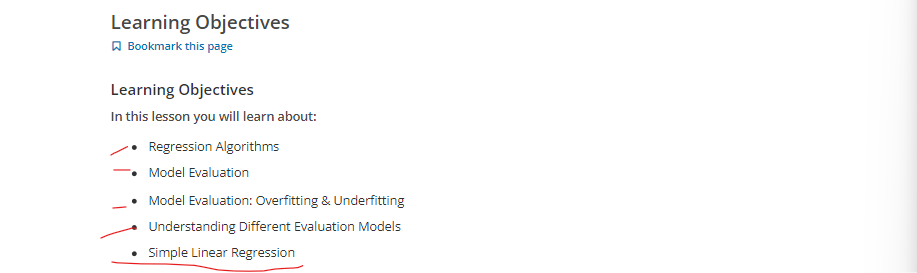
**[Module 2 - Regression](https://courses.excelr.skillsnetwork.site/courses/course-v1:CognitiveClass+ML0101ENv3+v4/course/" \l "block-v1:CognitiveClass+ML0101ENv3+v4+type@chapter+block@bd64ccdf56ad4ea1afe870e26d583038) 24/09**

**1st Slide:**

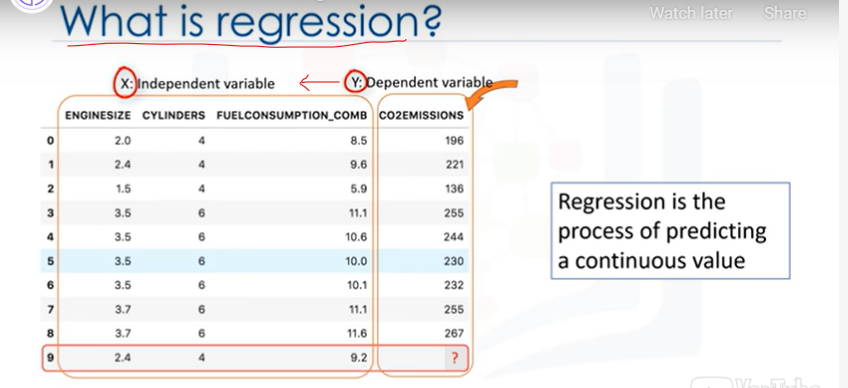
****

****

**A brief introduction to regression.**

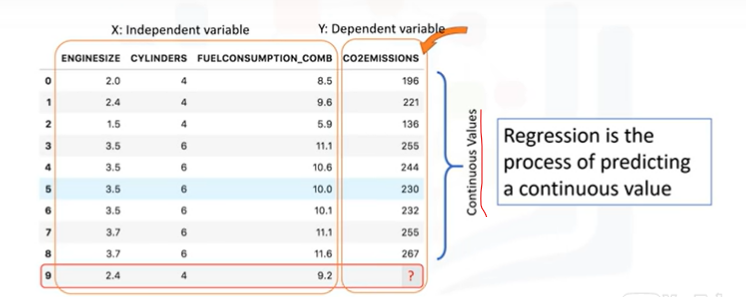
**Look at this dataset.**

1. It's related to Co2 emissions from different cars.
2. It includes Engine size, number of Cylinders, Fuel Consumption and Co2 emission from various automobile models.
3. The question is, "Given this dataset, can we predict the Co2 emission of a car using other fields, such as EngineSize or Cylinders?"
4. Let’s assume we have some historical data from different cars, and assume that a car, such as in row 9, has not been manufactured yet, but we're interested in estimating its approximate Co2 emission, after production Is it possible ?

****

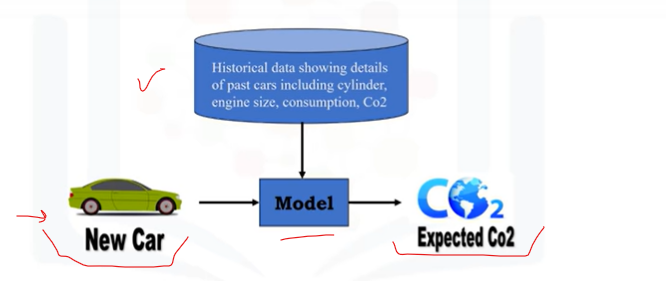
1. We can use regression methods to predict a continuous value, such as CO2 Emission, using some other variables.
2. Indeed, regression is the process of predicting a continuous value.
3. In regression there are two types of variables: a dependent variable and one or more independent variables.
4. The dependent variable can be seen as the "state", "target" or "final goal" we study and try to predict, and the independent variables, also known as explanatory variables, can be seen as the "causes" of those "states".
5. The independent variables are shown conventionally by x; and the dependent variable is notated by y.
6. A regression model relates y, or the dependent variable, to a function of x, i.e., the independent variables.
7. 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 measured on either a categorical or continuous measurement scale.**

****

**So, what we want to do here is to use the historical data of some cars, using one or more of their features, and from that data, make a model.**

1. We use regression to build such a regression/estimation model.
2. Then the model is used to predict the expected Co2 emission for a new or unknown car.

****

**Basically there are 2 types of regression models: simple regression and multiple regression.**

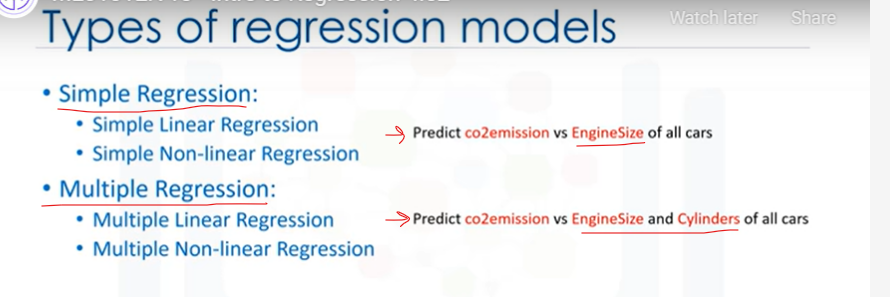
1. Simple regression is when one independent variable is used to estimate a dependent variable. It can be either linear on non-linear.

For example, predicting Co2emission using the variable of EngineSize.

1. Linearity of regression is based on the nature of relationship between independent and dependent variables.
2. When more than one independent variable is present, the process is called multiple linear regression.

For example, predicting Co2emission using EngineSize and the number of Cylinders in any given car.

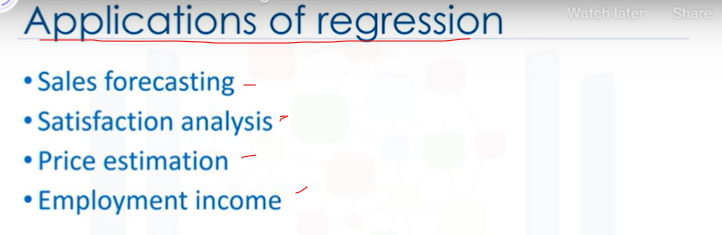
1. Again, depending on the relation between dependent and independent variables, it can be either linear or non-linear regression.



**Let’s examine some sample applications of regression.**

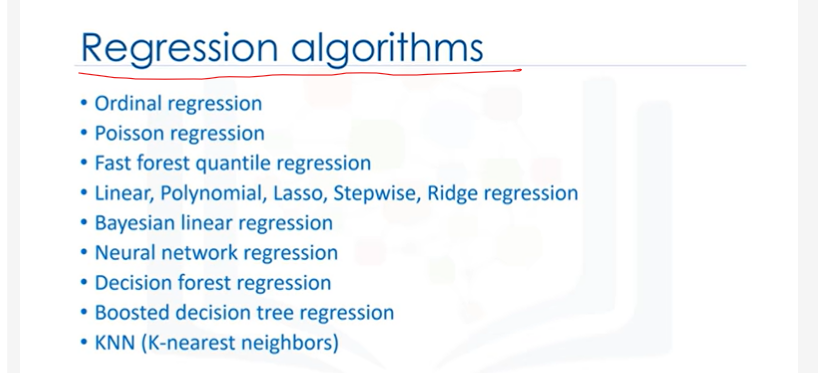
Essentially, we use regression when we want to estimate a continuous value.

1. For instance, one of the applications of regression analysis could be in the area of sales forecasting. You can try to predict a salesperson's total yearly sales from independent variables such as age, education, and years of experience.
2. It can also be used in the field of psychology, for example, to determine individual satisfaction based on demographic and psychological factors.
3. 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.
4. 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.
5. Indeed, you can find many examples of the usefulness of regression analysis in these and many other fields or domains, such as finance, healthcare, retail, and more.

****

**We have many regression algorithms.**

1. **Each of them has its own importance and a specific condition to which their application** is best suited.
2. And while we've covered just a few of them in this course, it gives you enough base knowledge for you to explore different regression techniques.

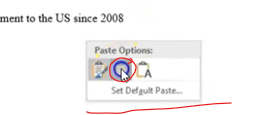


**You can Remove the Grey Background color in word document by selecting the Area and In Paragraphs Pane select that (Icon which is Pouring Color and Select “No Color”)**

Or Immediate method is

While Pasting you see a below Symbol->click the dorpdown!



Click on Middle (Merge Formatting Option)

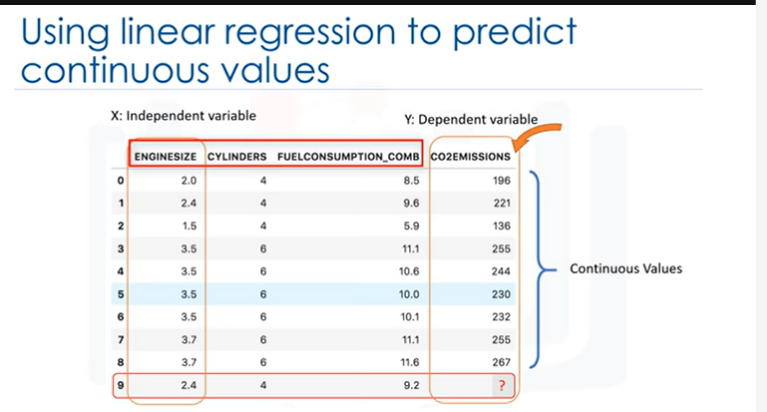
**Second Video!**

****

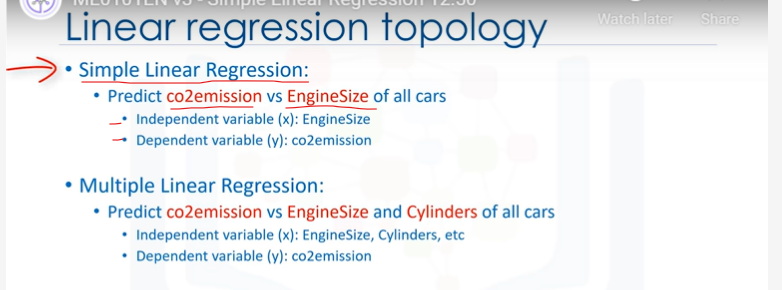
We’ll be covering linear regression.

1. You don’t need to know any linear algebra to understand topics in linear regression.
2. This high-level introduction will give you enough background information on linear regression to be able to use it effectively on your own problems.
3. Let’s take a look at this dataset. It’s related to the Co2 emission of different cars. It includes Engine size, Cylinders, Fuel Consumption and Co2 emissions for various car models.

**The question is: Given this dataset, can we predict the Co2 emission of a car, using another field, such as Engine size?**

-> Quite simply, yes! We can use linear regression to predict a continuous value such as Co2 Emission, by using other variables.

1. Linear regression is the approximation of a linear model used to describe the relationship between two or more variables.
2. In simple linear regression, there are two variables: a dependent variable and an independent variable.
3. The key point in the linear regression is that our dependent value should be continuous and cannot be a discreet value.
4. However, the independent variable(s) can be measured on either a categorical or continuous measurement scale.

****

1. **There are two types of linear regression models. They are: simple regression and multiple regression.**
2. **Simple linear regression is when one independent variable is used to estimate a dependent variable.**

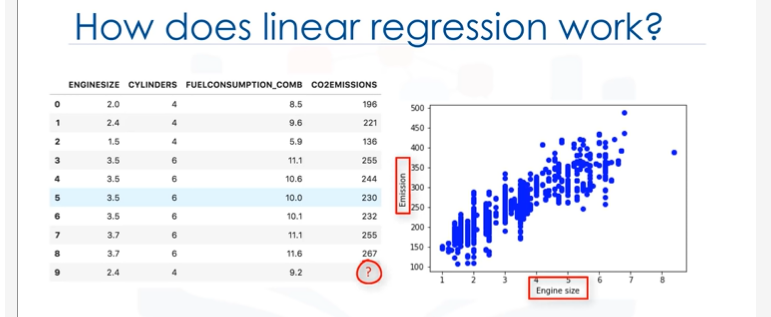
**For example, predicting Co2 emission using the EngineSize variable.**

1. **When more than one independent variable is present, the process is called multiple linear regression.**

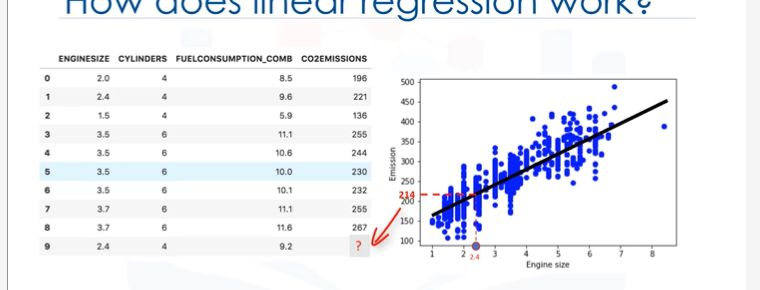
**For example, predicting Co2 emission using EngineSize and Cylinders of cars.**

**Our focus will be on simple linear regression.**

**let’s see how linear regression works. OK, so let’s look at our dataset again.**

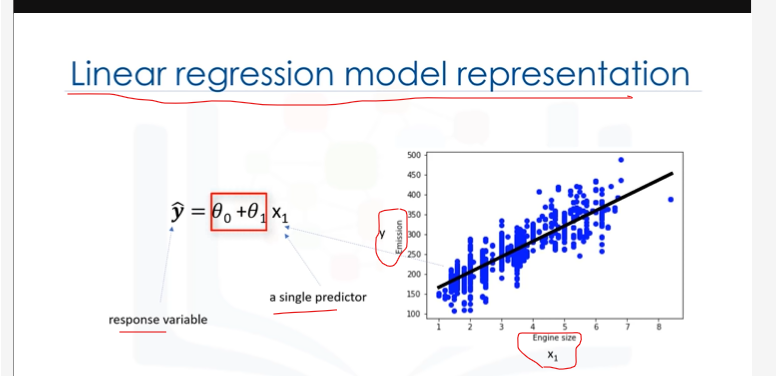


1. To understand linear regression, we can plot our variables here.
2. We show Engine size as an independent variable, and Emission as the target value that we would like to predict.
3. A scatterplot clearly shows the relation between variables where changes in one variable "explain" or possibly "cause" changes in the other variable.
4. Also, it indicates that these variables are linearly related.
5. With linear regression you can fit a line through the data.
6. For instance, as the EngineSize increases, so do the emissions.
7. With linear regression, you can model the relationship of these variables.
8. A good model can be used to predict what the approximate emission of each car is.
9. How do we use this line for prediction now? Let us assume, for a moment, that the line is a good fit of data.
10. We can use it to predict the emission of an unknown car. For example, for a sample car, with engine size 2.4, you can find the emission is 214.



**Now, let’s talk about what this fitting line actually is.**

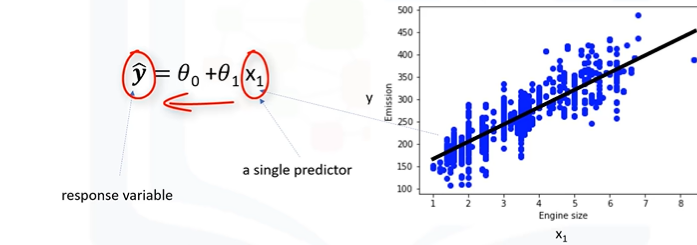
1. **We’re going to predict the target value, y.**
2. In our case, using the independent variable, "Engine Size," represented by x1.
3. The fit line is shown traditionally as a polynomial. In a simple regression problem (a single x),
4. the form of the model would be θ0 +θ1 x1. In this equation, y ̂ is the dependent variable or the predicted value, and x1 is the independent variable; θ0 and θ1 are the parameters of the line that we must adjust.



1. **θ1 is known as the "slope" or "gradient" of the fitting line and**
2. **θ0 is known as the "intercept."**

**θ0 and θ1 are also called the coefficients of the linear equation.**

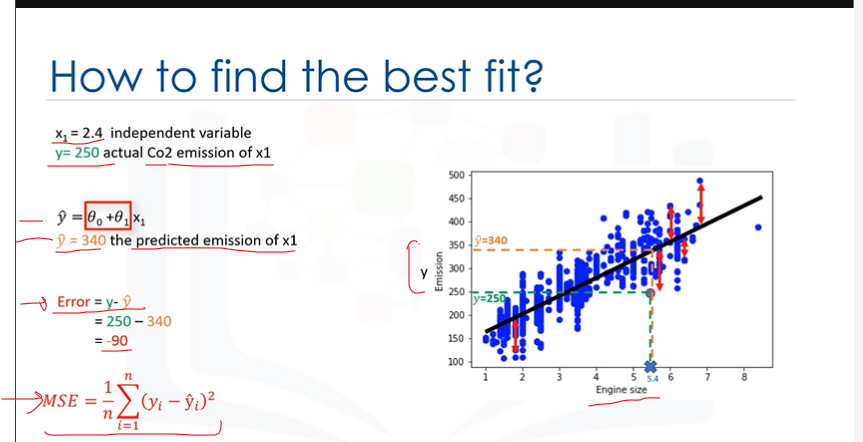
**You can interpret this equation as y ̂ being a function of x1, or y ̂ being dependent of x1.**

****

**Now the questions are: "How would you draw** a line through the points?" And, "How do you determine which line ‘fits best’?"

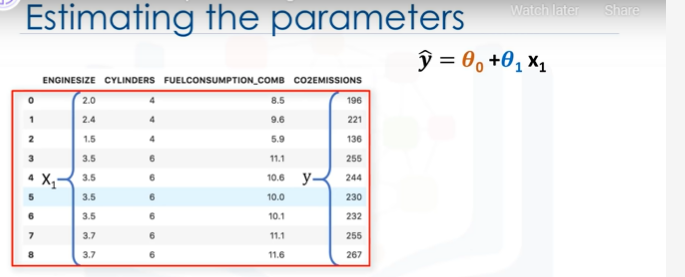
1. Linear regression estimates the coefficients of the line.
2. This means we must calculate θ0 and θ1 to find the best line to ‘fit’ the data.
3. This line would best estimate the emission of the unknown data points.
4. Let’s see how we can find this line

**Let’s see how we can find this line, or to be more precise, how we can adjust the parameters to make the line the best fit for the data.**

****

1. For a moment, let’s assume we’ve already found the best fit line for our data.
2. Now, let’s go through all the points and check how well they align with this line.
3. Best fit, here, means that if we have, for instance, a car with engine size x1=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.
4. But, if we use the fit line, or better to say, using our polynomial with known parameters to predict the Co2 emission, it will return y ̂ =340.
5. Now, if you compare the actual value of the emission of the car with what we predicted using our model, you will find out that we have a 90-unit error.
6. **This means our prediction line is not accurate. This error is also called the Residual error.**
7. So, we can say the error is the distance from the data point to the fitted regression line.
8. The mean of all residual errors shows how poorly the line fits with the whole dataset.
9. **Mathematically, it can be shown by the equation, mean squared error, shown as (MSE).**
10. Our objective is to find a line where the mean of all these errors is minimized.
11. In other words, the mean error of the prediction using the fit line should be minimized.
12. Let’s re-word it more technically. The objective of linear regression is to minimize this MSE equation, and to minimize it, we should find the best parameters, θ0 and θ1.
13. **Now, the question is, how to find θ0 and θ1 in such a way that it minimizes this error? How can we find such a perfect line? Or, said another way, how should we find the best parameters for our line? Should we move the line a lot randomly and calculate the MSE value every time, and choose the minimum one?**
14. Not really! Actually, we have two options here:
15. Option 1 - We can use a mathematic approach. Or, Option 2 - We can use an optimization approach.

**Let’s see how we can easily use a mathematic formula to find the θ0 and θ1.**

****

1. As mentioned before, θ0 and θ1, in the simple linear regression, are the coefficients of the fit line. We can use a simple equation to estimate these coefficients.
2. That is, given that it’s a simple linear regression, with only 2 parameters, and knowing that θ0 and θ1 are the intercept and Slope of the line, we can estimate them directly from our data.
3. It requires that we calculate the mean of the independent and dependent or target columns, from the dataset.

Notice that all of the data must be available **to traverse and calculate the parameters.**

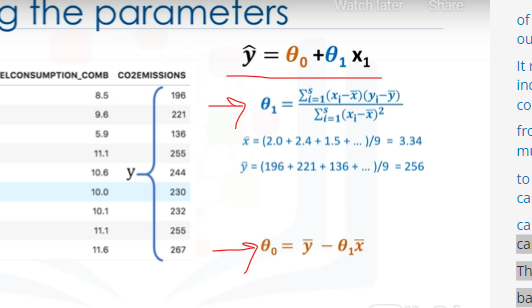
1. **It can be shown that the intercept and slope** can be calculated using these equations.

**Y = Θ.o + Θ.1x**

**θ1 is known as the "slope" or "gradient" of the fitting line and θ0 is known as the "intercept."**

**We can start off by estimating the value for θ1.**

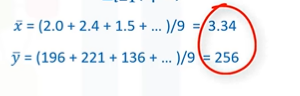
This is how you can find the slope of a line based on the data.



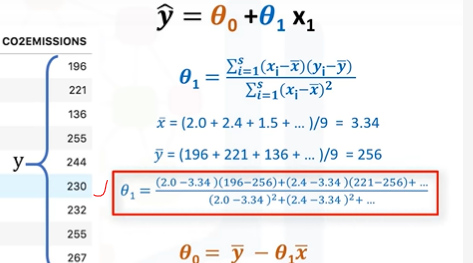
1. **x ̅ is the average value for the engine size** in our dataset.

****

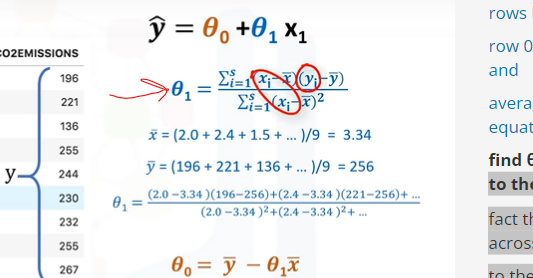
1. Please consider that we have 9 rows here,row 0 to 8.
2. First, we calculate the average of x1 andaverage of y.

****

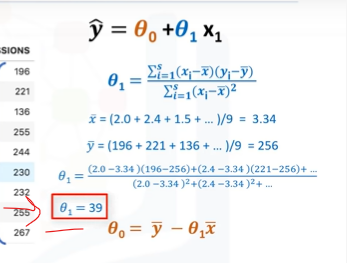
1. Then we plug it into the slope equation, tofind θ1.

****

1. **The xi and yi in the equation refer to the** fact that we need to repeat these calculations across all values in our dataset and i refersto the i’th value of x or y.

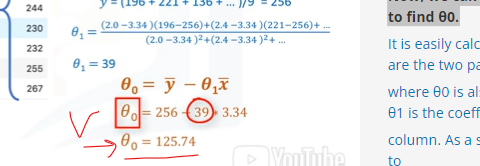
****

**Applying all values, we find θ1=39; it is our second parameter.**

****

**It is used to calculate the first parameter, which is the intercept of the line.**

**Now, we can plug θ1 into the line equation to find θ0.**

****

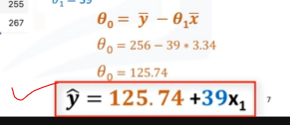
**It is easily calculated that θ0=125.74.**

**NOTE:**  So, these are the two parameters for the line, where θ0 is also called the bias coefficient and θ1 is the coefficient for the Co2 Emission **column.**

As a side note, you really don’t need to remember the formula for calculating these parameters, as most of the libraries used for machine learning in Python, R, and Scala can easily find these parameters for you.

But it’s always good to understand how it works.

1. **Now, we can write down the polynomial of the line.**



So, we know how to find the best fit for our data, and its equation.

**Que) Now the question is: "How can we use it to predict the emission of a new car based on its engine size?"**

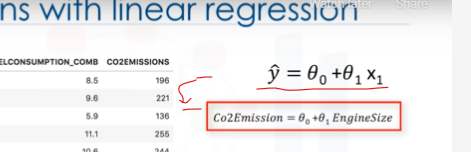
* **After we found the parameters of the linear equation, making predictions is as simple** as solving the equation for a specific set of inputs.

Imagine we are predicting Co2 Emission(y) from EngineSize(x) for the Automobile in record **number 9.**

****

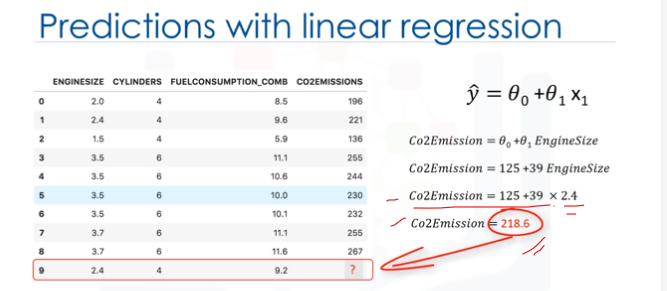
**Our linear regression model representation** for this problem would be**: y ̂ = θ0 + θ1 x1.**

Or if we map it to our dataset, it would be Co2Emission = θ0 + θ1 EngineSize.



**For example, let’s use θ0=125 and θ1=39. So, we can rewrite the linear model as 𝐶𝑜2𝐸𝑚𝑖𝑠𝑠𝑖𝑜𝑛=125+39𝐸𝑛𝑔𝑖𝑛𝑒𝑆𝑖𝑧𝑒.**

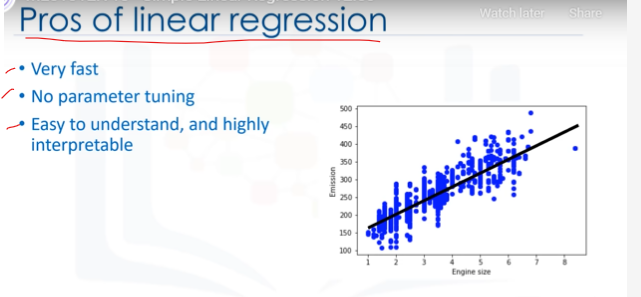
1. Now, let’s plug in the 9th row of our dataset and calculate the Co2 Emission for a car with an EngineSize of 2.4. So Co2Emission = 125 + 39 × 2.4.
2. Therefore, we can predict that the Co2 Emission for this specific car would be 218.6.



**WHY Linear Regression?**

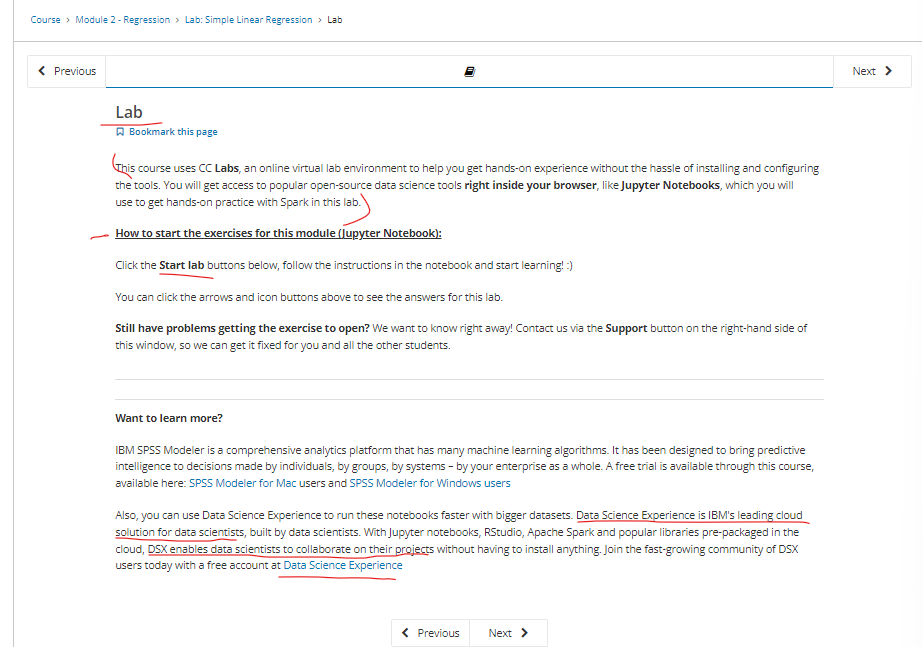
Let’s talk a bit about why Linear Regression is so useful.

1. **Quite simply, it is the most basic regression to use and understand.**
2. In fact, one reason why Linear Regression is so useful is that it’s fast!
3. It also doesn’t require tuning of parameters. So, something like tuning the K parameter in K-Nearest Neighbors or the learning rate in Neural Networks isn’t something to worry about.
4. Linear Regression is also easy to understand and highly interpretable.



END of LR Discussion

**Third Step after Video**

****

**Want to learn more?**

IBM SPSS Modeler is a comprehensive analytics platform that has many machine learning algorithms. It has been designed to bring predictive intelligence to decisions made by individuals, by groups, by systems – by your enterprise as a whole. A free trial is available through this course, available here: [SPSS Modeler for Mac](https://cocl.us/ML0101EN_SPSSMod_mac) users and [SPSS Modeler for Windows users](https://cocl.us/ML0101EN_SPSSMod_win)

Also, you can use Data Science Experience to run these notebooks faster with bigger datasets. Data Science Experience is IBM's leading cloud solution for data scientists, built by data scientists. With Jupyter notebooks, RStudio, Apache Spark and popular libraries pre-packaged in the cloud, DSX enables data scientists to collaborate on their projects without having to install anything. Join the fast-growing community of DSX users today with a free account at [Data Science Experience](https://cocl.us/ML0101EN_DSX)

Create a 30 Day Free Tier Account in IBM SPSS Modeler for windows!