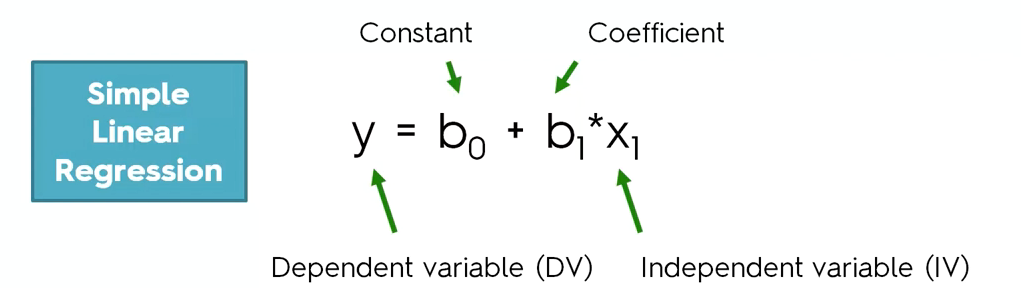
Regression

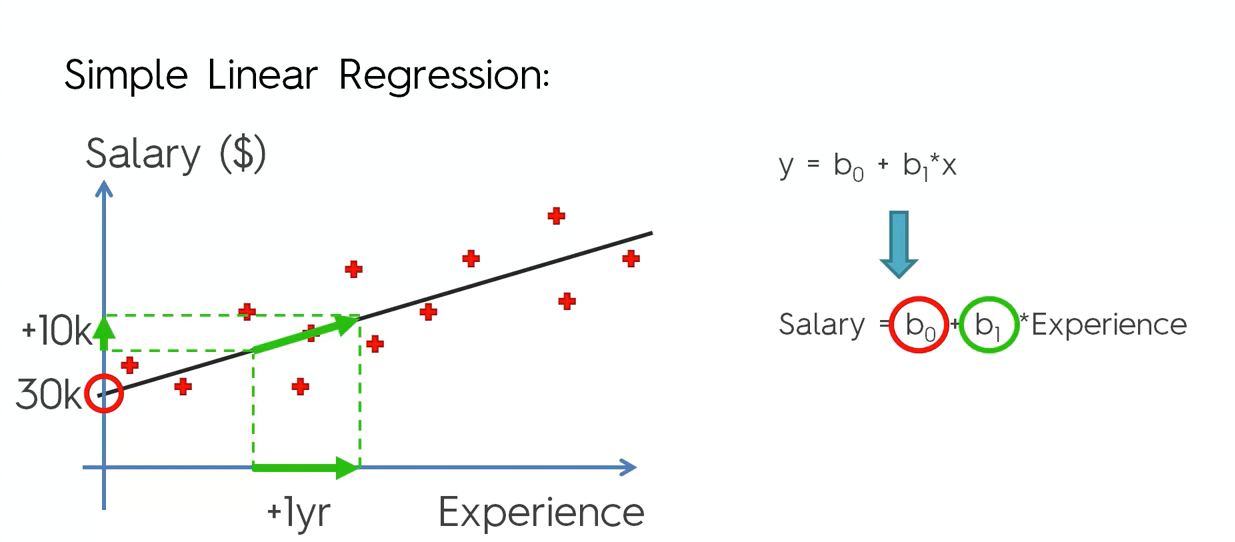


**Simple Linear Regression:**

It is like a formula of a slope

Eg scenario: how does a **salary** (Dependent) changes with **years of experience** (Independent Variable)

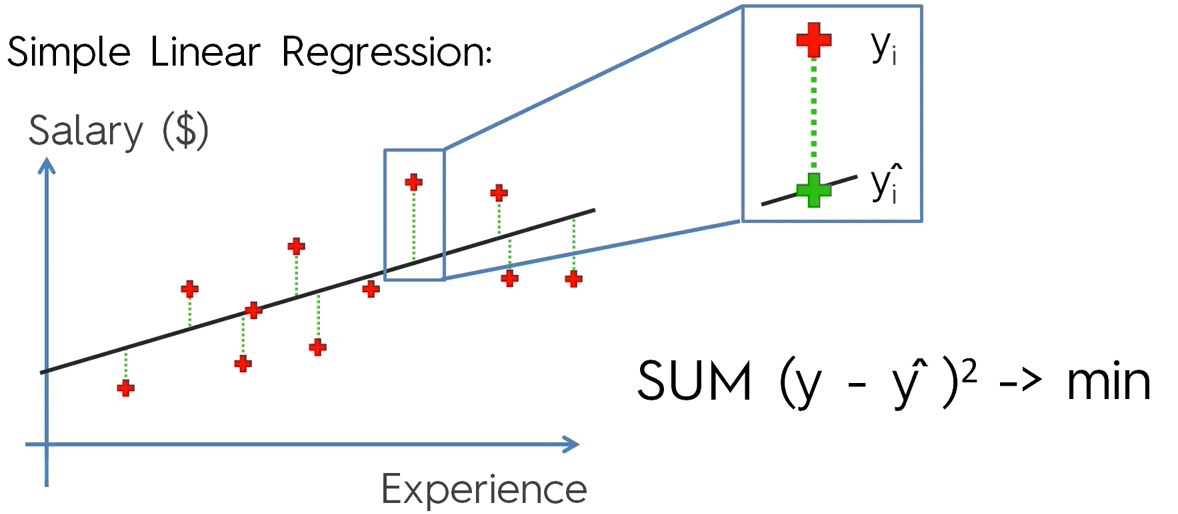
Coefficient🡪 Unit change in independent Variable (like multiplier🡪connection between x1 and y)



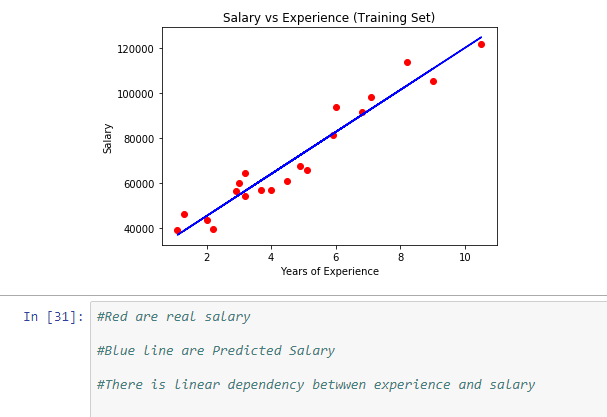
b0🡪Constant (30,000$)

b1🡪Coefficient (+10,000$ for 1 year)

1. How to find the best fitting line(line in the graph)



**Example**



Correlation🡪The correlation coefficient is a **statistical measure** that calculates the strength of the relationship between the relative movements of the two variables

It measures the strength (qualitatively) and direction of the **linear** relationship between two or more variables. The Pearson correlation coefficient measures the strength of the **linear** association between two variables

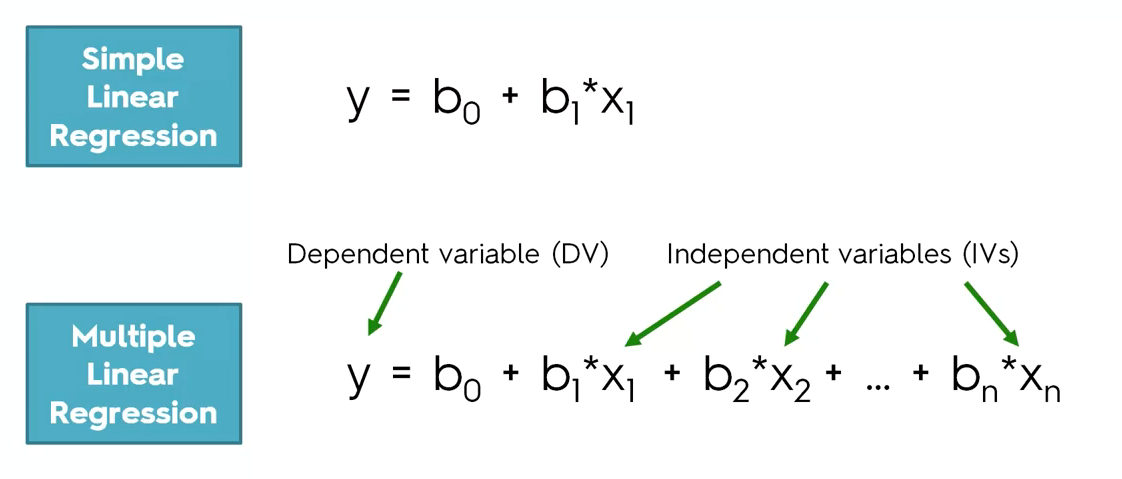
**DIFFERENCE BETWEEN SIMPLE AND LINEAR REGRESSION:**

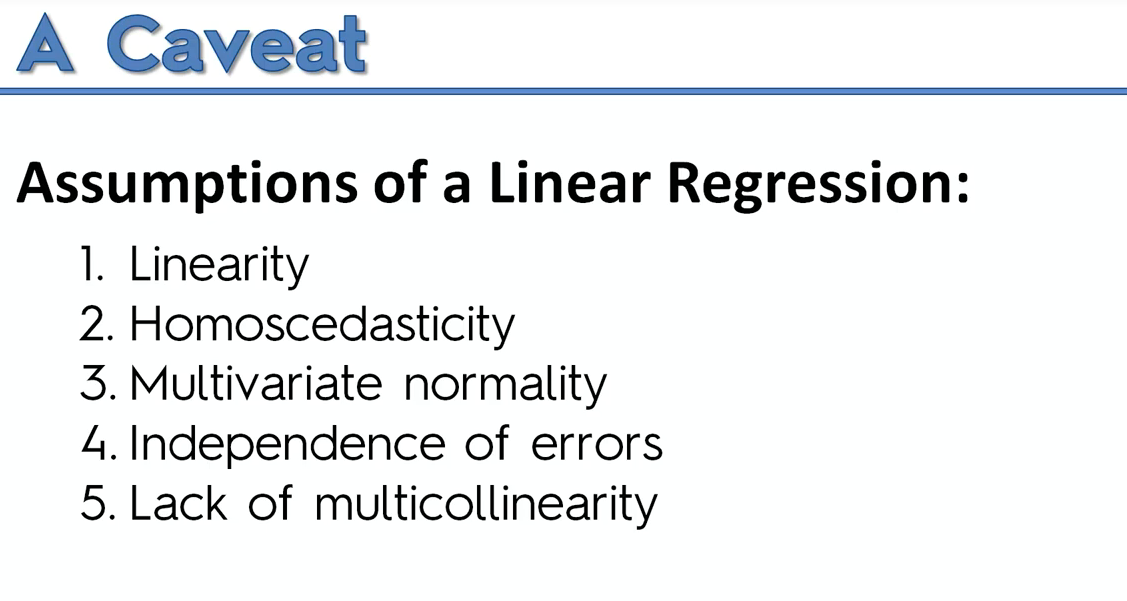
**Simple linear regression:**a single independent variable is used to predict the value of a dependent variable.

Equation: **y=A+BX**

**Multiple linear regression:** two or more independent variables are used to predict the value of a dependent variable. The difference between the two is the number of independent variables.

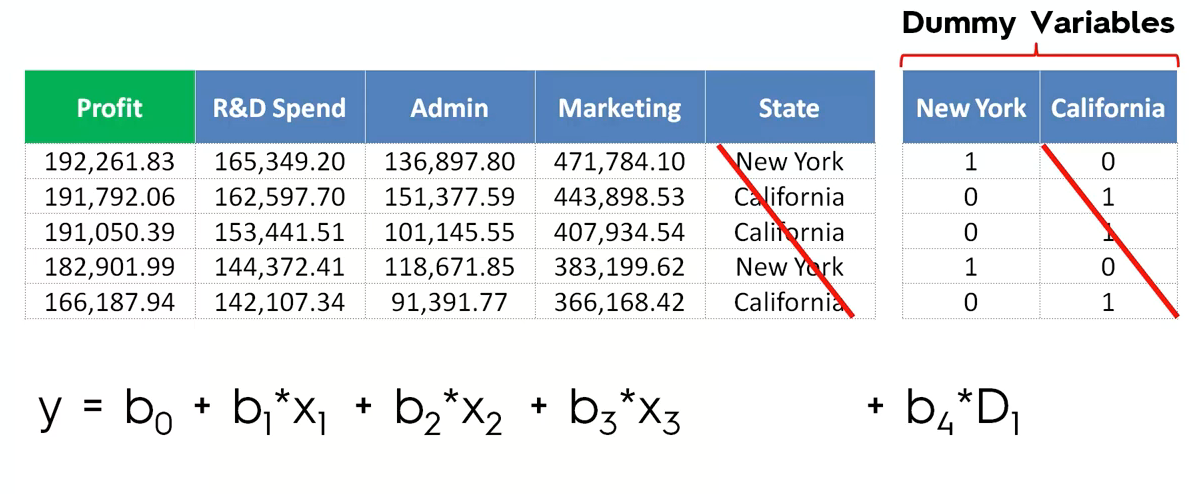
Equation: **y=A+BX1+CX2+DX3**





**Check the above factors before jumping into building linear regression**

**Multiple Linear Regression:**



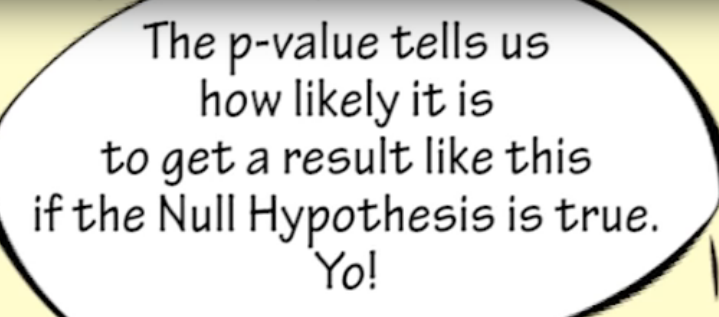
Green color(Profit)🡪dependent variable(y)

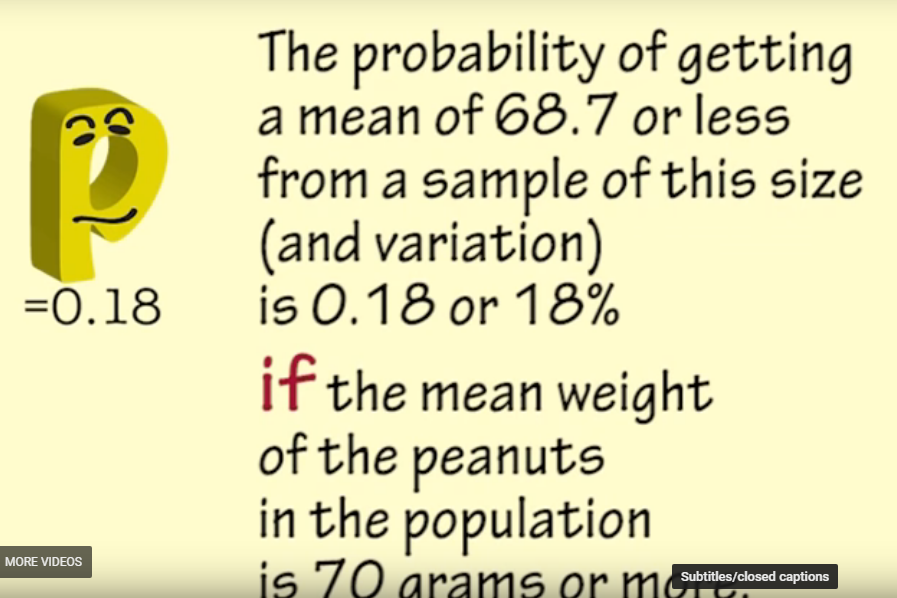
Blue color tabular col🡪independent variable(b1—Coefficient,x1🡪variable)

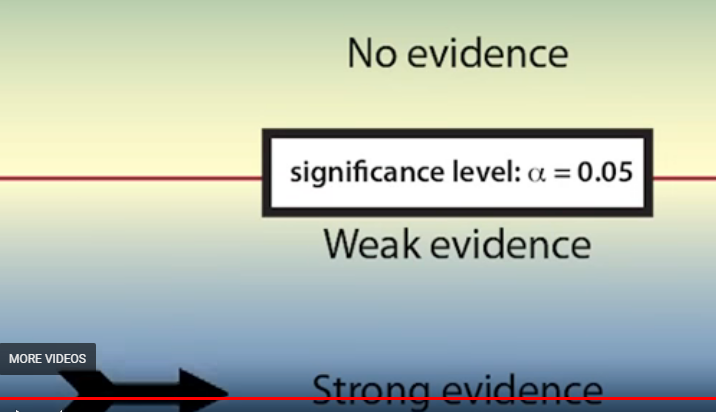
B0🡪 constant

State🡪categorical data, hence we are using dummy variable.

**P-Value:**

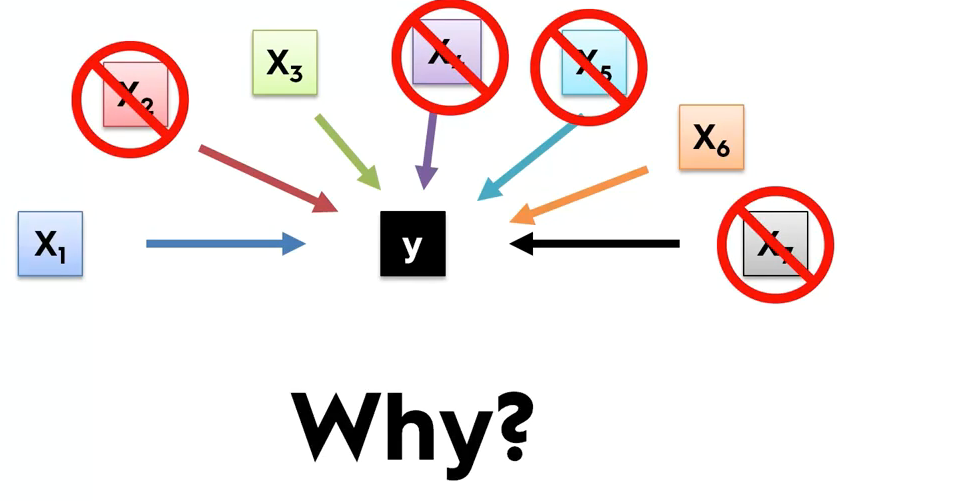






**Building a model step by step:**

We need to remove some independent variable,

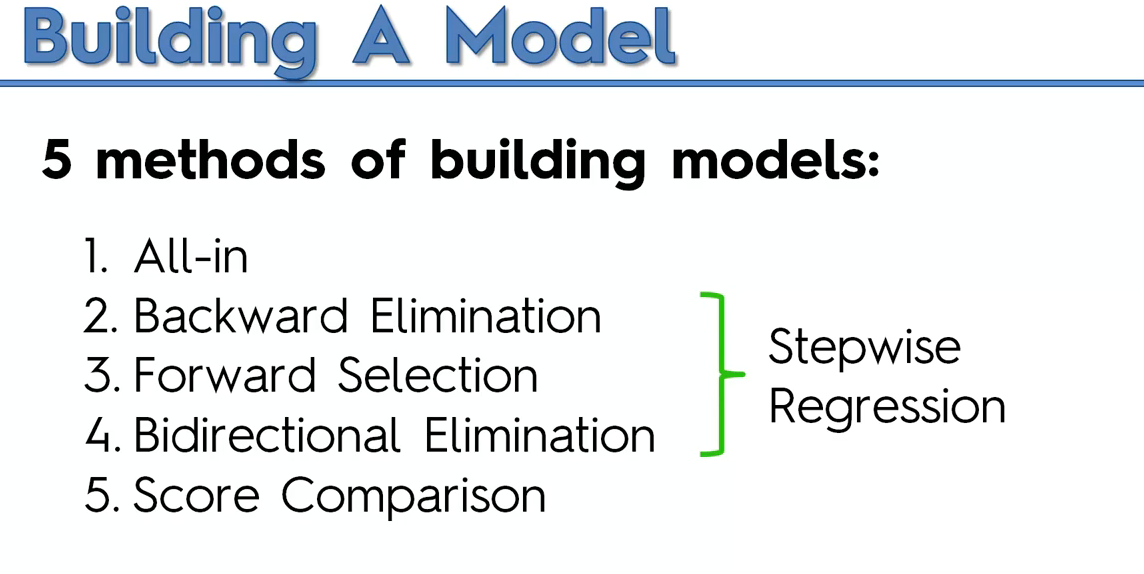


**Reason 1:**

Garbage in –Garbage out🡪 if you through in lot of stuff, then It won’t be good and accurate and it will be a garbage model.

**Reason 2:**

It may confuse the system, hence we should select the independent variable carefully.

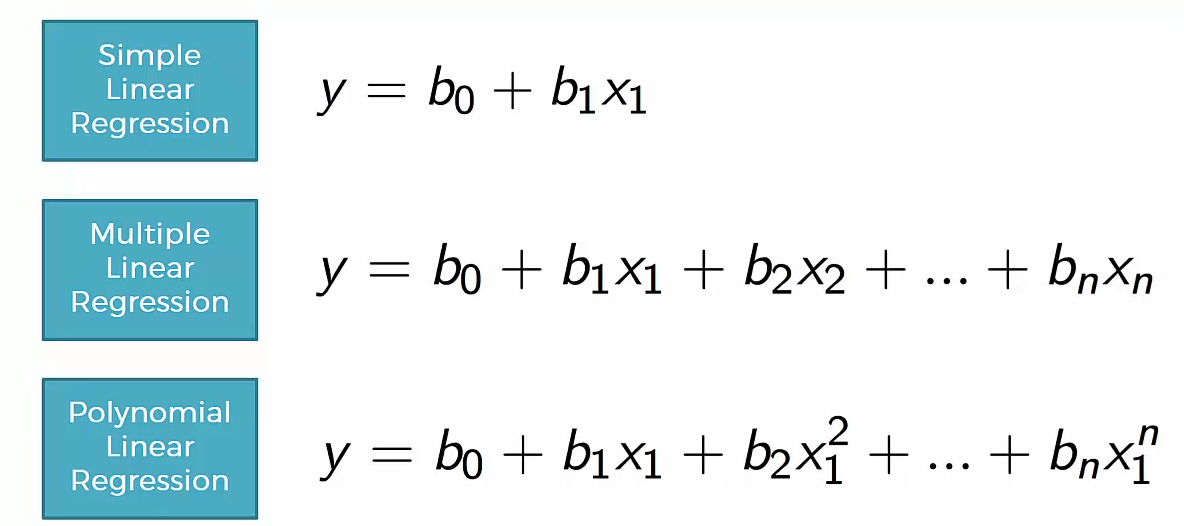


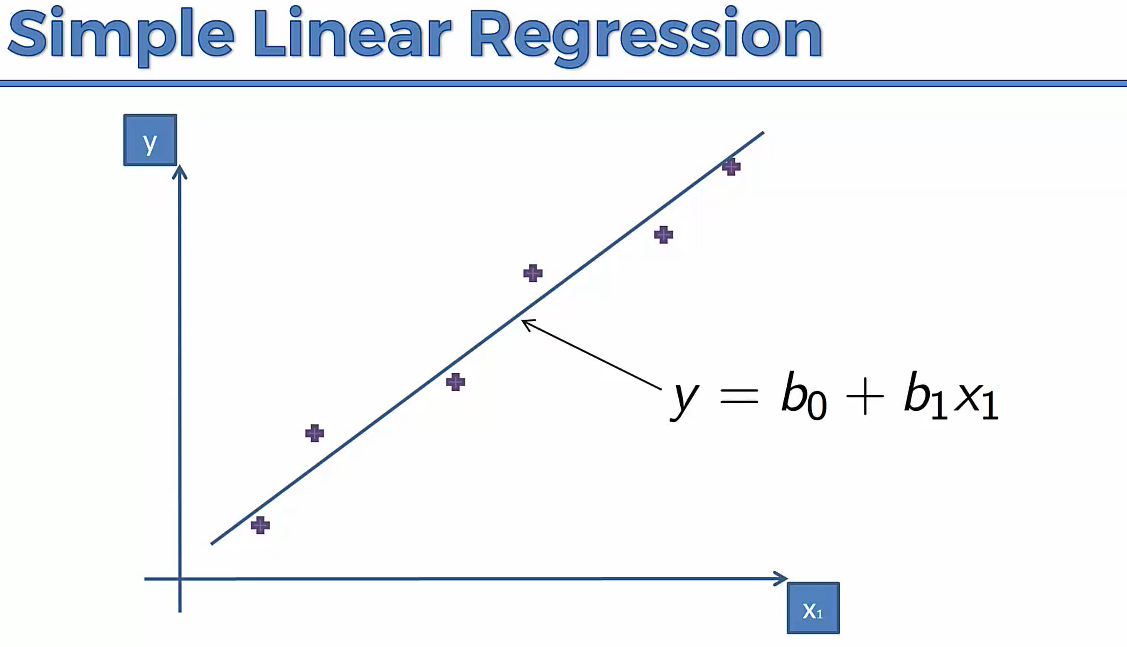
Refer the “**Building a model**” PDF to know more

Backward Elimination is the best and efficient.

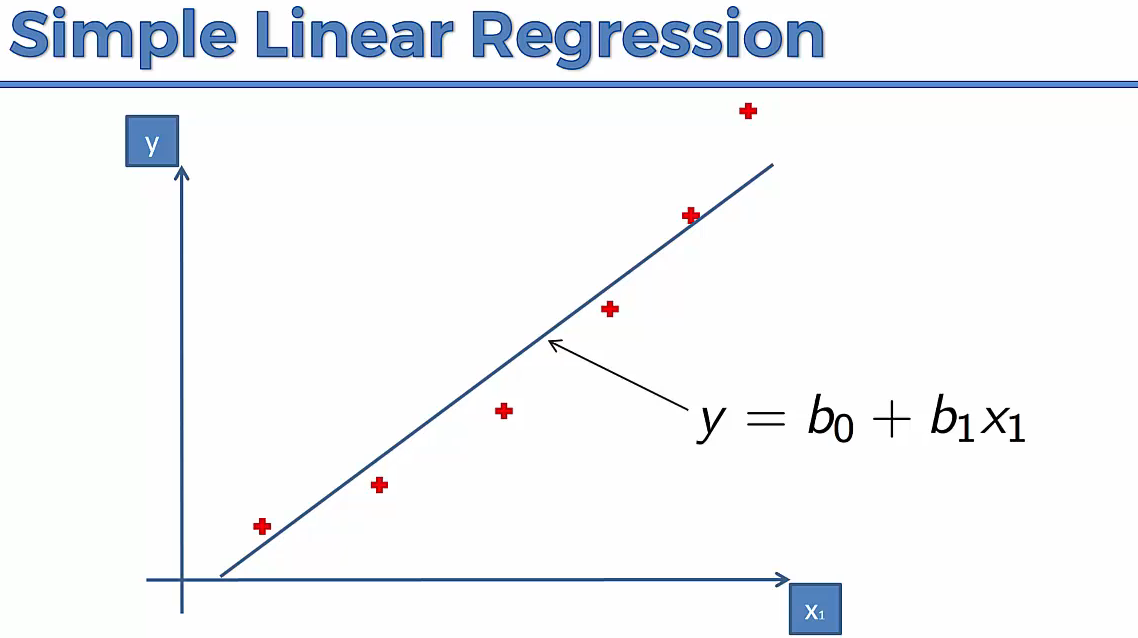
Score comparison consumes so much of the resource and the complexity grows exponentially .

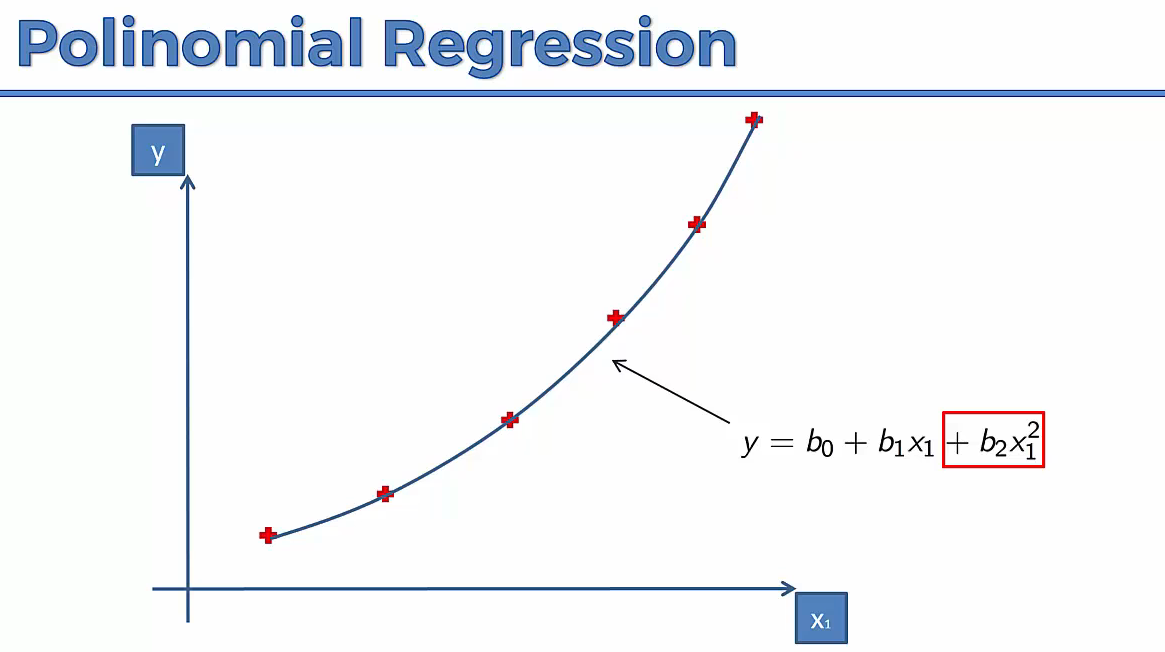
**Polynomial Linear regression:**





If we have plots, like the below, Simple linear regression won’t work

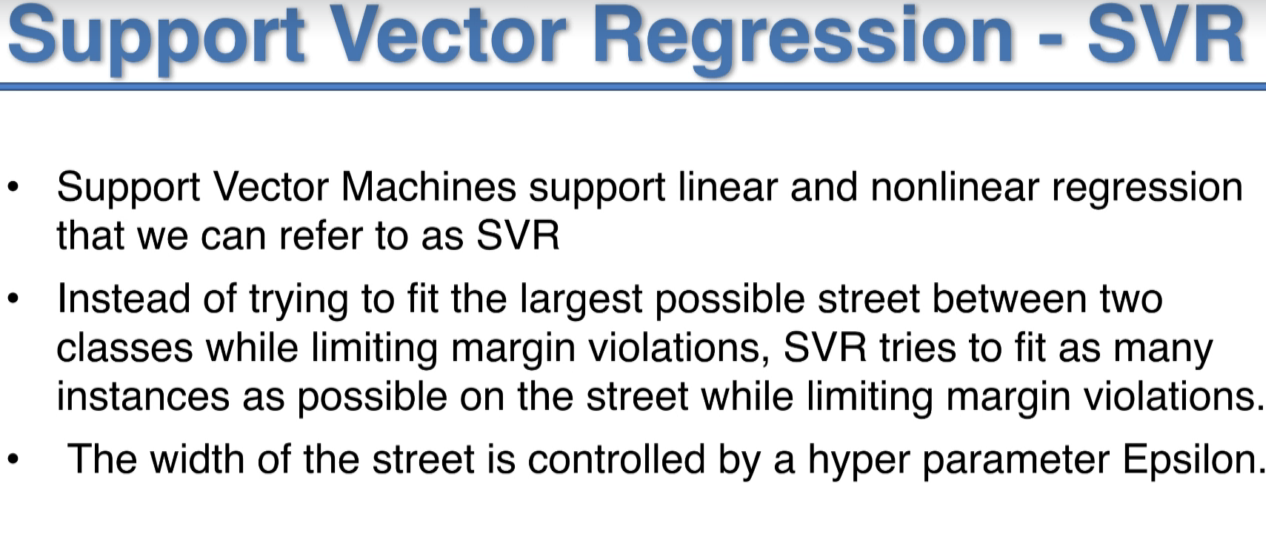


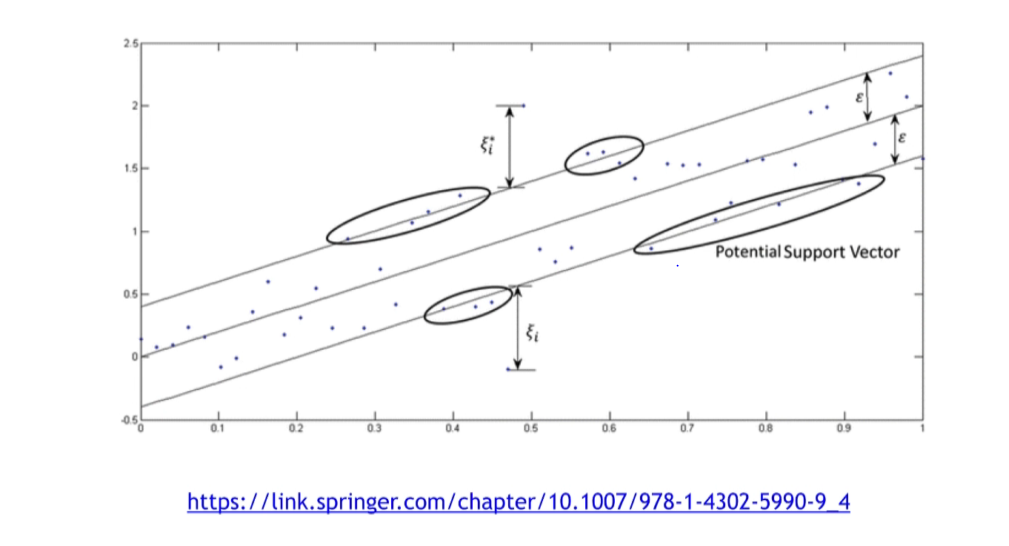


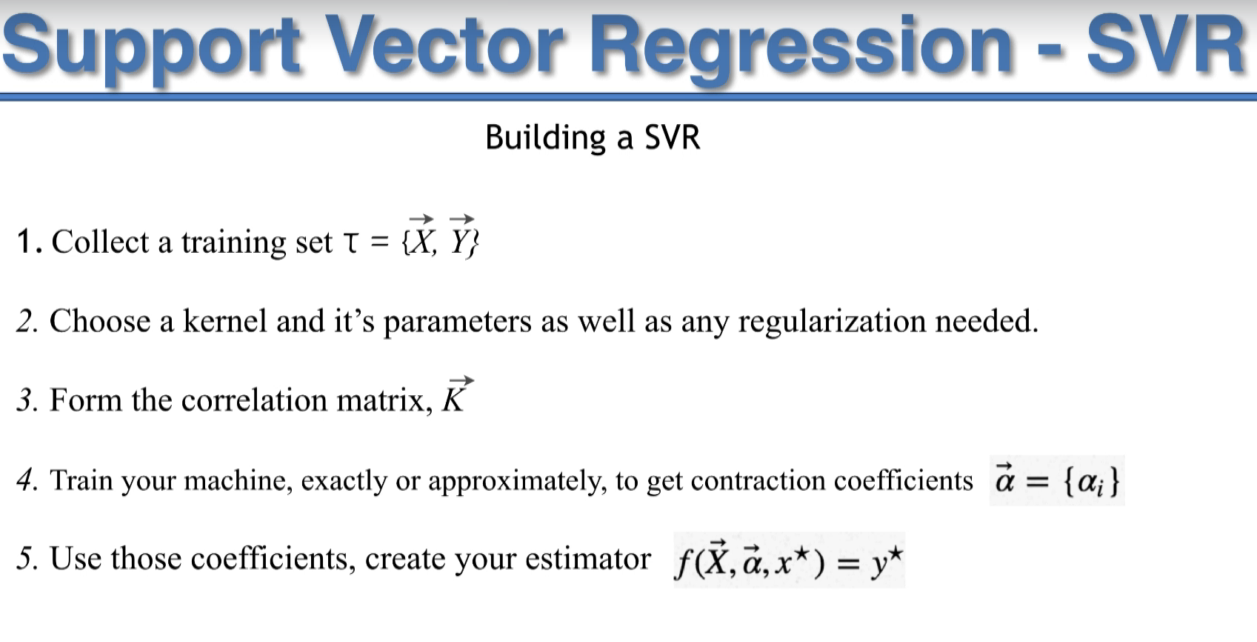
**WHY POLINOMIAL IS STILL LINEAR REGRESSION?**

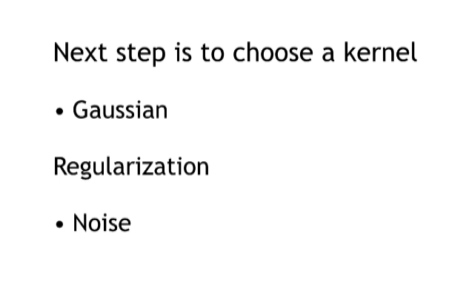
---We are cared on Co-efficient like b0,b1,b2….It is a special case of linear regression.

**SUPPORT VECOTR REGRESSION-SVR:**

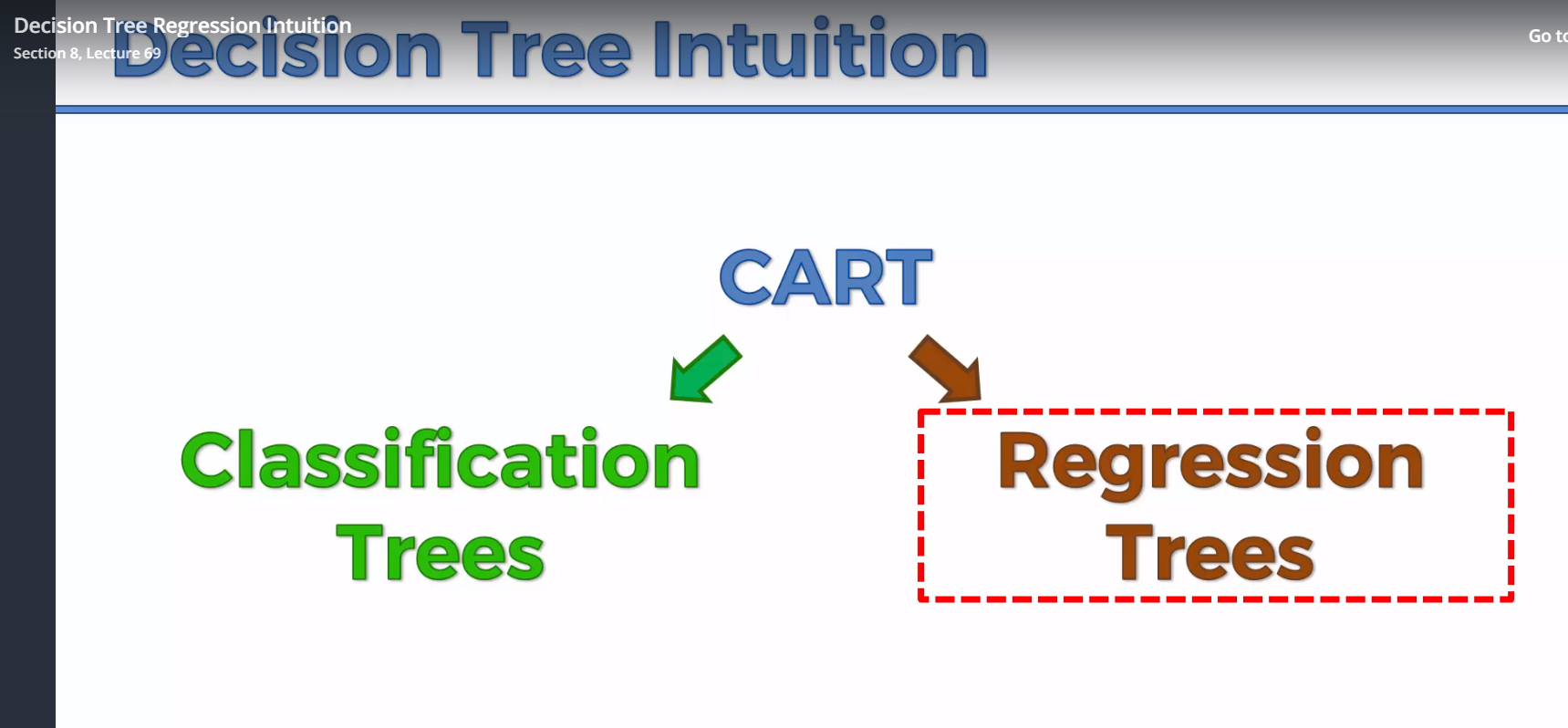


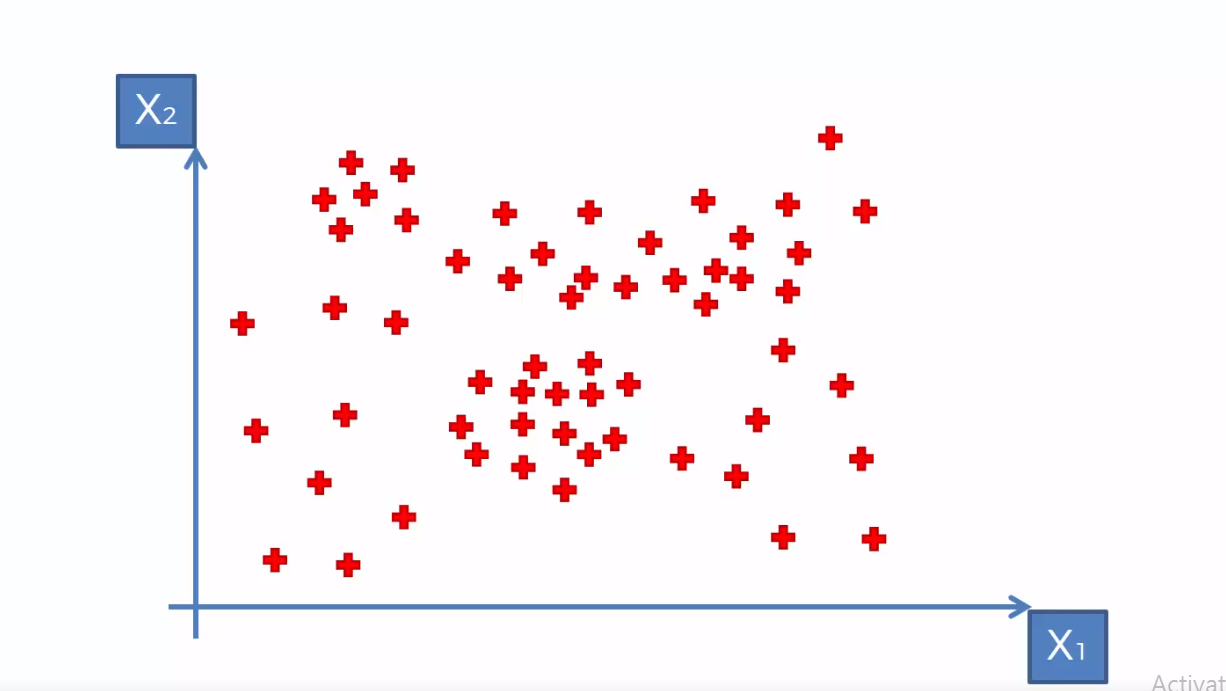




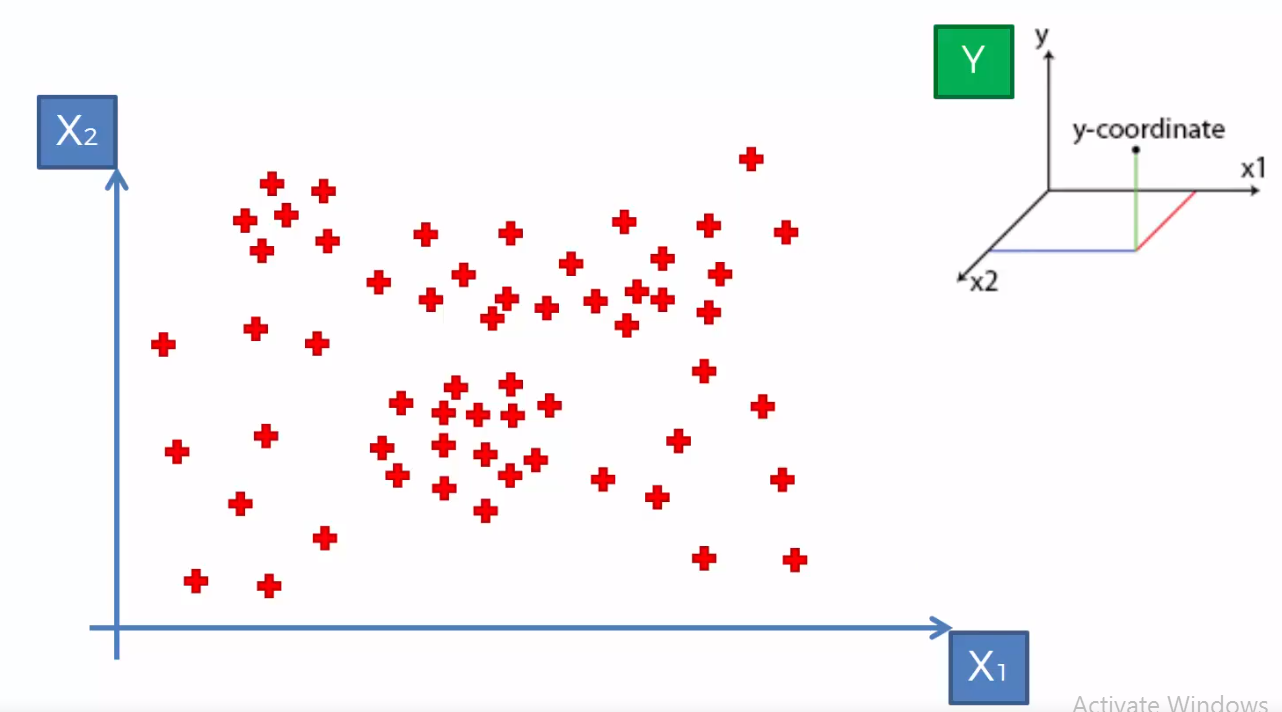


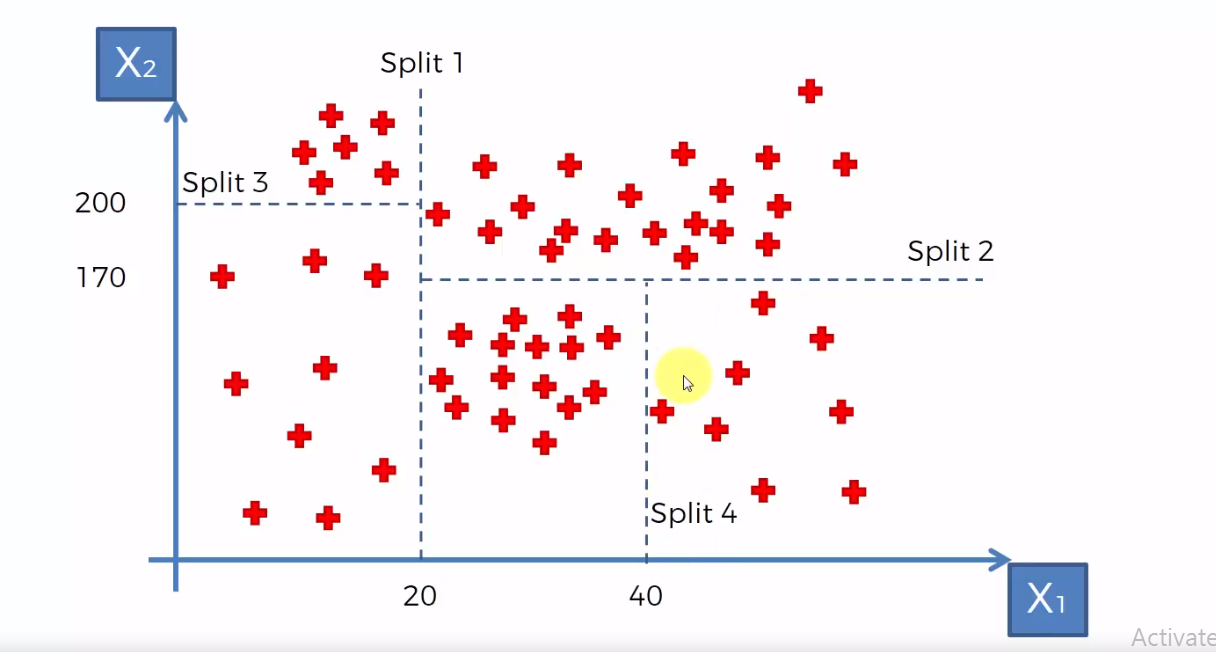
Decision Tree Regression:



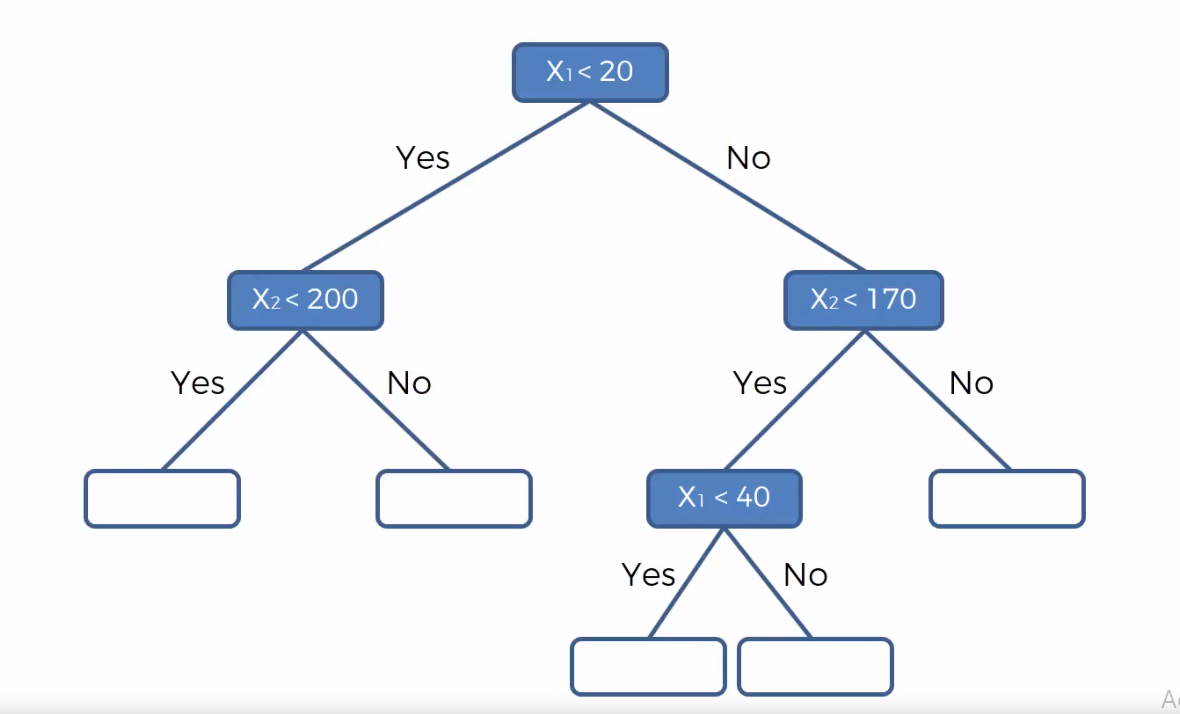


X0,x1 are the independent variable with which we have to predict y in the 3-Dimension (dependent variable)



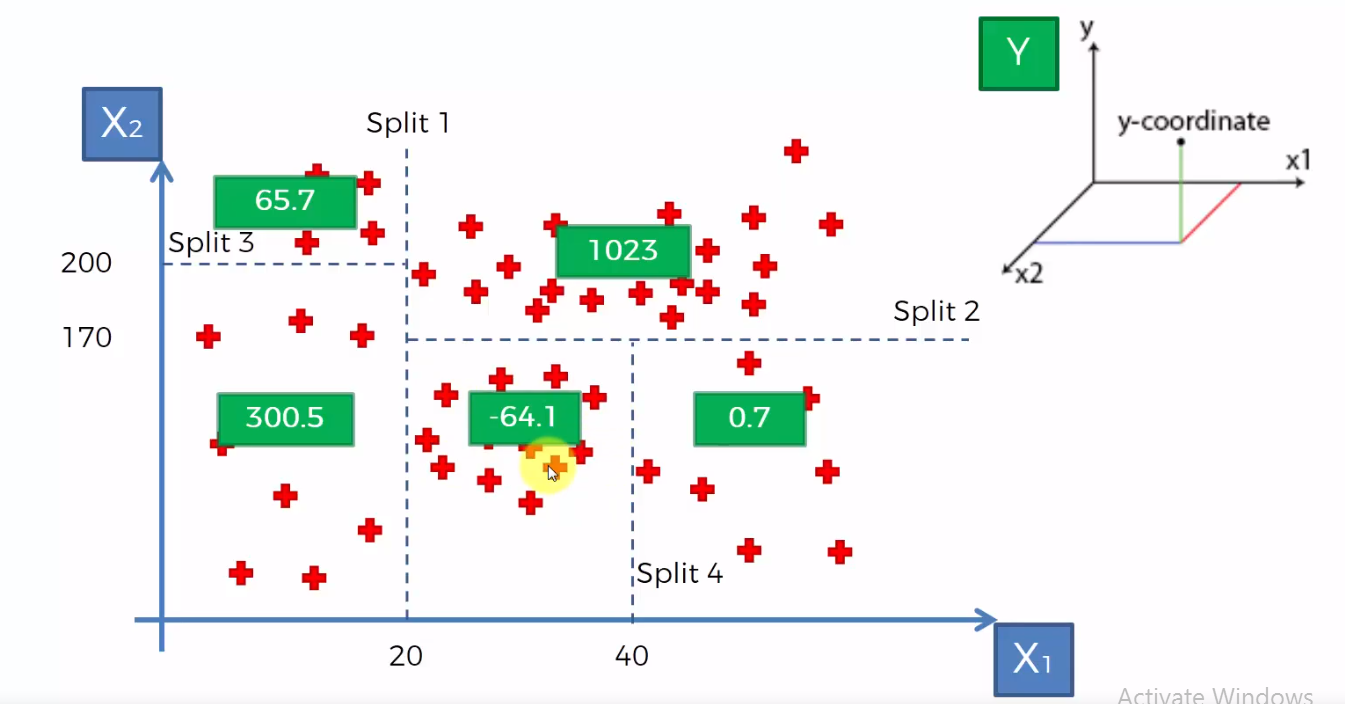


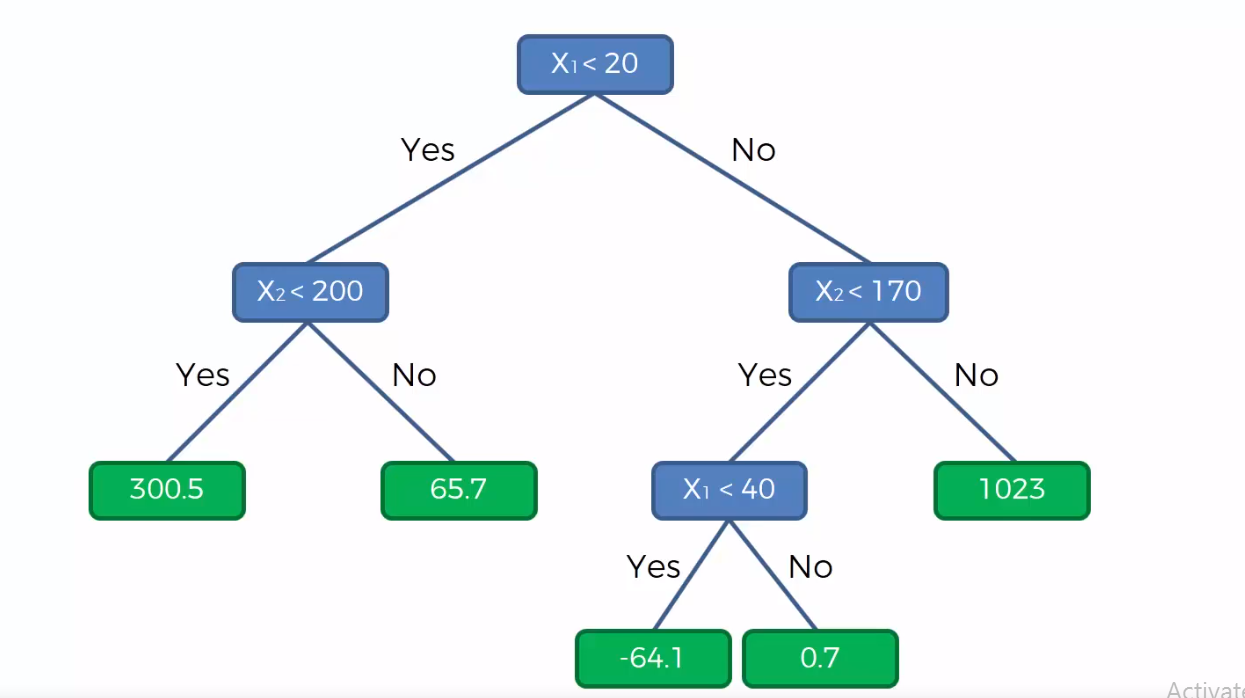
Based on the split, the decision tree is drawn,



The average is taken in the split data, the prediction is done based on the fall of the split position.

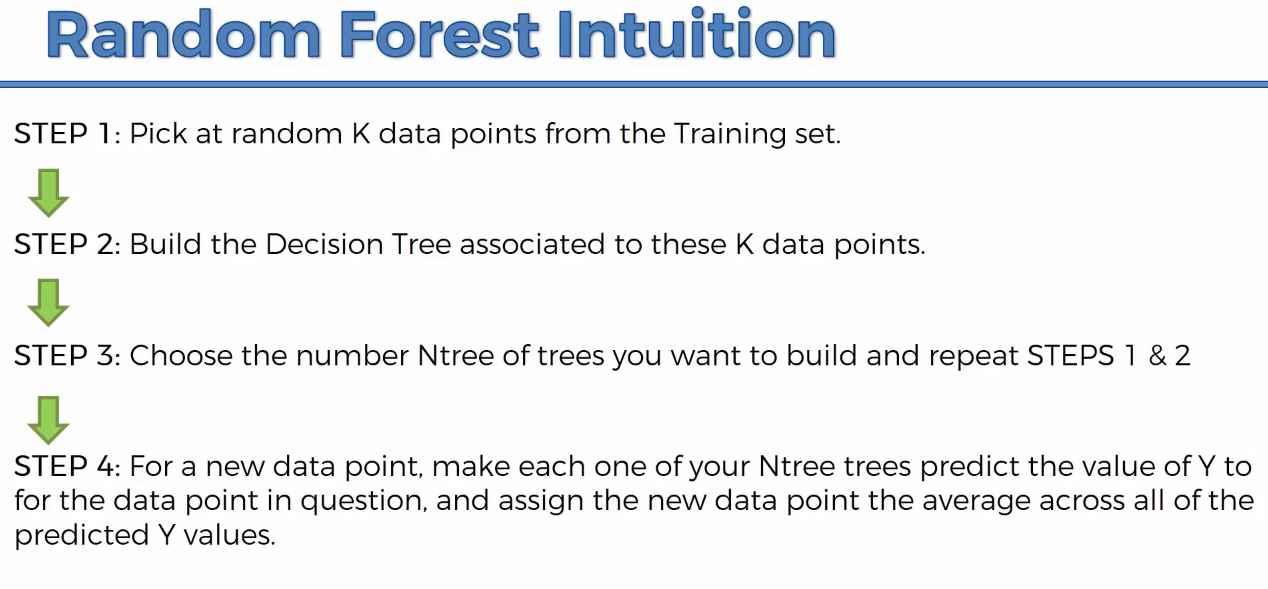
Eg: if x1 is greater than 20 less than 40 and x2 is less the 170 then the prediction is -64.1





**Random Forest:**

Random forest is Ensemble Learning.

Steps in building Random Forest, 

It comprises of many trees like forest. It is more stable, hence ensemble is more powerful.

Example:



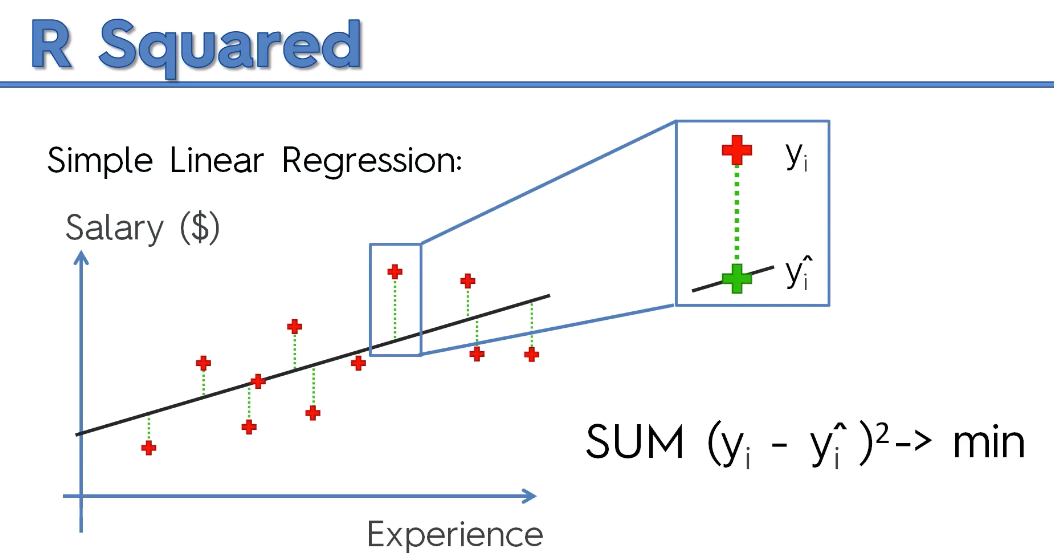
Take all the wild guess and take average it and take the median.

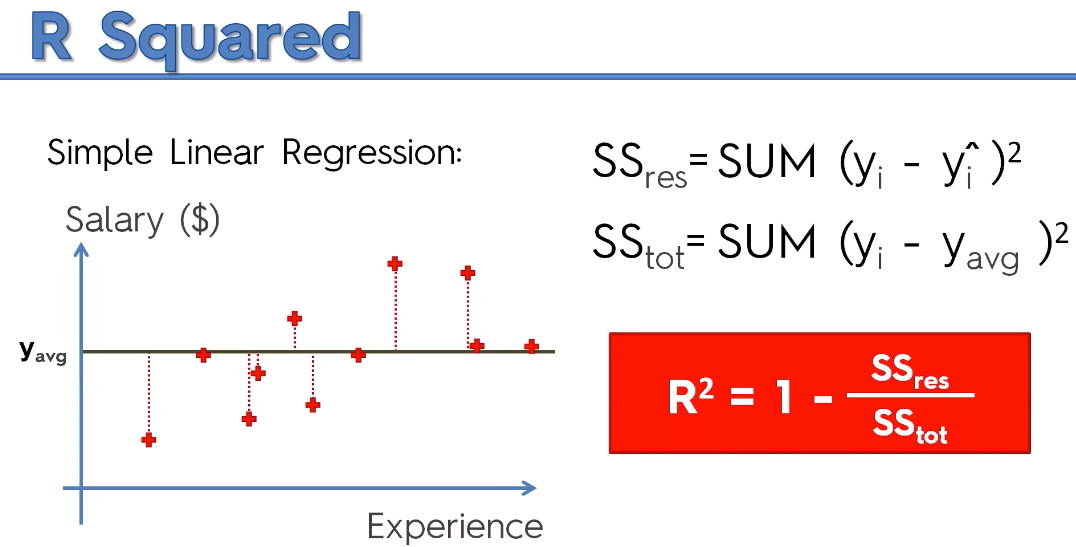
The likelihood of mean we got will be correct.

1. It is a non-linear regression model.

2. It is a team of decision tree. Each will do prediction of dependent variable and then the average of the different prediction of all the different tree in the forest.

**R-Squared:**



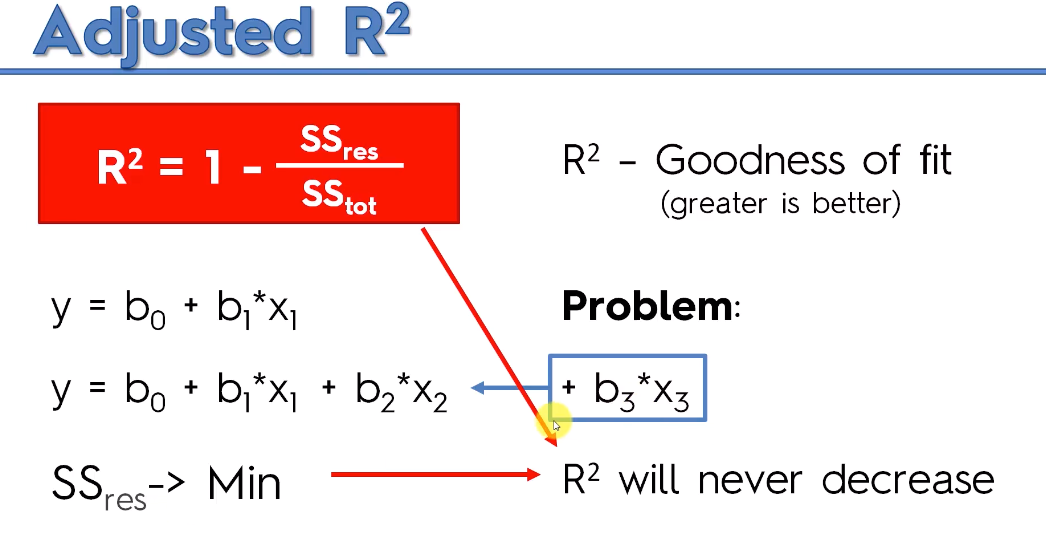


R Squared will give, how good the best fitting line is ?

SSresiduals=0, then Rsquare is 1.Then it passes through all the points.

R^2- Goodness of fit(greater is better)

If we add a new variable with co-efficient,



1. R^2 will increase or

2. R^2 will remain the same.

It will difficult to find the best fit, hence we use Adjuster R^2.

