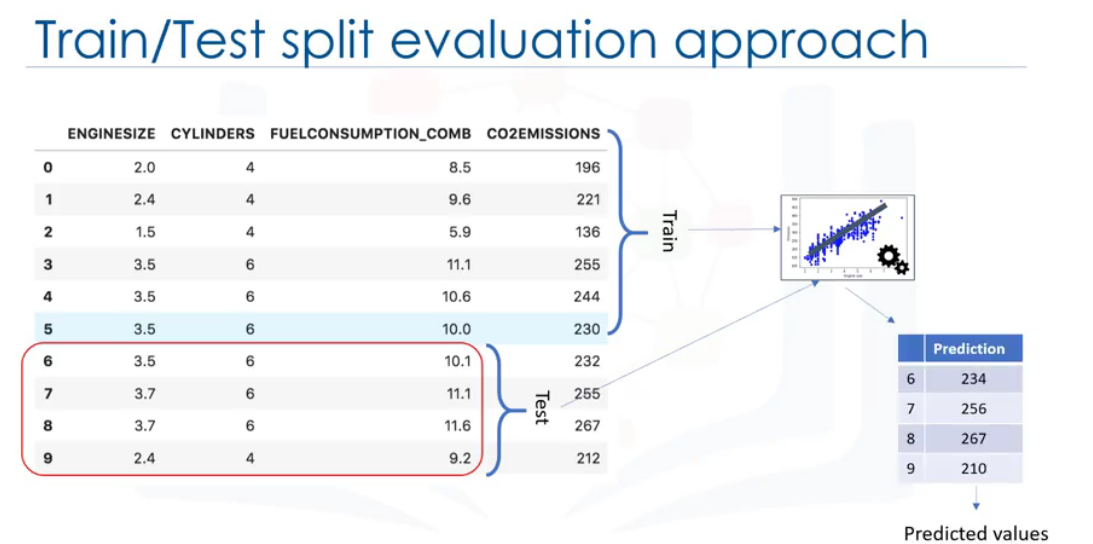
**Over fit**: Model is overly trained in the training dataset which may lead to noise and produce non-generalized model.

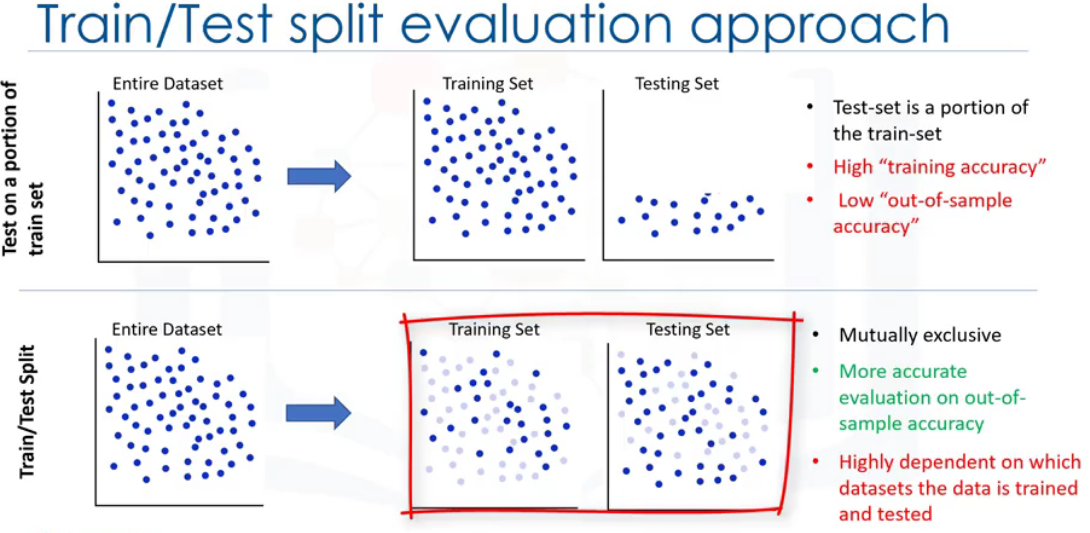
Out of sample accuracy:

This method has very less out of sample accuracy which is not good for a model.

Better approach is train-test split.

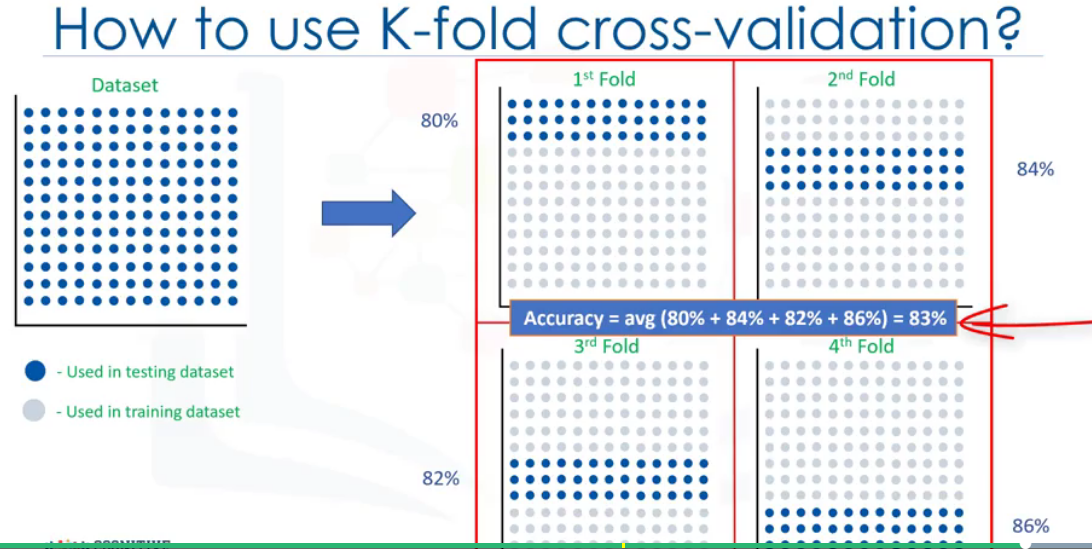
1. **Train and test split**





To overcome the disadvantage of train-test split which heavily dependent on dataset, we use k-fold cross validation approach.

K-fold approach performs multiple train test split with each split is different and finally the accuracy is averaged.

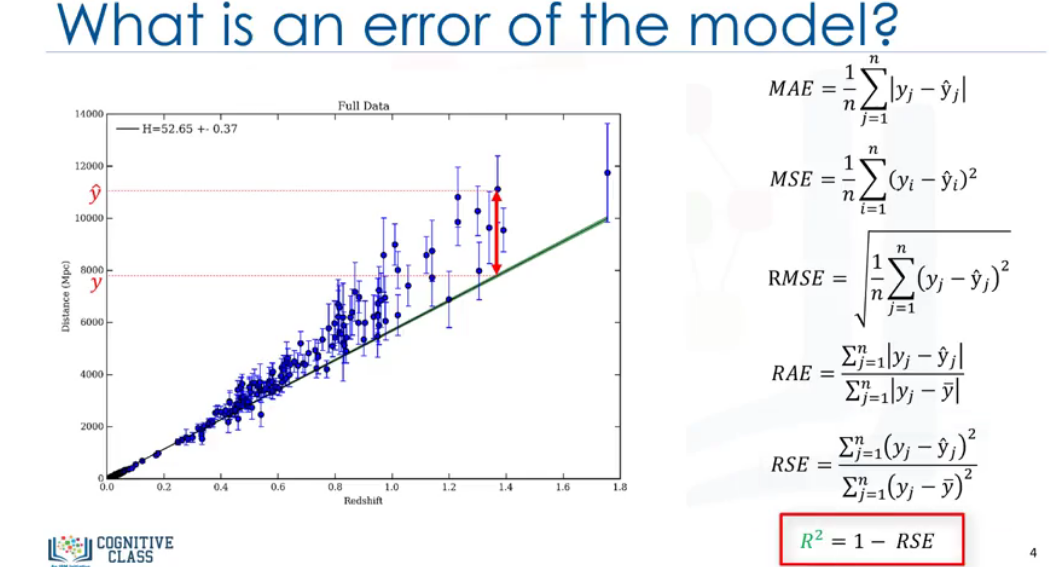


**Evaluation methods for the model**

1. **Mean Actual Error.**
2. **Mean Squared Error.**
3. **Root mean square Error.**
4. **Relative absolute Error or Residual sum of square.**
5. **Relative Squared Error (Used to calculate R2).**

R2 is not an error, it is metric for accuracy of model.

Higher the R2, better the model designed.



**Multiple Linear Regression:**

Process of predicting price for continuous value. It deals with two variables.

X: Independent variable which can be one or more.

Y: Dependent variable {State, target or final goal).

Equation: **Y = Ɵ0 + Ɵ1 X1 + Ɵ X2 +…..**

**Independent variable effectiveness on prediction.**

**Predicting impacting of the changes.**

**P – value:**

*P value* is a statistical measure that helps scientists determine whether their hypotheses are correct. P values are used to determine whether the results of their experiment are within the normal range of values for the events being observed. Usually, if the P value of a data set is below a certain pre-determined amount (like, for instance, 0.05), scientists will reject the "null hypothesis" of their experiment - in other words, they'll rule out the hypothesis that the variables of their experiment had *no* meaningful effect on the results. Today, p values are usually found on a reference table by first calculating a *chi square* value.

Steps to calculate P-value:

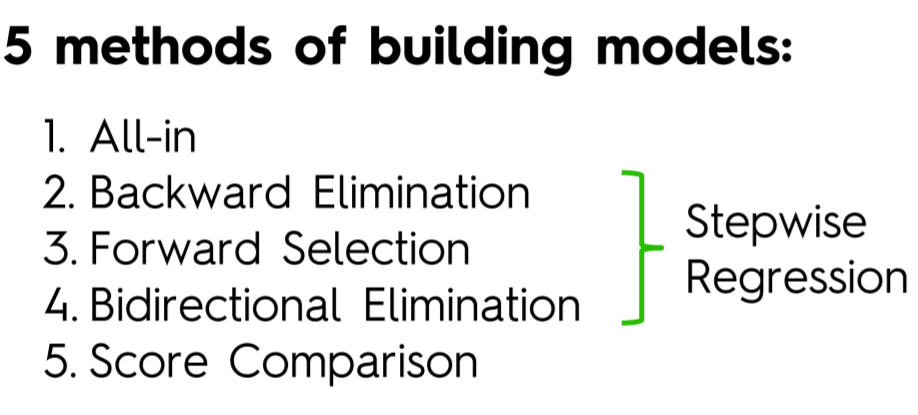
* Determine expected result (e)
* Determine the Observed result (o)
* Determine the degree of freedom d = n-1 (N= different category of observation)
* Calculate x^2 = sum((o – e) ^ 2 / e )
* Compare expected results to observed results with *chi square*.
* Choose the significance level. Significance levels are written as a decimal (such as 0.01), By convention, scientists usually set the significance value for their experiments at 0.05, or 5 percent
* Use a chi square distribution table to approximate your p-value.

**Link :** [**https://www.wikihow.com/Calculate-P-Value**](https://www.wikihow.com/Calculate-P-Value)

**Build a model:**

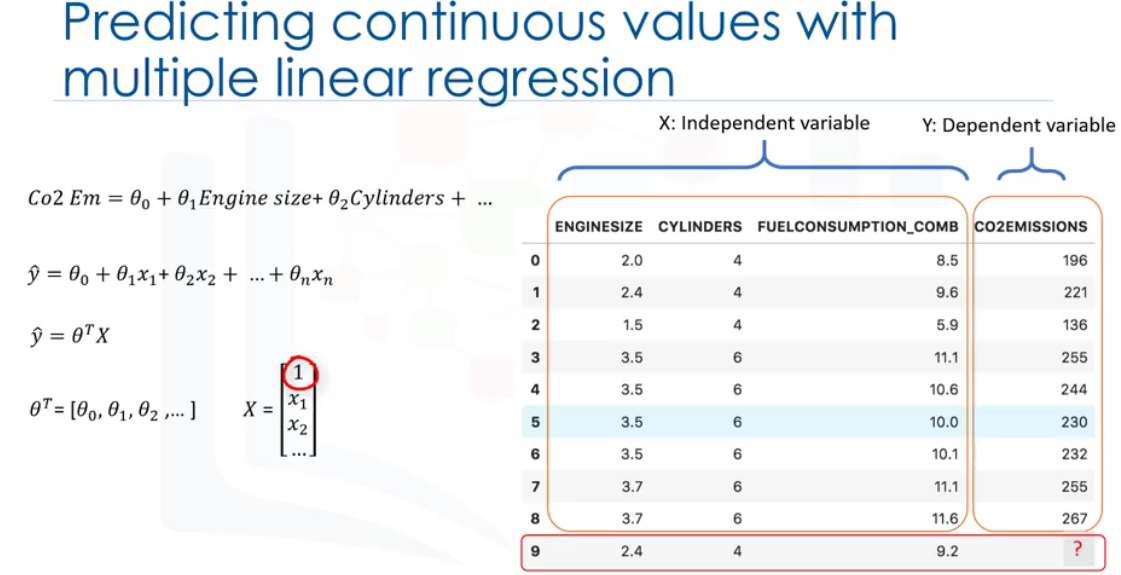
Use only relevant variable which will result in your prediction value i.e dependent variable

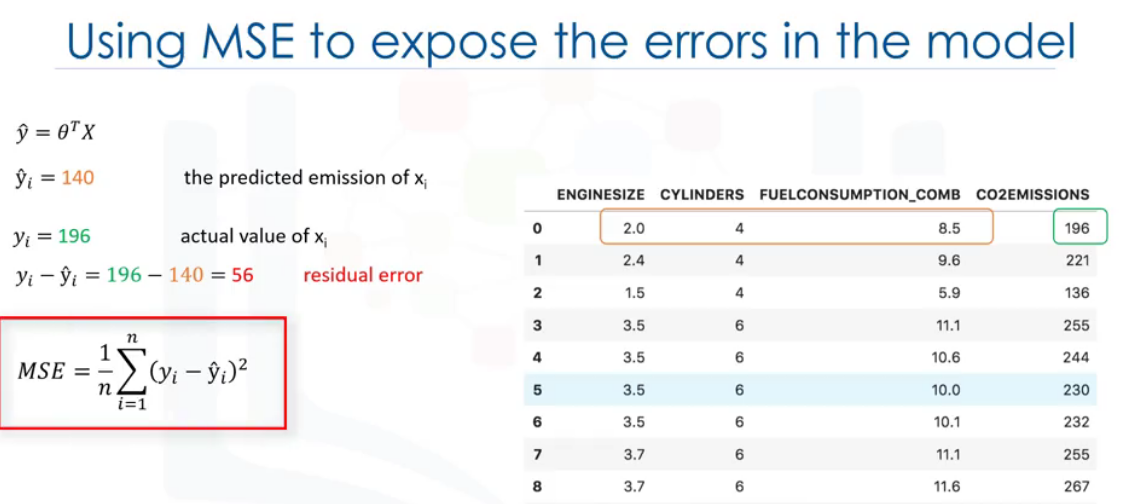
**Note: Use of multiple independent variable without any theoretical justification will result in Overfitting which is a bigger problem in building a model.**

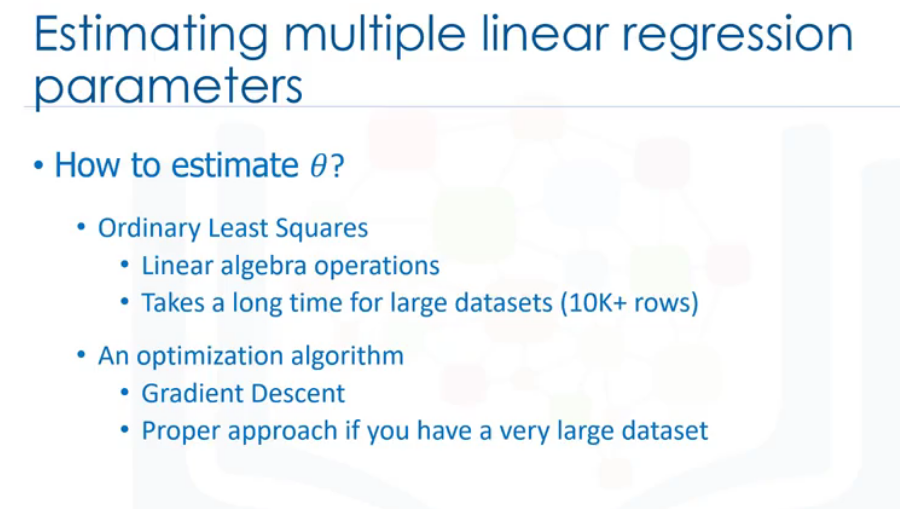


**Attachment below:**

****



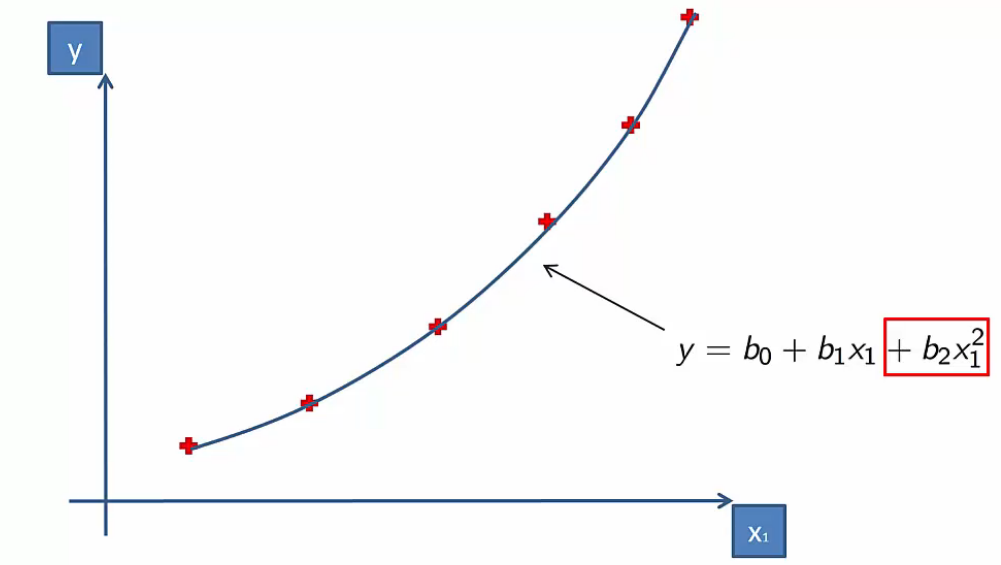




**Polynomial linear regression:**

Fit the observation in a scatter plot. If you find the scatter plot shows a curvy trend then you have do prediction using polynomial regression.

Equation: **Y = b0 + b1 X1 + b2 X1^2 +…..**



**Why it is called linear:**

Here, b0, b1, b2 determines the linearity i.e coefficient of x1. Thus, it’s a special case multiple linear regression.

# References

Aghabozorgi, S. (n.d.). *coursera.* Retrieved from www.coursera.org: https://www.coursera.org/learn/machine-learning-with-python