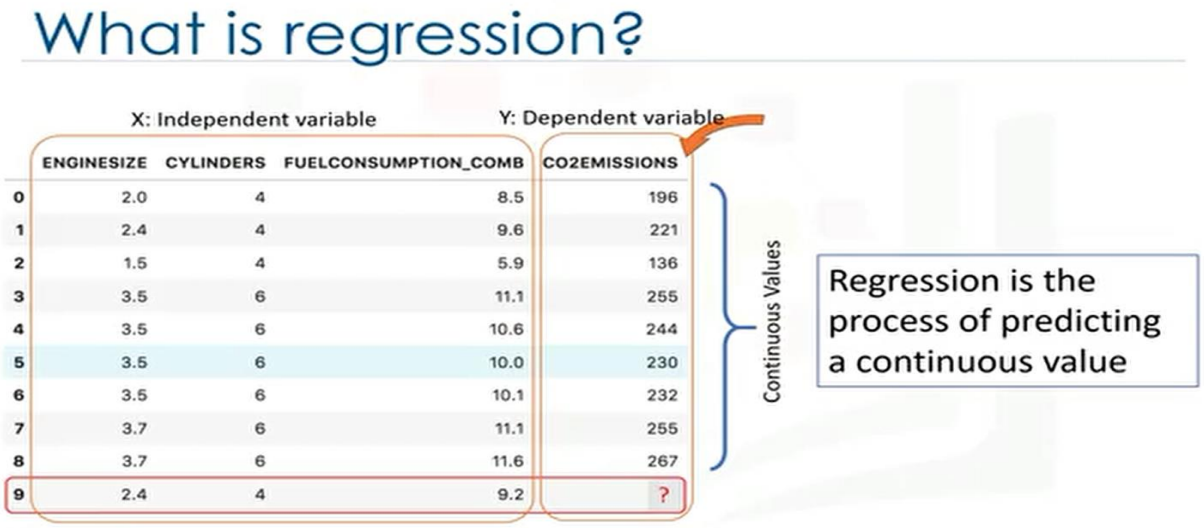
Week 2 Regression

Week 2 index 1: Introduction to Regression

It's related to co2 emissions from different cars. It includes engine size, number of cylinders, fuel consumption, and co2 emission from various automobile models.



The question is: given this data set can we predict the co2 emission of a car using other fields such as engine size or cylinders?

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? We can use regression methods to predict a continuous value such as co2 emission using some other variables.

Indeed, regression is the process of predicting a continuous value.

In regression there are two types of variables:

- ✓ a dependent variable and
- ✓ one or more independent variables.

Dependent variable: The dependent variable can be seen as the state, target, or final goal we study and try to predict.

Independent variables: the independent variables, also known as explanatory variables, can be seen as the causes of those states.

The independent variables are shown conventionally by X and the dependent variable is notated by Y.

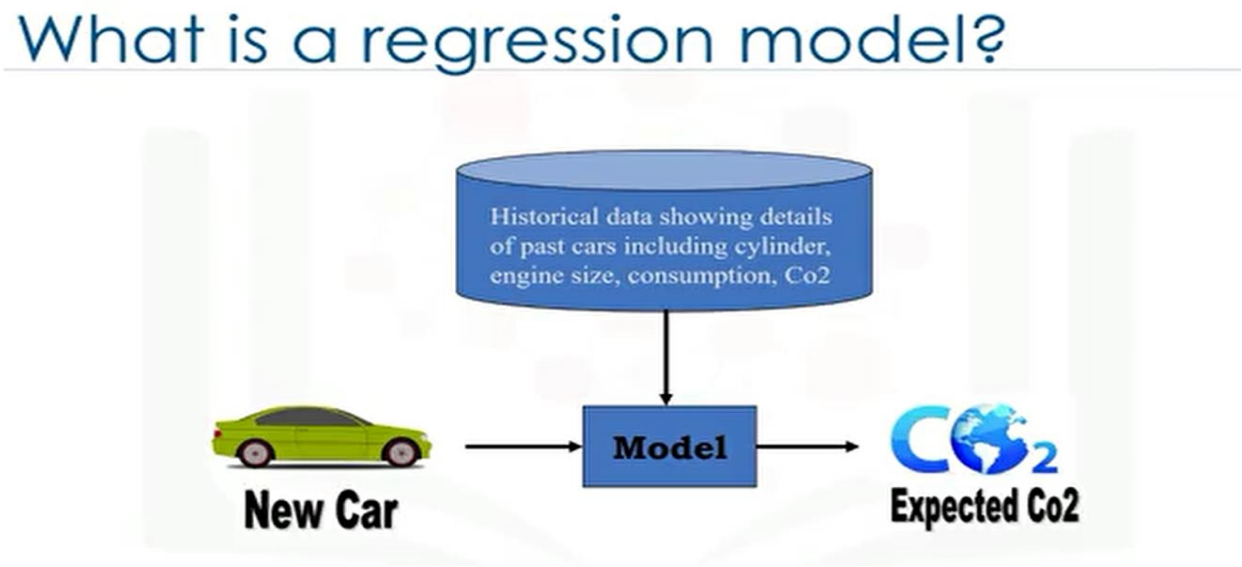
A regression model relates Y or the dependent variable to a function of X i.e. the independent variables.

Note: The key point in the regression is that our dependent value should be continuous and cannot be a discrete value.

However, the independent variable, or variables, can be measured on either a categorical or continuous measurement scale.

Goal: 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.

We use regression to build such a regression estimation model; then the model is used to predict the expected co2 emission for a new or unknown car.



There are two types of regression models:

- ✓ Simple regression
- ✓ Multiple regression.

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

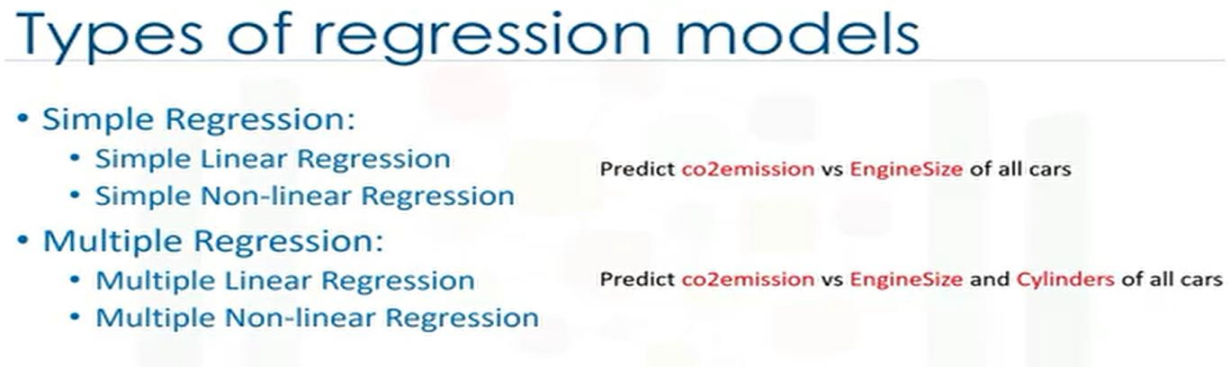
For example, predicting co2 emission using the variable of engine size.

Note: Linearity of regression is based on the nature of relationship between independent and dependent variables.

Multiple regression: When more than one independent variable is present the process is called multiple linear regression.

For example, predicting co2 emission using engine size and the number of cylinders in any given car.

Note: 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.

- **For instance**, one of the applications of regression analysis could be in the area of sales forecasting. We can try to predict a sales person's total yearly sales from independent variables such as age, education, and years of experience.

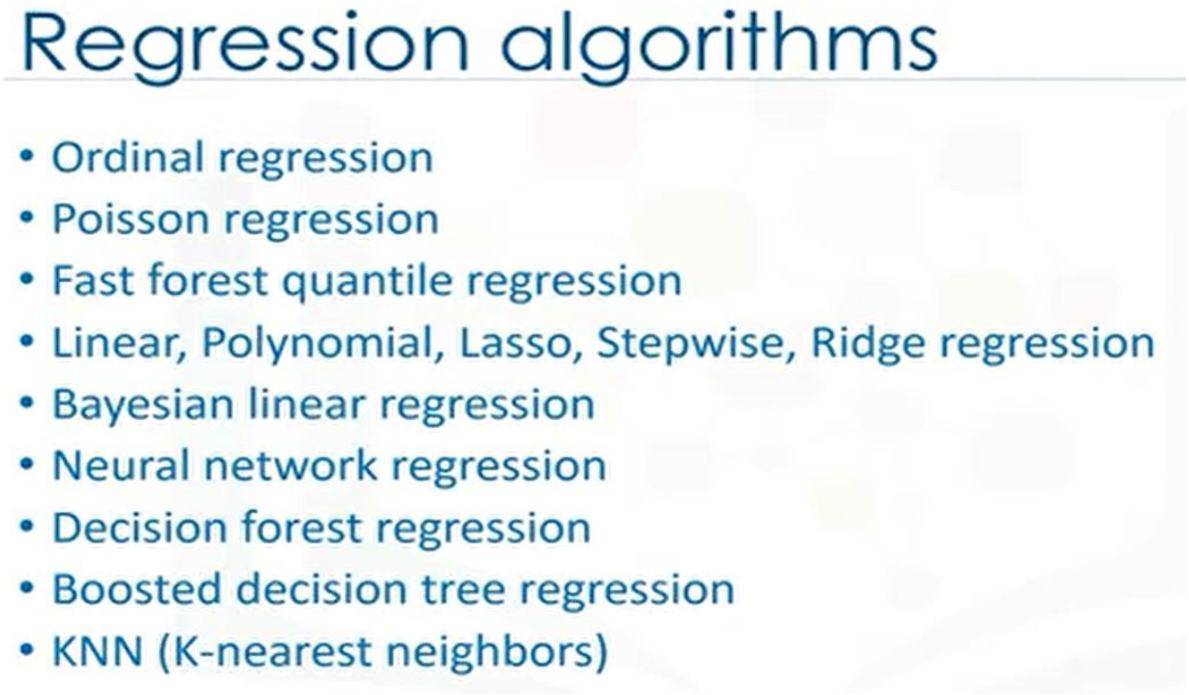
- It can also be used in the field of psychology, for example, to determine individual satisfaction, based on demographic and psychological factors.

- 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.

- 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.

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.

Note: We have many regression algorithms, each of them has its own importance and a specific condition to which their application is best suited.



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.

Week 2 index 2: Simple Linear Regression

Q: The question is, given this data set, 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.

Linear regression: Linear regression is the approximation of a linear model used to describe the relationship between two or more variables.

In simple linear regression, there are two variables,

- a dependent variable and
- an independent variable.

The key point in the linear regression is that our dependent value should be continuous and cannot be a discrete value.

Note: the independent variables can be measured on either a categorical or continuous measurement scale.

There are two types of linear regression models. They are:

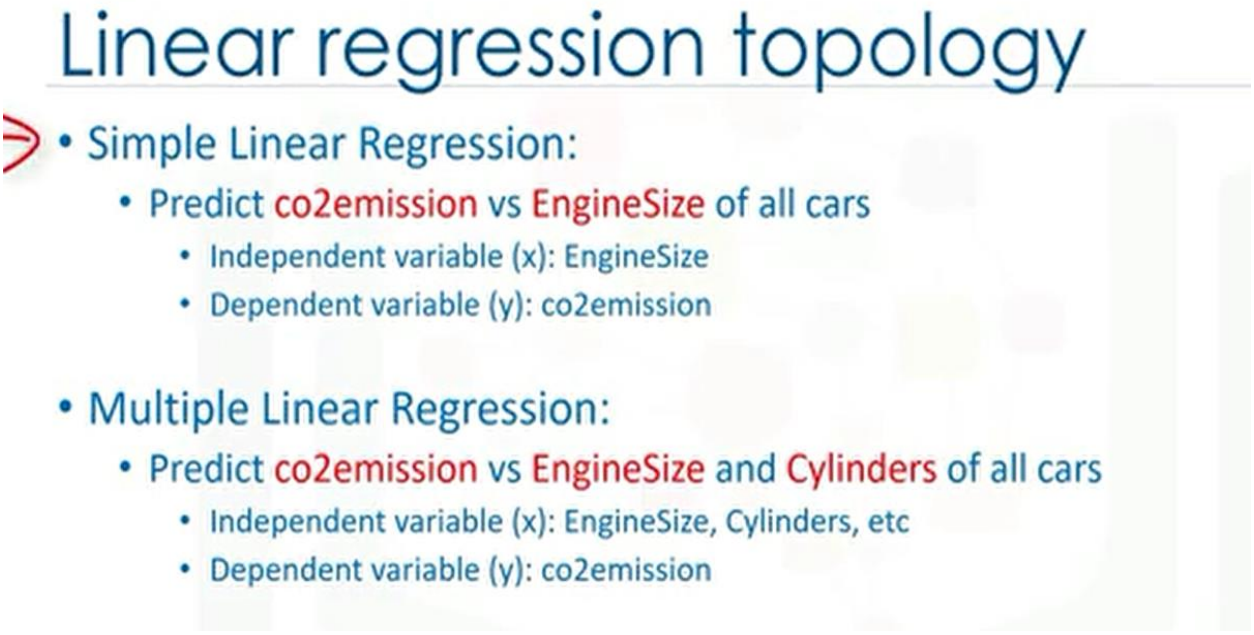
- Simple regression
- Multiple regression.

Simple linear: Simple linear regression is when one independent variable is used to estimate a dependent variable.

For example, predicting Co2 emission using the engine size variable.

Multiple linear: When more than one independent variable is present the process is called multiple linear regression.

For example, predicting Co2 emission using engine size and cylinders of cars.



Our focus in this video is on simple linear regression.

To understand linear regression, we can plot our variables here. We show engine size as an independent variable and emission as the target value that we would like to predict.

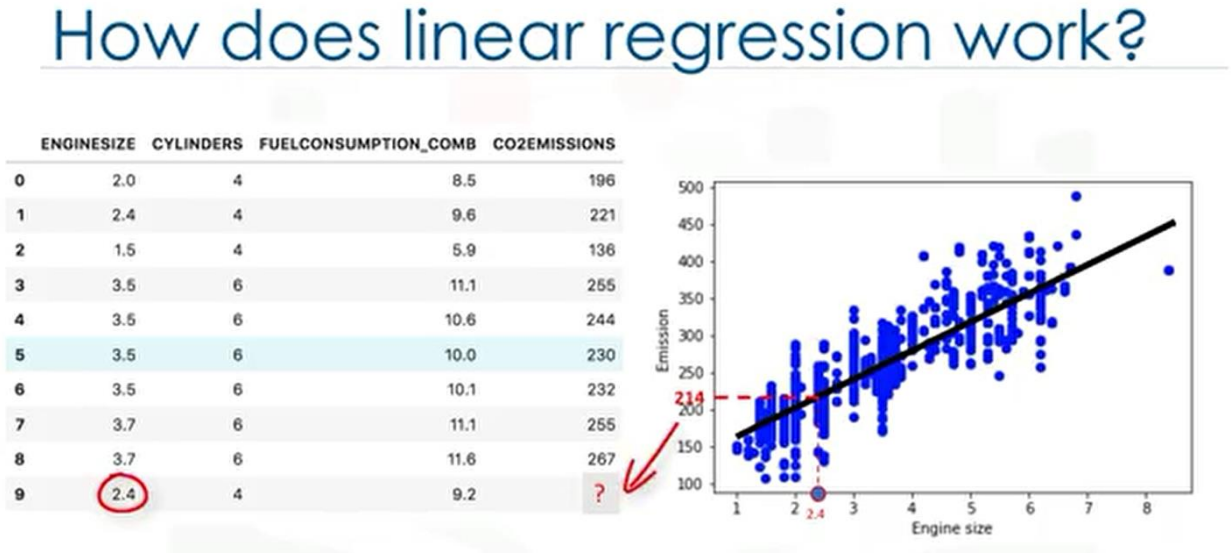
A scatter plot clearly shows the relation between variables where changes in one variable explain or possibly cause changes in the other variable.

Also, it indicates that these variables are linearly related. With linear regression you can fit a line through the data.

For instance, as the engine size increases, so do the emissions. With linear regression you can model the relationship of these variables.

A good model can be used to predict what the approximate emission of each car is.

For example, for a sample car with engine size 2.4, you can find the emission is 214.



Now, let's talk about what the fitting line actually is.

We're going to predict the target value y. In our case using the independent variable engine size represented by x1.

The fit line is shown traditionally as a polynomial.

$\hat{y} = \theta_0 + \theta_1 x_1$

In this equation,

- $\hat{y}$ is the dependent variable of the predicted value.
- x_1 is the independent variable.

Theta 0 and theta 1 are the parameters of the line that we must adjust.

- θ_1 is known as the slope or gradient of the fitting line and
- θ_0 is known as the intercept.

Note: θ_0 and θ_1 is also called the coefficients of the linear equation.

You can interpret this equation as $\hat{y}$ being a function of x_1, or $\hat{y}$ being dependent of x_1.

How would you draw a line through the points? And how do you determine which line fits best?

Linear regression estimates the coefficients of the line. This means we must calculate theta 0 and theta 1 to find the best line to fit the data. This line would best estimate the emission of the unknown data points.

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.

Now, let's go through all the points and check how well they align with this line.

Best fit here means that if we have, for instance, a car with engine size $x_1 = 5.4$ and actual $Co_2 = 250$, its Co_2 should be predicted very close to the actual value, which is $y = 250$ based on historical data.

But if we use the fit line, or better to say using our polynomial with known parameters to predict the Co_2 emission, it will return $\hat{y} = 340$.

Now if you compare the actual value of the emission of the car with what we've predicted using our model, you will find out that we have a 90-unit error.

This means our prediction line is not accurate. This error is also called the residual error. So, we can say the error is the distance from the data point to the fitted regression line.

The mean of all residual errors shows how poorly the line fits with the whole data set.

Mathematically it can be shown by the equation Mean Squared Error, shown as MSE.

Objective: Our objective is to find a line where the mean of all these errors is minimized. In other words, the mean error of the prediction using the fit line should be minimized.

Let's reword it more technically. The objective of linear regression, is to minimize this MSE equation and to minimize it, we should find the best parameters θ_0 and θ_1.

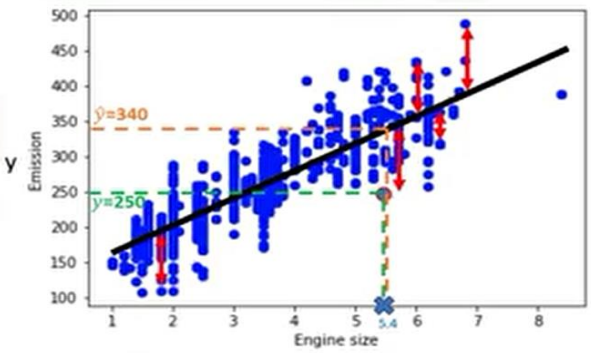
How to find the best fit?

$x_1 = 5.4$ independent variable
$y = 250$ actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$
$\hat{y} = 340$ the predicted emission of x1

Error = $y - \hat{y}$
$= 250 - 340$
$= -90$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



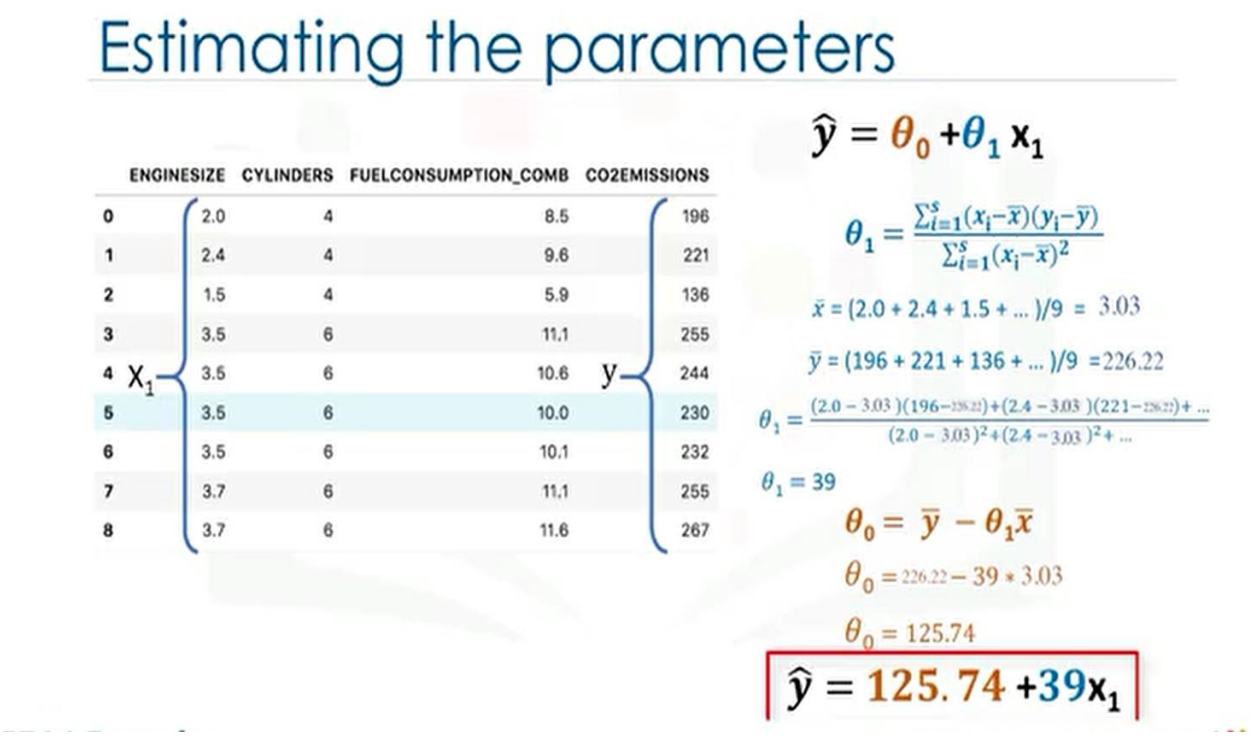
Big Question: Now the question is how to find θ_0 and θ_1 in such a way that it minimizes this error?

we have two options here.

- Option one, we can use a mathematic approach, or
- option two, we can use an optimization approach.

θ_0 and θ_1 in the simple linear regression are the coefficients of the fit line. We can use a simple equation to estimate these coefficients.

Let's see how we could easily use a mathematic formula to find the θ_0 and θ_1



θ_0 and θ_1 are the intercept and slope of the line; we can estimate them directly from our data.

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

We can start off by estimating the value for θ_1. This is how you can find the slope of a line based on the data. $\bar{x}$ is the average value for the engine size and $\bar{y}$ is the average value for the CO_2 emission in our data set.

First, we calculate the average of x_1 and of y, then we plug it into the slope equation to find θ_1.

The x_i and y_i in the equation refer to the fact that we need to repeat these calculations across all values in our data set. And i refers to the ith value of x or y. Applying all values, we find θ_1=39. It is our second parameter(slope). It is used to calculate the first parameter which is the intercept (θ_0) of the line.

Now we can plug θ_1 into the line equation to find θ_0. It is easily calculated that θ_0 =125.74. So, these are the two parameters for the line, where θ_0 is also called the bias coefficient, and $\theta_1$is the coefficient for the Co2 emission column.

Imagine we are predicting CO_2 emission, or y, from engine-size, or x for the automobile in record number 9. Our linear regression model representation for this problem would be $\hat{y}= \theta_0+ \theta_1 x_1$. Or if we map it to our data set, it would be CO_2Emission = $\theta_0+ \theta_1$Engine-Size.

For example,

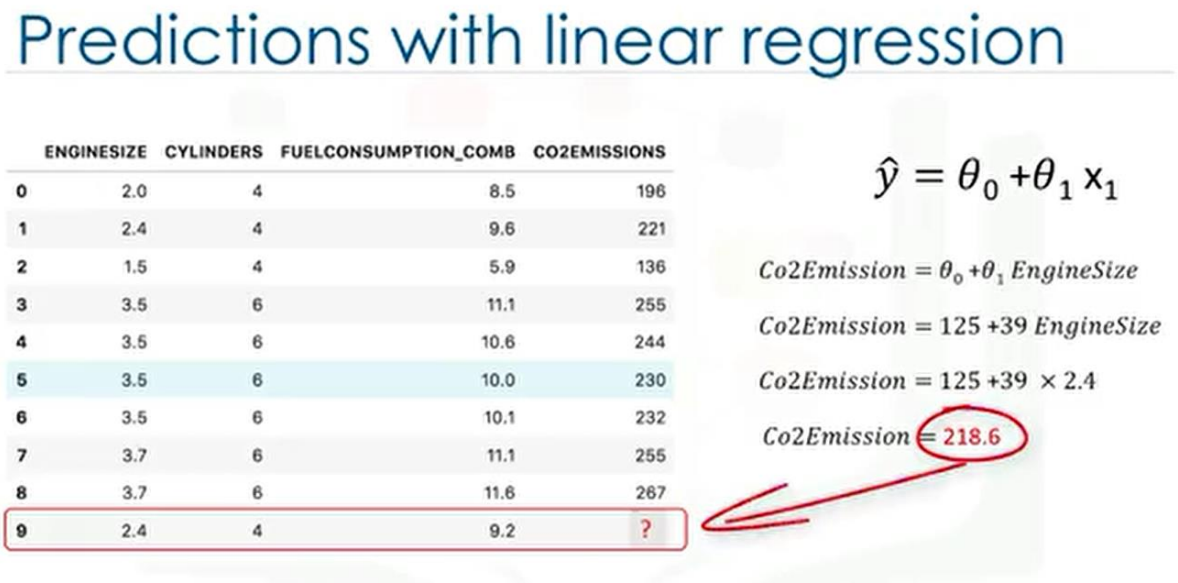
let's use θ_0 =125.74 and

θ_1=39

Engine-Size=2.4

CO2Emission = 125 + 39 * 2.4

= 218.6

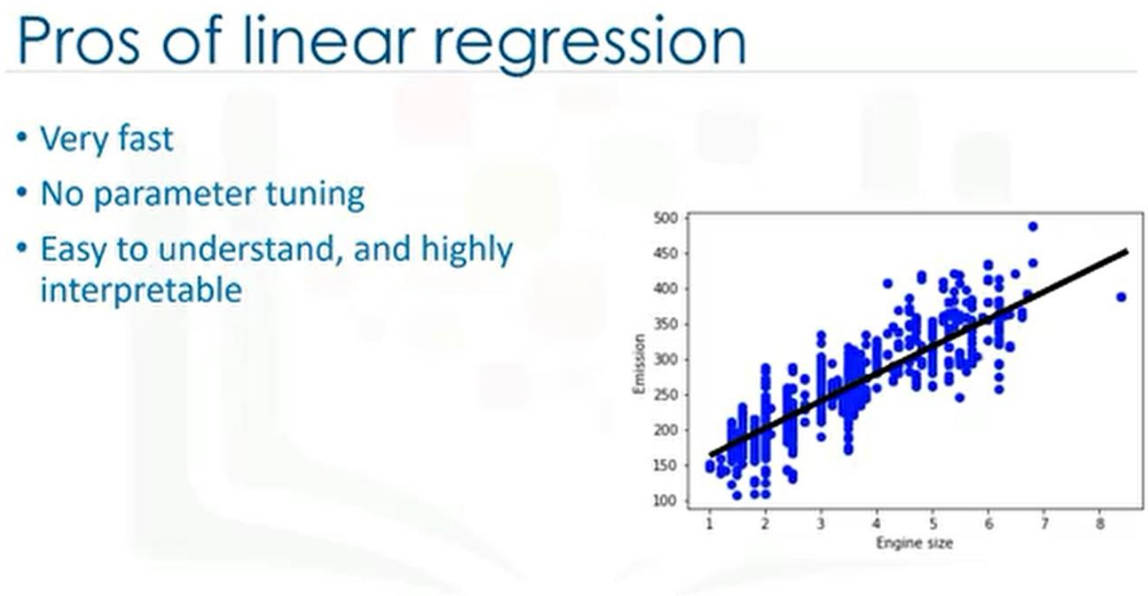


Therefore, we can predict that the Co2Emission for this specific car would be 218.6.

Let's talk a bit about why linear regression is so useful. Quite simply, it is the most basic regression to use and understand.

- it's fast.
- **It also doesn't require tuning of parameters.** So, something like tuning the K parameter and K nearest neighbors, or the learning rate in neural networks isn't something to worry about.

- Linear regression is also easy to understand, and highly interpretable.



Week 2 index 3: Model Evaluation in Regression Models

Here, we'll be covering model evaluation. So, let's get started.

Note: The goal of regression is to build a model to accurately predict an unknown case.

To this end, we have to perform regression evaluation after building the model.

we'll introduce and discuss two types of evaluation approaches that can be used to achieve this goal.

These approaches are

- train and test on the same dataset and
- train/test split.

We'll talk about what each of these are, as well as the **pros and cons** of using each of these models.

Also, we'll introduce some metrics for accuracy of regression models

Let's look at the first approach.

When considering evaluation models, we clearly want to choose the one that will give us the most accurate results.

Q: how can we calculate the accuracy of our model?

In other words, how much can we trust this model for prediction of an unknown sample using a given dataset and having built a model such as linear regression?

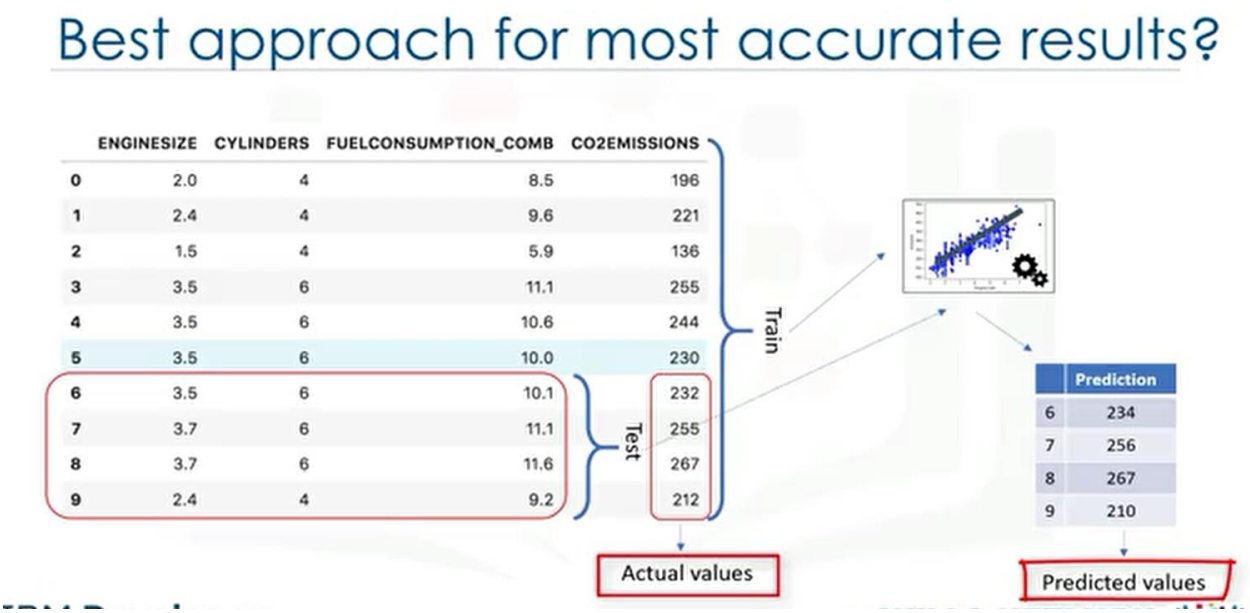
One of the solutions is to select a portion of our dataset for testing.

For instance, assume that we have **10 records in our dataset.**

We use the entire dataset for training, and we build a model using this training set. Now, we select a small portion of the dataset, such as row number **six to nine**, but without the labels. This set is called a test set,

Note: which has the labels, but the labels are not used for prediction and is used only as ground truth. **The labels are called actual values of the test set**.

Now we pass the feature set of the testing portion to our built model and predict the target values. Finally, we compare the predicted values by our model with the actual values in the test set. This indicates how accurate our model actually is.

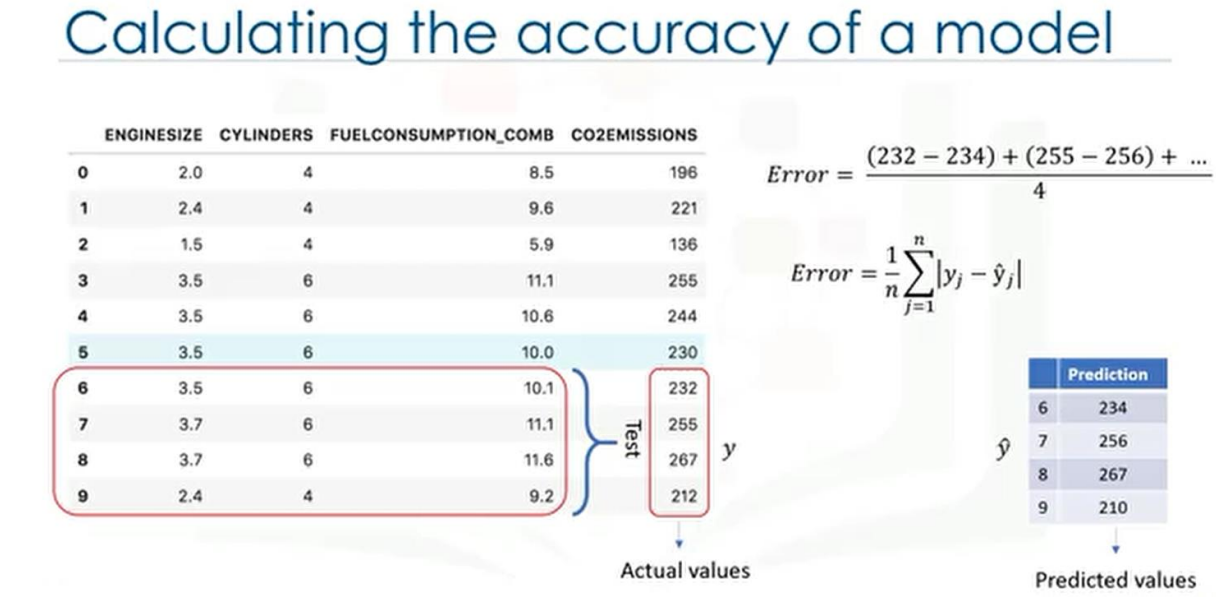


There are different metrics to report the accuracy of the model, but most of them work generally based on the similarity of the predicted and actual values.

Let's look at one of the simplest metrics to calculate the accuracy of our regression model.

As mentioned, we just compare the actual values y with the predicted values, which is noted as $\hat{y}$ for the testing set.

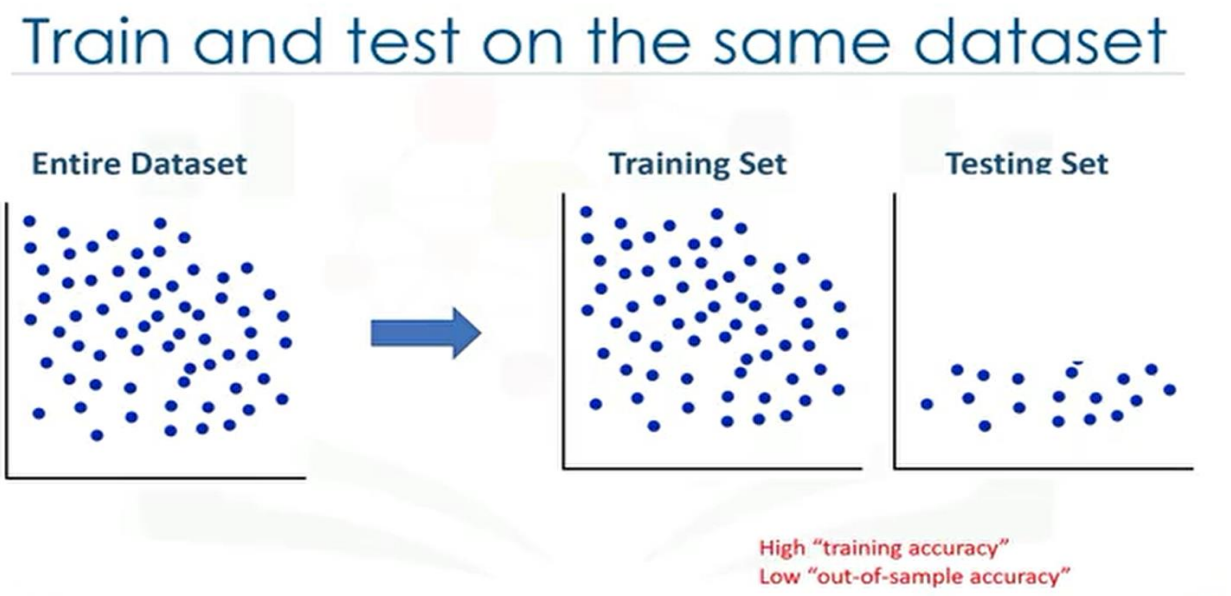
The error of the model is calculated as the average difference between the predicted and actual values for all the rows.



So, the first evaluation approach we just talked about is the simplest one, train and test on the same dataset. We train the model on the entire dataset, then you test it using a portion of the same dataset.

In a general sense, when you test with a dataset in which you know the target value for each data point, you're able to obtain a percentage of accurate predictions for the model.

This evaluation approach would most likely have a high training accuracy and the low out-of-sample accuracy since the model knows all of the testing data points from the training.



Q: What is training accuracy and out-of-sample accuracy?

We said that training and testing on the same dataset produces a high training accuracy, but what exactly is training accuracy?

Training accuracy is the percentage of correct predictions that the model makes when using the test dataset.

However, a high training accuracy isn't necessarily a good thing.

For instance, having a high training accuracy may result in an over-fit the data. This means that the model is overly trained to the dataset, which may capture noise and produce a non-generalized model.

Out-of-sample accuracy is the percentage of correct predictions that the model makes on data that the model has not been trained on.

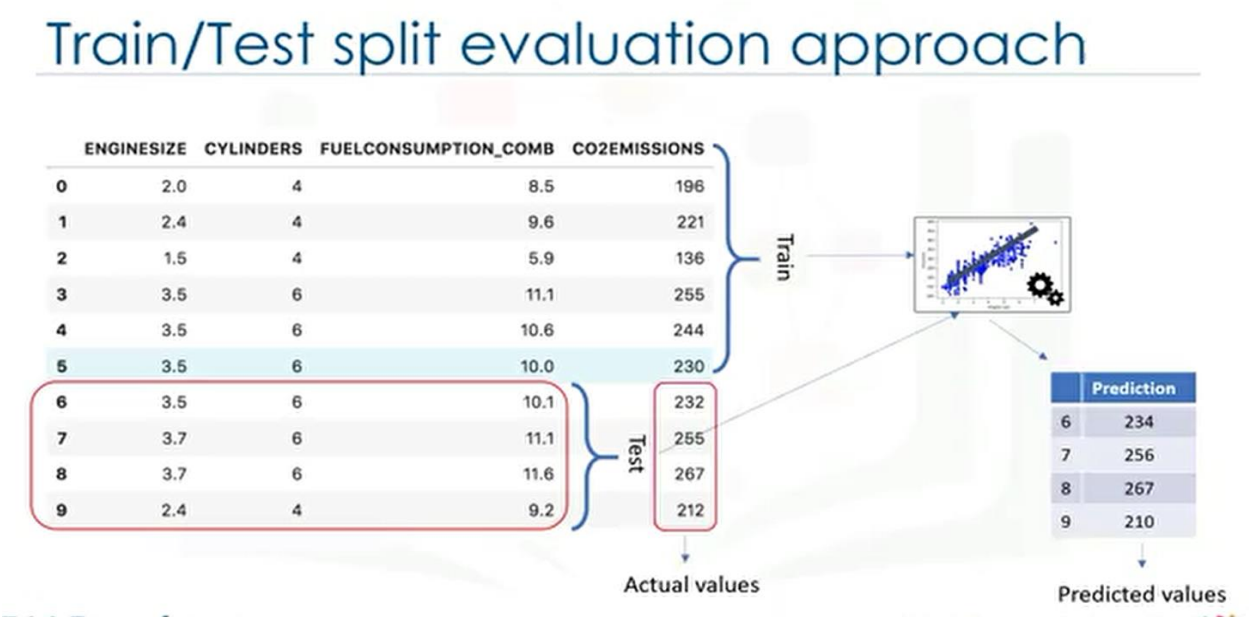
Note: Doing a train and test on the same dataset will most likely have low out-of-sample accuracy due to the likelihood of being over-fit.

It's important that our models have high out-of-sample accuracy because the purpose of our model is, of course, to make correct predictions on unknown data.

train/test split

So, how can we improve out-of-sample accuracy? One way is to use another evaluation approach called train/test split

In this approach, we select a portion of our dataset for training, for example, row zero to five, and the rest is used for testing, for example, row six to nine.



The model is built on the training set. Then, the test feature set is passed to the model for prediction.

Finally, the predicted values for the test set are compared with the actual values of the testing set. The second evaluation approach is called train/test split.

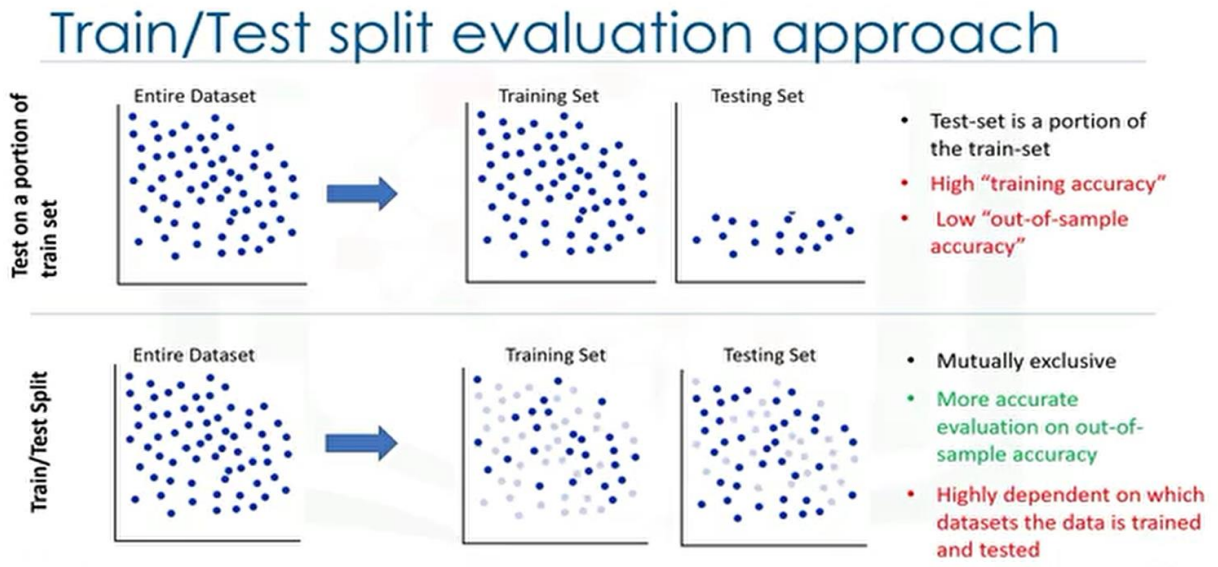
Train/test split involves splitting the dataset into training and testing sets respectively, which are mutually exclusive.

After which, We train with the training set and test with the testing set.

This will provide a more accurate evaluation on out-of-sample accuracy because the testing dataset is not part of the dataset that has been used to train the data.

It is more realistic for real-world problems. This means that we know the outcome of each data point in the dataset, making it great to test with.

Since this data has not been used to train the model, the model has no knowledge of the outcome of these data points.

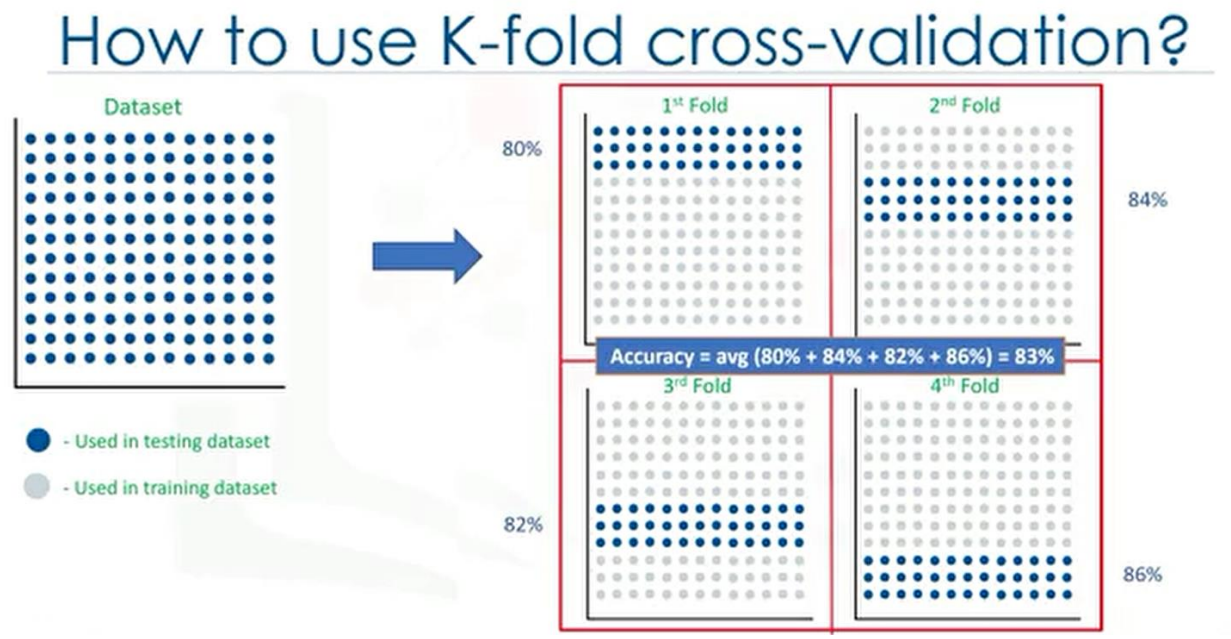


So, in essence, it's truly out-of-sample testing. However, please ensure that you train your model with the testing set afterwards, as you don't want to lose potentially valuable data.

The issue with train/test split is that it's highly dependent on the datasets on which the data was trained and tested.

The variation of this causes train/test split to have a better out-of-sample prediction than training and testing on the same dataset, but it still has some problems due to this dependency.

K-fold cross-validation



Another evaluation model, called K-fold cross-validation, resolves most of these issues.

Q: How do you fix a high variation that results from a dependency?

Well, we average it.

If we have K equals four folds, then we split up this dataset as shown here.

In the first fold for example, we use the first 25 percent of the dataset for testing and the rest for training.

The model is built using the training set and is evaluated using the test set.

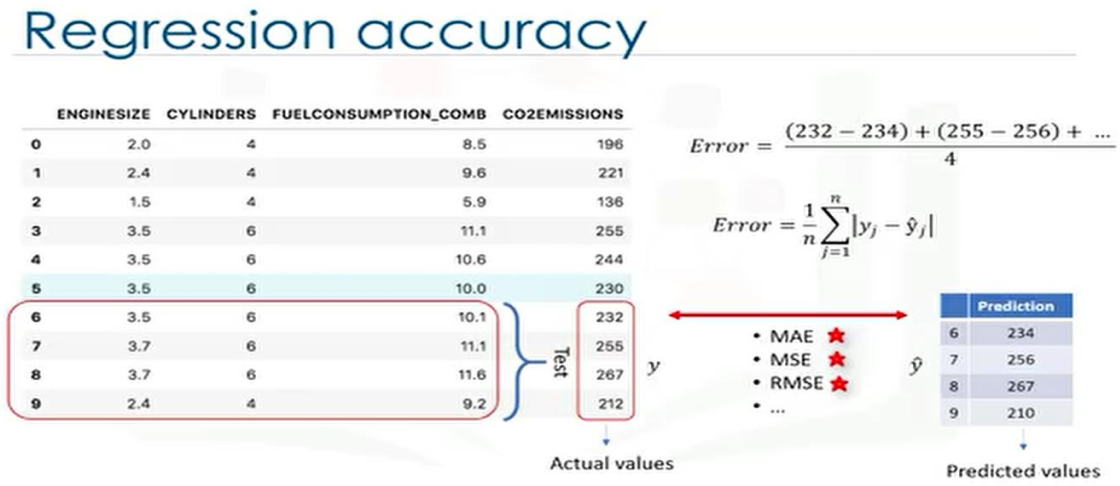
Then, in the next round or in the second fold, the second 25 percent of the dataset is used for testing and the rest for training the model. Again, the accuracy of the model is calculated. We continue for all folds. Finally, the result of all four evaluations are averaged.

Note: That is, the accuracy of each fold is then averaged, keeping in mind that each fold is distinct, where no training data in one-fold is used in another.

K-fold cross-validation in its simplest form performs multiple train/test splits, using the same dataset where each split is different. Then, the result is average to produce a more consistent out-of-sample accuracy.

Week 2 index 4: Evaluation Metrics in Regression Models

In this video, we'll be covering accuracy metrics for model evaluation.



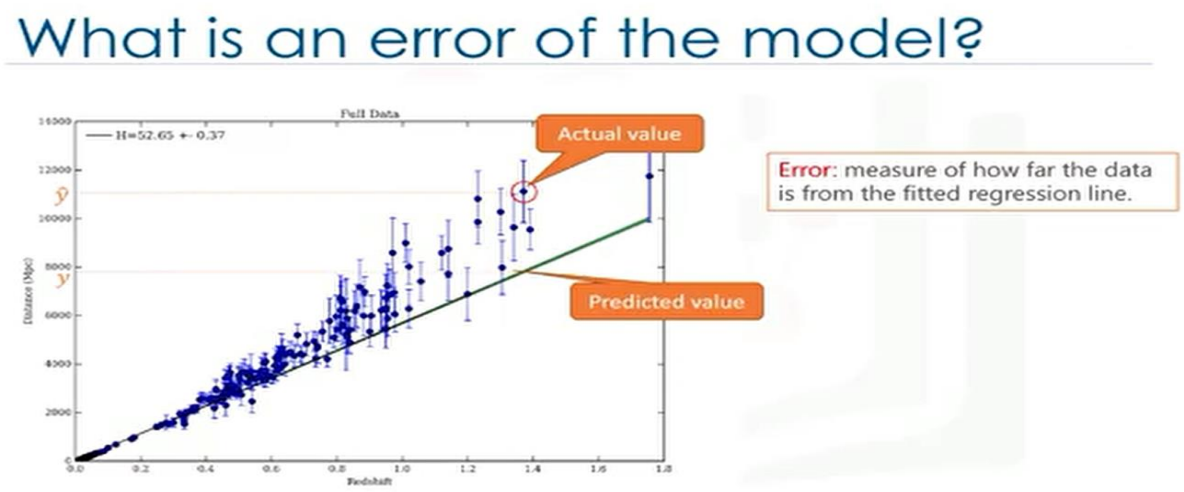
let's get started. Evaluation metrics are used to explain the performance of a model. Let's talk more about the model evaluation metrics that are used for regression.

basically, we can compare the actual values and predicted values to calculate the accuracy of our regression model.

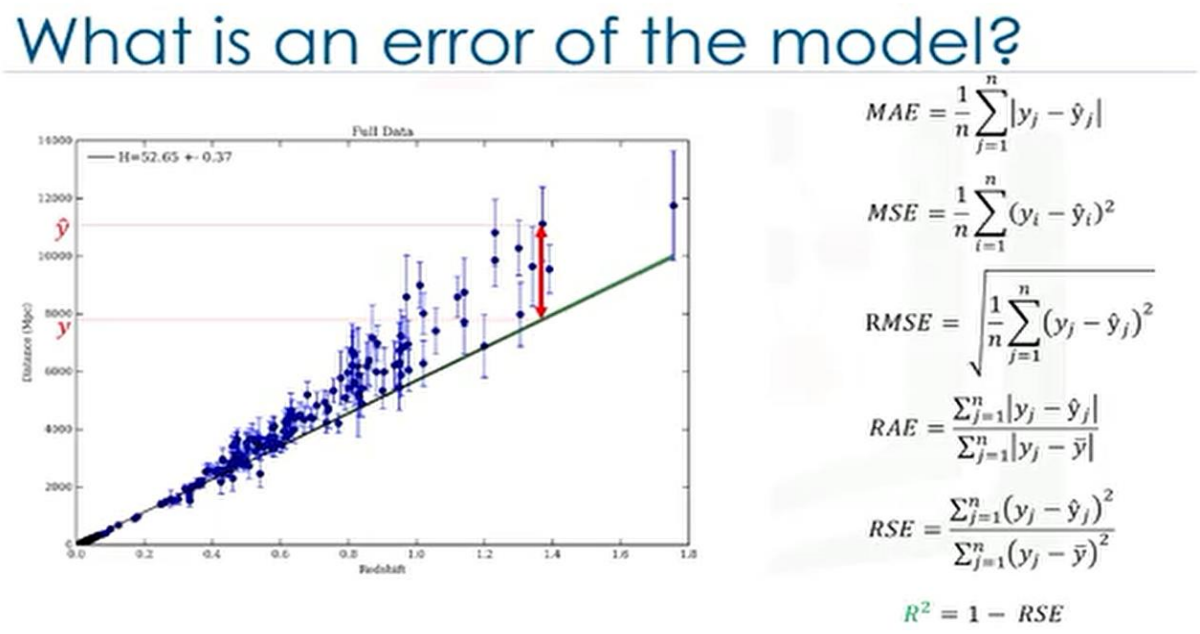
Evaluation metrics provide a key role in the development of a model as it provides insight to areas that require improvement.

We'll be reviewing a number of model evaluation metrics, including Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, but before we get into defining these, we need to define what an error actually is.

Error: In the context of regression, the error of the model is the difference between the data points and the trend line generated by the algorithm.



Since there are multiple data points, an error can be determined in multiple ways.



MAE: Mean Absolute Error is the mean of the absolute value of the errors. This is the easiest of the metrics to understand, since it's just the average error.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

MSE: Mean Squared Error is the mean of the squared error. It's more popular than Mean Absolute Error because the focus is geared more towards large errors. This is due to the squared term, exponentially increasing larger errors in comparison to smaller ones.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

RMSE: Root Mean Squared Error is the square root of the mean squared error. This is one of the most popular of the evaluation metrics because Root Mean Squared Error is interpretable in the same units as the response vector or Y units, making it easy to relate its information.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

RAE: Relative absolute error, also known as residual sum of square, where $\bar{y}$ is a mean value of Y, takes the total absolute error and normalizes it. By dividing by the total absolute error of the simple predictor.

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

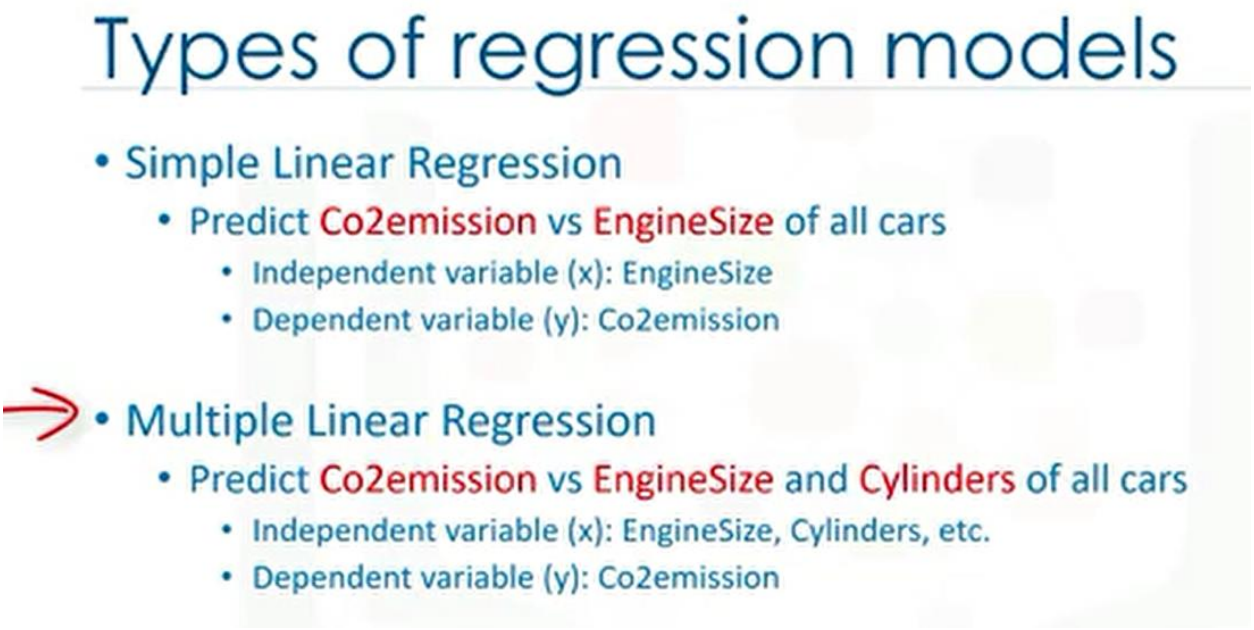
R^2: R-squared is not an error per say but is a popular metric for the accuracy of your model. It represents how close the data values are to the fitted regression line. The higher the R-squared, the better the model fits your data.

$$R^2 = 1 - RSE$$

Each of these metrics can be used for quantifying of your prediction. The choice of metric completely depends on the type of model your data type and domain of knowledge.

Week 2 index 5: Multiple Linear Regression

There are multiple variables that predict the CO_2 emission. When multiple independent variables are present, the process is called multiple linear regression. **For example**, predicting CO_2 emission using engine size and the number of cylinders in the car's engine.



The good thing is that multiple linear regression is the extension of the simple linear regression model.

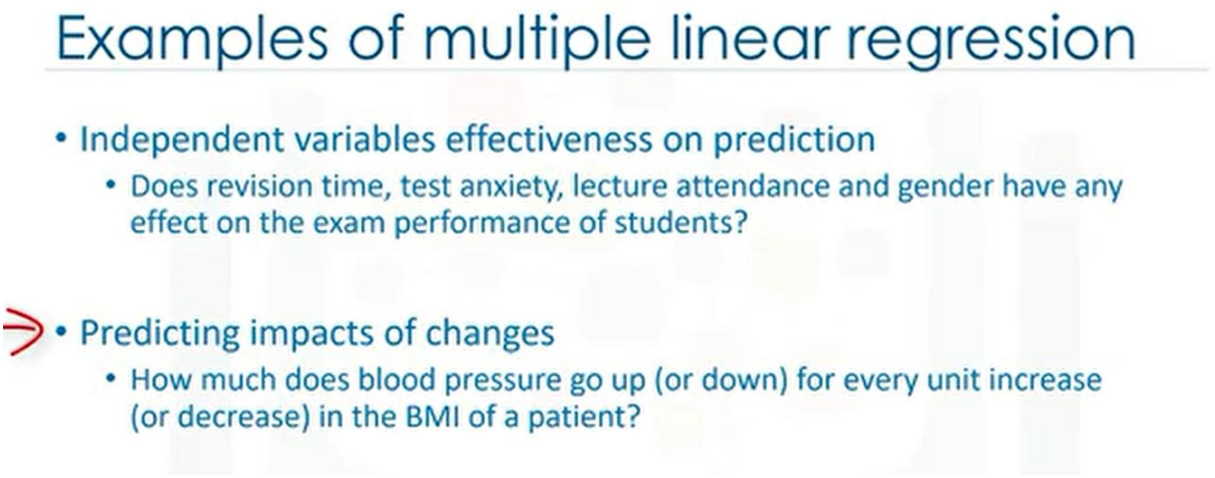
Basically, there are two applications for multiple linear regression.

First, it can be used when we would like to identify the strength of the effect that the independent variables have on the dependent variable.

For example, does revision time, test anxiety, lecture attendance and gender have any effect on exam performance of students?

Second, it can be used to predict the impact of changes, that is, to understand how the dependent variable changes when we change the independent variables.

For example, if we were reviewing a person's health data, a multiple linear regression can tell you how much that person's blood pressure goes up or down for every unit increase or decrease in a patient's body mass index holding other factors constant.



As is the case with simple linear regression, multiple linear regression is a method of predicting a continuous variable.

It uses multiple variables called independent variables or predictors that best predict the value of the target variable which is also called the dependent variable.

In multiple linear regression, the target value Y, is a linear combination of independent variables X.

For example, you can predict how much CO_2 a car might admit due to independent variables such as the car's engine size, number of cylinders, and fuel consumption.

.

- ✓ Multiple linear regression is very useful because you can examine which variables are significant predictors of the outcome variable.
- ✓ Also, you can find out how each feature impacts the outcome variable.

Generally, the model is of the form $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots. + \theta_n x_n$

Mathematically, we can show it as a vector form $\hat{y} = \theta^T x$

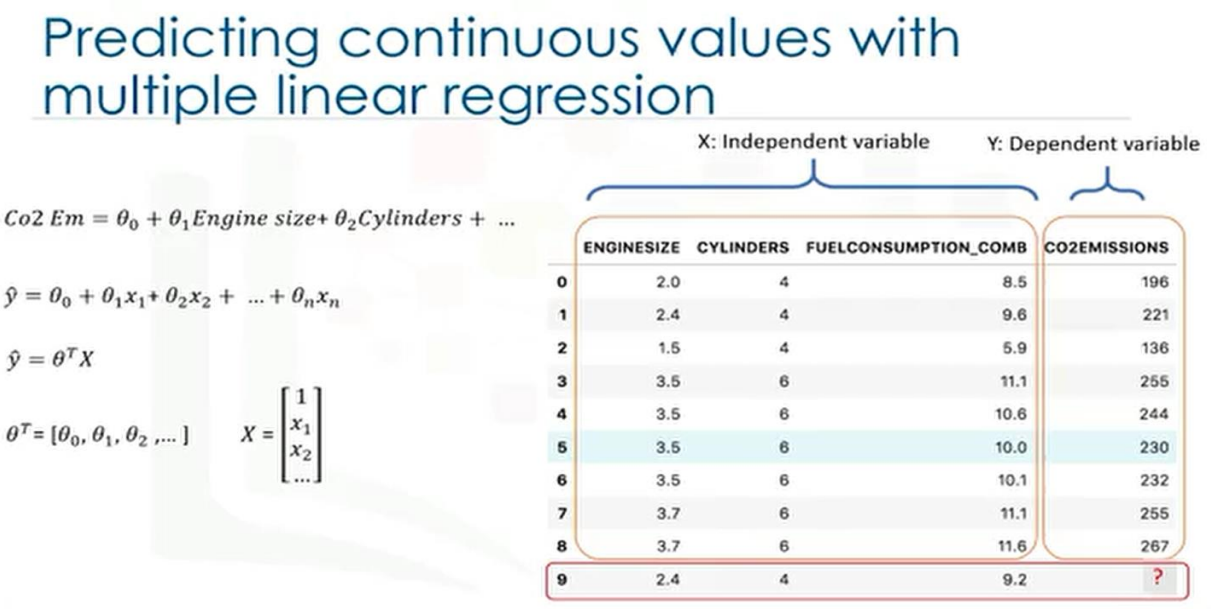
This means it can be shown as a dot product of two vectors;

- ✓ the parameters vector and
- ✓ the feature set vector.

Generally, we can show the equation for a multidimensional space as **theta transpose** x, where theta is an n by one vector of unknown parameters in a multi-dimensional space, and x is the vector of the featured sets, as theta is a vector of coefficients and is supposed to be multiplied by x.

Conventionally, it is shown as transpose theta [$\theta^T = \theta_0, \theta_1, \theta_2,.....$]. Theta is also called the parameters or weight vector of the regression equation. Both these terms can be used interchangeably, and x is the feature set which represents a car. $x = \begin{vmatrix} 1 \\ x_1 \\ x_2 \end{vmatrix}$

For example, $x_1$for engine size or x_2 for cylinders, and so on.



The first element of the feature set would be set to one, because it turns that theta zero into the intercept or biased parameter when the vector is multiplied by the parameter vector.

Note: theta transpose x in a one-dimensional space is the equation of a line, it is what we use in simple linear regression.

In higher dimensions when we have more than one input or x the line is called a plane or a hyperplane, and this is what we use for multiple linear regression.

So, the whole idea is to find the best fit hyperplane for our data. To this end and as is the case in linear regression, we should estimate the values for theta vector that best predict the value of the target field in each row. To achieve this goal, we have to minimize the error of the prediction.

Now, the **question** is, how do we find the optimized parameters?

To find the optimized parameters for our model, we should first understand what the optimized parameters are, then we will find a way to optimize the parameters. In short, optimized parameters are the ones which lead to a model with the fewest errors.

Let's assume for a moment that we have already found the parameter vector of our model; it means we already know the values of theta vector.

Now we can use the model and the feature set of the first row of our dataset to predict the CO_2 emission for the first car, correct?
If we plug the feature set values into the model equation, we find $\hat{y}$.

Let's say for example,

it returns $\hat{y}_i$=140 as the predicted value for this specific row,

what is the actual value? y_i=196

here, $y_i - \hat{y}_i = 196 - 140 = 56$ [residual error]

This is the error of our model only for one row or one car in our case.

Using MSE to expose the errors in the model

$\hat{y} = \theta^T X$

$\hat{y}_i = 140$ the predicted emission of x_i

$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

As is the case in linear regression, we can say the error here is the distance from the data point to the fitted regression model.

The mean of all residual errors shows how bad the model is representing the data set, it is called the mean squared error, or MSE.

Note: While this is not the only way to expose the error of a multiple linear regression model, it is one of the most popular ways to do so.

Note: The best model for our data set is the one with minimum error for all prediction values. So, the objective of multiple linear regression is to minimize the MSE equation.

To minimize it, we should find the best parameters theta, but how?

Okay, how do we find the parameter or coefficients for multiple linear regression? There are many ways to estimate the value of these coefficients.

However, the most common methods are

- ✓ ordinary least squares and
- ✓ optimization approach

Ordinary least squares: Ordinary least squares try to estimate the values of the coefficients by minimizing the mean square error.

This approach uses the data as a matrix and uses linear algebra operations to estimate the optimal values for the theta.

Problem: The problem with this technique is the time complexity of calculating matrix operations as it can take a very long time to finish.

When the number of rows in your data set is less than 10,000, you can think of this technique as an option. However, for greater values, you should try other faster approaches.

Optimization: The second option is to use an optimization algorithm to find the best parameters. That is, you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

For example, you can use gradient descent which starts optimization with random values for each coefficient, then calculates the errors and tries to minimize it through y's changing of the coefficients in multiple iterations.

Note: Gradient descent is a proper approach if you have a large data set.

Please understand however, that there are other approaches to estimate the parameters of the multiple linear regression that you can explore on your own.

Estimating multiple linear regression parameters

- How to estimate θ?
  - Ordinary Least Squares
    - Linear algebra operations
    - Takes a long time for large datasets (10K+ rows)
  - An optimization algorithm
    - Gradient Descent
    - Proper approach if you have a very large dataset

After you find the best parameters for your model, you can go to the prediction phase.

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 CO_2 emission or Y from other variables for the automobile in record number nine.

Our linear regression model representation for this problem would be $\hat{y} = \theta^T x$. Once we find the parameters, we can plug them into the equation of the linear model

Making predictions with multiple linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$\hat{y} = \theta^T X$

$\theta^T = [125, 6.2, 14, \ldots]$

$\hat{y} = 125 + 6.2x_1 + 14x_2 +$

$Co2Em = 125 + 6.2EngSize + 14\ \text{Cylinders} + \ldots$

$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 + \ldots$

$Co2Em = 214.1$

As you can see, multiple linear regression estimates the relative importance of predictors.

Now, let me address some concerns that you might already be having regarding multiple linear regression.

As you saw, you can use multiple independent variables to predict a target value in multiple linear regression. It sometimes results in a better model compared to using a simple linear regression which uses only one independent variable to predict the dependent variable.

Now the **question** is

- ✓ how many independent variables should we use for the prediction?
- ✓ Should we use all the fields in our data set?
- ✓ Does adding independent variables to a multiple linear regression model always increase the accuracy of the model?

Basically, adding too many independent variables without any theoretical justification may result in an overfit model. An overfit model is a real problem because it is too complicated for your data set and not general enough to be used for prediction. So, it is recommended to avoid using many variables for prediction.

The next **question** is, should independent variables be continuous?

Basically, categorical independent variables can be incorporated into a regression model by converting them into numerical variables.

For example, given a binary variable such as car type, the code dummy zero for manual and one for automatic cars.

Note: As a last point, remember that multiple linear regression is a specific type of linear regression. So, there needs to be a linear relationship between the dependent variable and each of your independent variables.

There are a number of ways to check for linear relationship.

For example, you can use scatter plots and then visually checked for linearity. If the relationship displayed in your scatter plot is not linear, then you need to use non-linear regression

Week 2 index 6: Non-Linear Regression

In this video, we'll be covering non-linear regression basics. So, let's get started. These data points correspond to China's gross domestic product or GDP from 1960-2014. The first column is the years and the second is China's corresponding annual gross domestic income in US dollars for that year.

This is what the data points look like.

Now, we have a couple of interesting questions.

- ✓ First, can GDP be predicted based on time?
- ✓ Second, can we use a simple linear regression to model it?

Indeed. If the data shows a curvy trend, then linear regression would not produce very accurate results when compared to a non-linear regression.

Simply because, as the name implies, linear regression presumes that the data is linear.

The scatter plot shows that there seems to be a strong relationship between GDP and time, but the relationship is not linear.

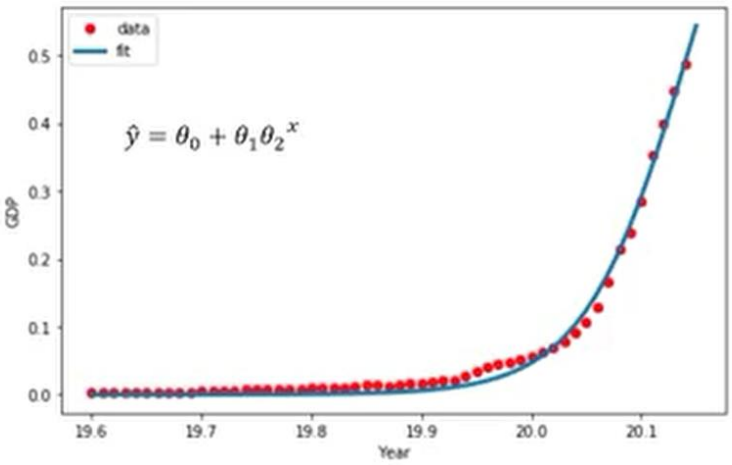
As you can see, the growth starts off slowly, then from 2005 onward, the growth is very significant. Finally, it decelerates slightly in the 2010s. It looks like either a logistic or exponential function.

So, it requires a special estimation method of the non-linear regression procedure.

For example, if we assume that the model for these data points are exponential functions, such as $\hat{y} = \theta_o + \theta_1 \theta_2^x$ (transpose X or to the power of X),

Should we use linear regression?

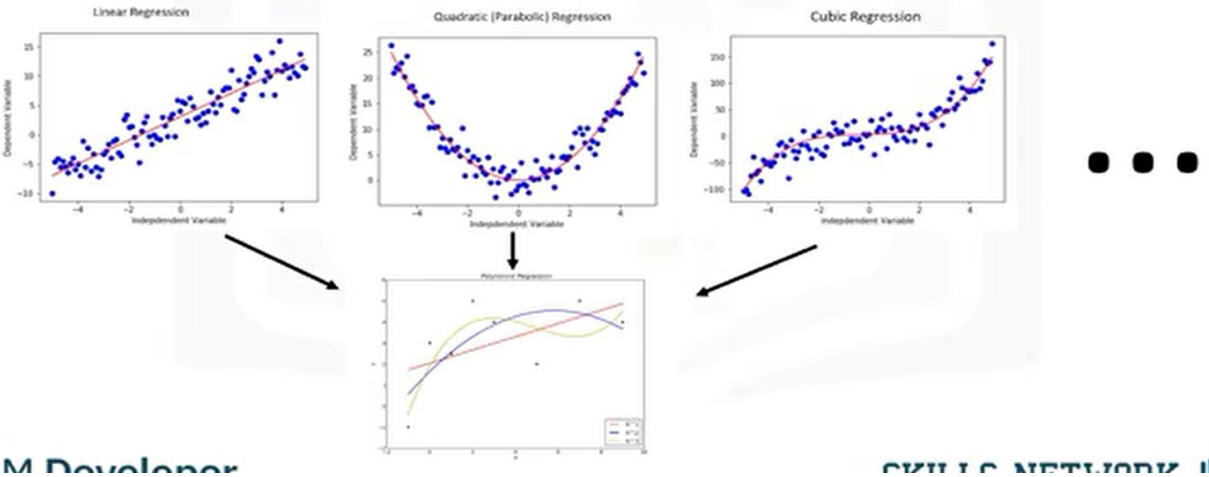
	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10
...	.....	.....



Our job is to estimate the parameters of the model, i.e., Thetas, and use the fitted model to predict GDP for unknown or future cases.

In fact, many different regressions exist that can be used to fit whatever the dataset looks like. You can see a quadratic and cubic regression lines here, and it can go on and on to infinite degrees.

Different types of regression



In essence, we can call all of these polynomial regressions, where the relationship between the independent variable X and the dependent variable Y is modeled as an

Nth degree polynomial in X. With many types of regression to choose from, there's a good chance that one will fit your dataset well. Remember, it's important to pick a regression that fits the data the best.

So, what is polynomial regression?

Polynomial regression fits a curve line to your data. A simple example of polynomial with degree three is shown as $\hat{y} = \theta_o + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$

where Thetas are parameters to be estimated that makes the model fit perfectly to the underlying data.

Note: Though the relationship between X and Y is non-linear here and polynomial regression can't fit them, a polynomial regression model can still be expressed as linear regression.

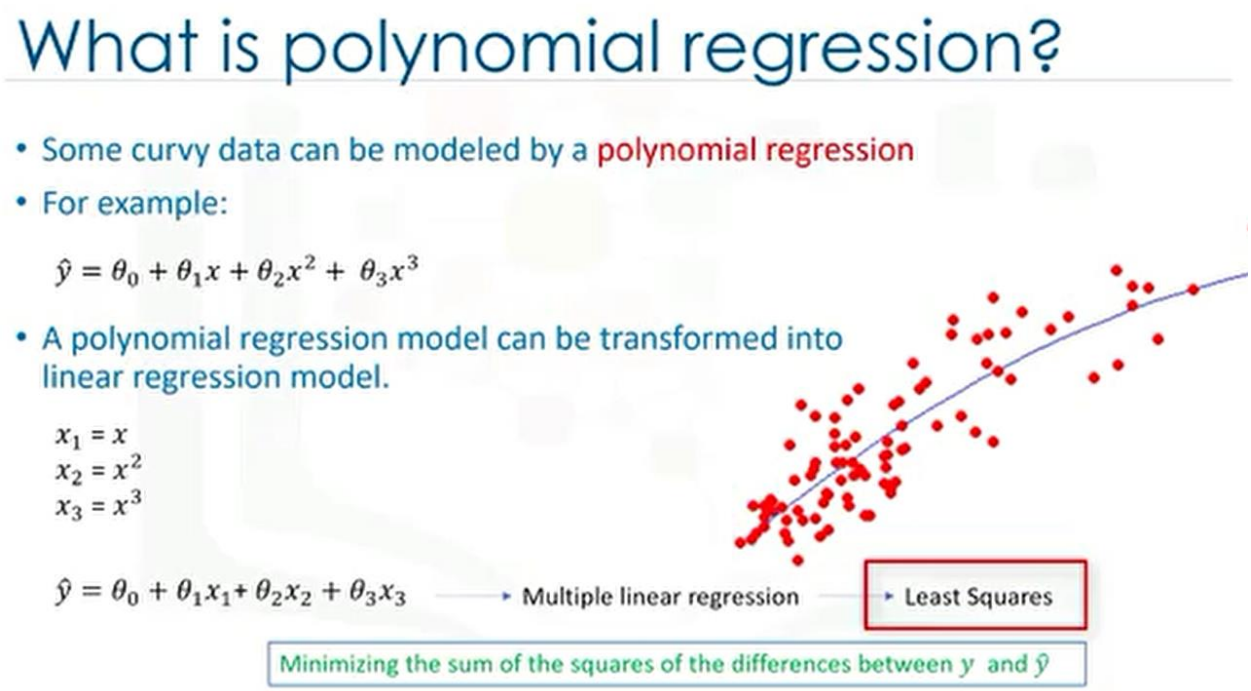
I know it's a bit confusing, but let's look at an example. Given the third degree polynomial equation, by defining $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$

The model is converted to a simple linear regression with new variables as $\hat{y} = \theta_o + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$

This model is linear in the parameters to be estimated, right?

Therefore, this polynomial regression is considered to be a special case of traditional multiple linear regression. So, you can use the same mechanism as linear regression to solve such a problem.

Therefore, polynomial regression models can fit using the model of least squares.
Least squares: Least squares is a method for estimating the unknown parameters in a linear regression model by minimizing the sum of the squares of the differences between the observed dependent variable in the given dataset and those predicted by the linear function.



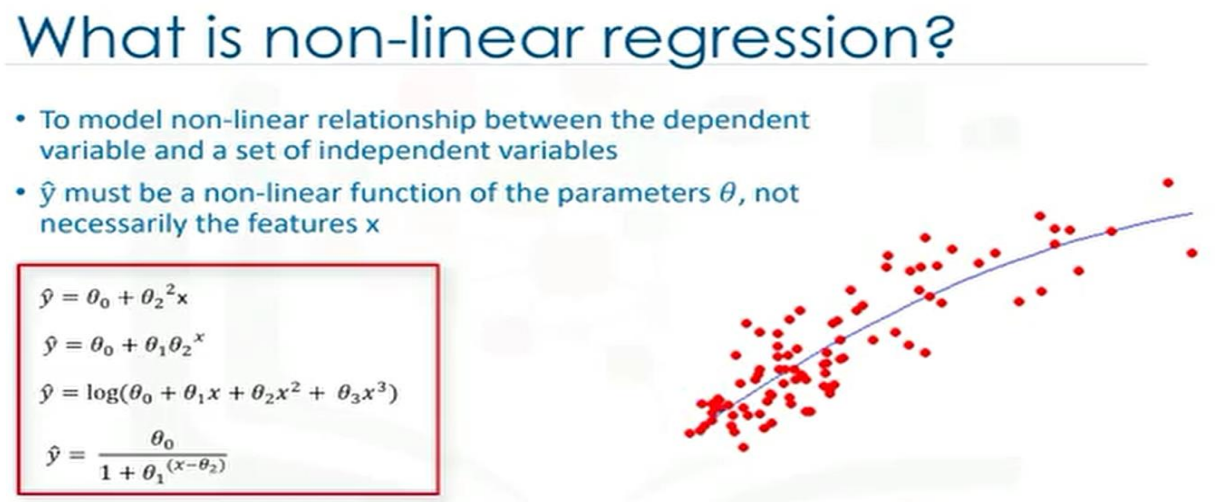
So, what is non-linear regression exactly?

- ✓ First, non-linear regression is a method to model a non-linear relationship between the dependent variable and a set of independent variables.
- ✓ Second, for a model to be considered non-linear, Y hat must be a non-linear function of the parameters Theta, not necessarily the features X.

Note: When it comes to non-linear equation, it can be the shape of exponential, logarithmic, and logistic, or many other types. As you can see in all of these equations, the change of $\hat{y}$ depends on changes in the parameters Theta, not necessarily on X only. That is, in non-linear regression, a model is non-linear by parameters.

Note: In contrast to linear regression, we cannot use the ordinary least squares method to fit the data in non-linear regression.

In general, estimation of the parameters is not easy.



Let me answer two important questions here.

Q: First, how can I know if a problem is linear or non-linear in an easy way?

To answer this question, we have to do two things.

First one:

- The first is to visually figure out if the relation is linear or non-linear.
- It's best to plot bivariate plots of output variables with each input variable.
- Also, you can calculate the correlation coefficient between independent and dependent variables, and if, for all variables, it is 0.7 or higher, there is a linear tendency and thus, it's not appropriate to fit a non-linear regression.

Second one:

- The second thing we have to do is to use non-linear regression instead of linear regression when we cannot accurately model the relationship with linear parameters.

Q: The second important question is, how should I model my data if it displays non-linear on a scatter plot?

- Well, to address this, you have to use either a polynomial regression, use a non-linear regression model, or transform your data.

Linear vs non-linear regression

- How can I know if a problem is linear or non-linear in an easy way?
  - Inspect visually
  - Based on accuracy

- How should I model my data, if it displays non-linear on a scatter plot?
  - Polynomial regression
  - Non-linear regression model
  - Transform your data