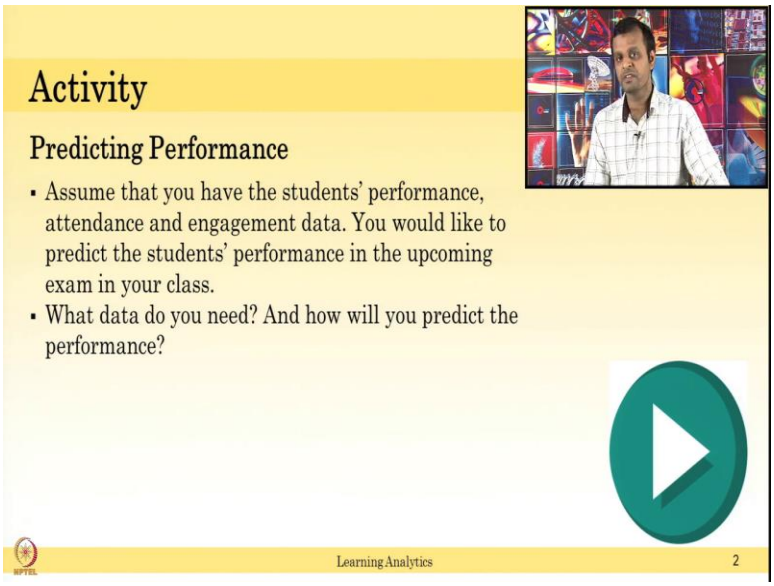
Learning Analytics Tools
Professor Ramkumar Rajendran
Department of Educational Technology
Indian Institute of Technology, Bombay
Lecture 8.2
Linear Regression

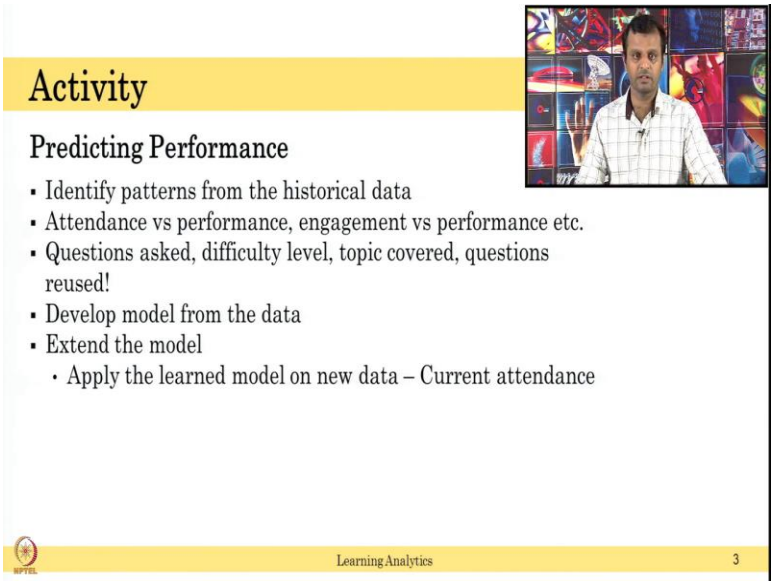
(Refer Slide Time: 00:22)



In this video we will talk about Linear Regression in Predictive Analytics. So, let us start with the activity, assume that you have student's performance, attendance and engagement data and you would like to predict the students performance in the upcoming exam in your class. You have historical data, of three years or four years, and you want to predict the students performance in the upcoming exam.

Given all this data, which data is important for you? And how will you predict the system? So, it can be regression or anything else? So, think about it, write down your answers after writing it down resume the video to continue.

(Refer Slide Time: 01:03)

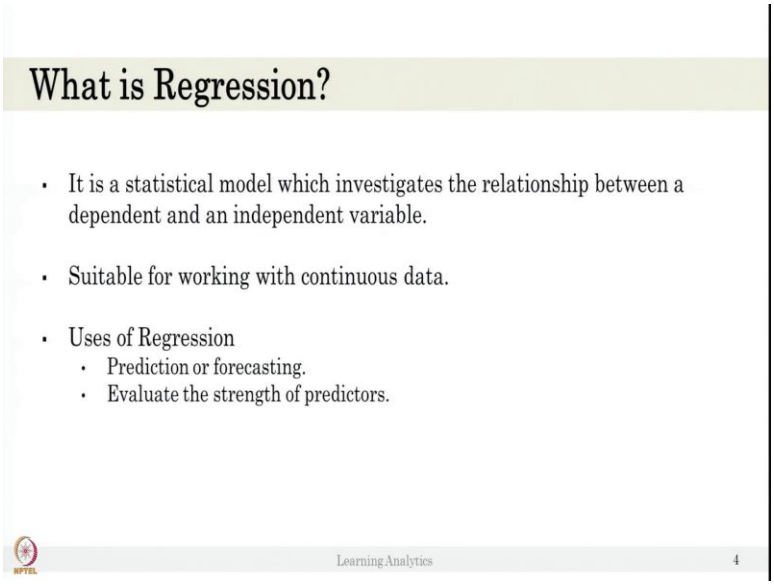


Predicting performance recovers identifying patterns from the historical data and you can use attendances versus performance or engagement versus performance, you can do those kind of data picking. So, you can also talk about, difficultly level of each questions, topic covered, time taken to cover the topics, Questions re-used, all these things can be also considered for the predicting.

So, the step is you have to develop the model from the given data.That will be the linear regression model and we are going to we will talk about it. After developing the model from the training data you have to extend the model to predict the future events that is apply the model to new data that is with the current attendance, and check what will happen?

Suppose you created a learner regression model, using learner's attendance versus performance from the historical data and that is a train model and you should test it. You going to use the model to predict the performance of the student's attendance in a current semester, so apply the learnt model to predict the future event that is a new data, that is a linear regression or any other predicting algorithm.

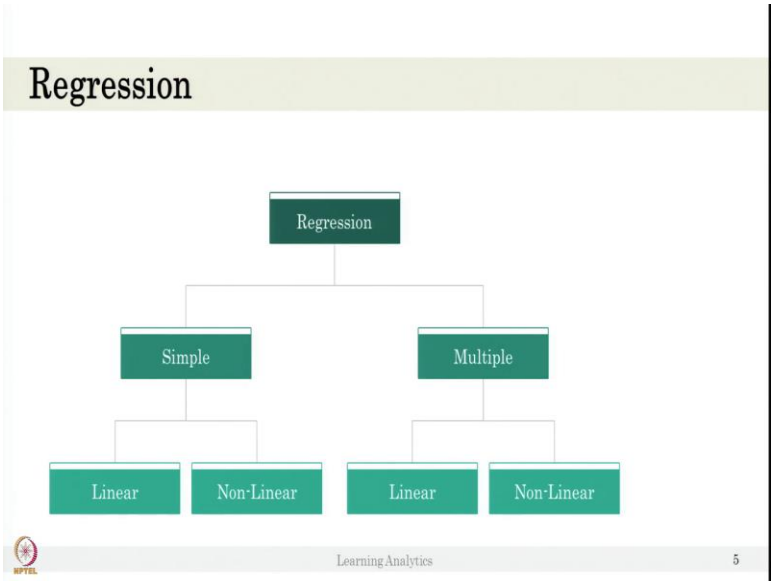
(Refer Slide Time: 02:24)



So, what is regression? We talk about regression in this video. It is a statistical model which investigates the relationship between a dependent and independent variables. We already saw a regression briefly, that it is to identify the linear relationship between dependent and independent variable, this assumption is important assumption for linear regression. And it is suitable for working with a continuous data, for example the marks, we saw that instead of classifying into bin A, bin B you want to predict in a continuous data, like 75, 76.5 something like that.

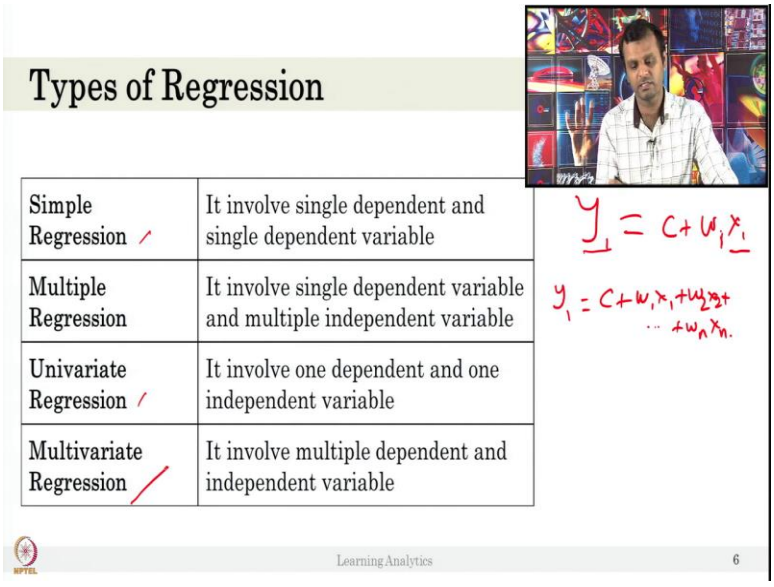
The use of regression is predicting or forecasting that is what will the student's performance. Or evaluate the strength of predictors, which particular independent variable is strongly associated with the dependent variable something like that also can be evaluated using linear regression model. It is used widely just because it is easy to understand.Thats why, it is used widely as a first step in a predictive models.

(Refer Slide Time: 03:40)



So, we saw this picture already, the regression can be classified into simple and multiple. Again simple can be classified into linear regression or non-linear regression, let us talk about simple linear regression, then we will see that one example of multiple linear regression.

(Refer Slide Time: 03:58)

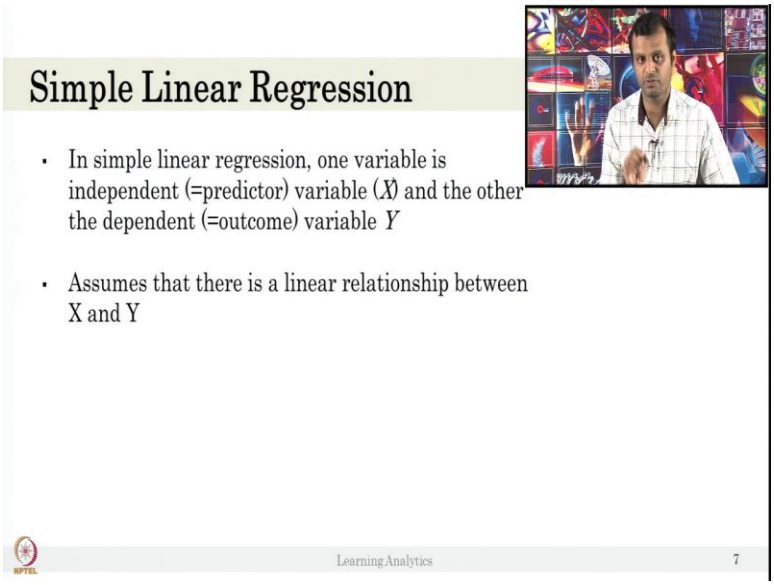


So, there are four types of regression, let us take this linear regression, first we look at simple regression, means it has one dependent value and one independent value. Let us see, what it means, it implies y is 1, there is no multiple things you are predicting, So, there is a only one independent and one dependent variable here, so that is called very simple regression, simple

regression. In multiple regression there is a one dependent value and number of independent variable, see y1 equals intercept plus w1 x1 plus w2 x2, so this can be right upto n variables.

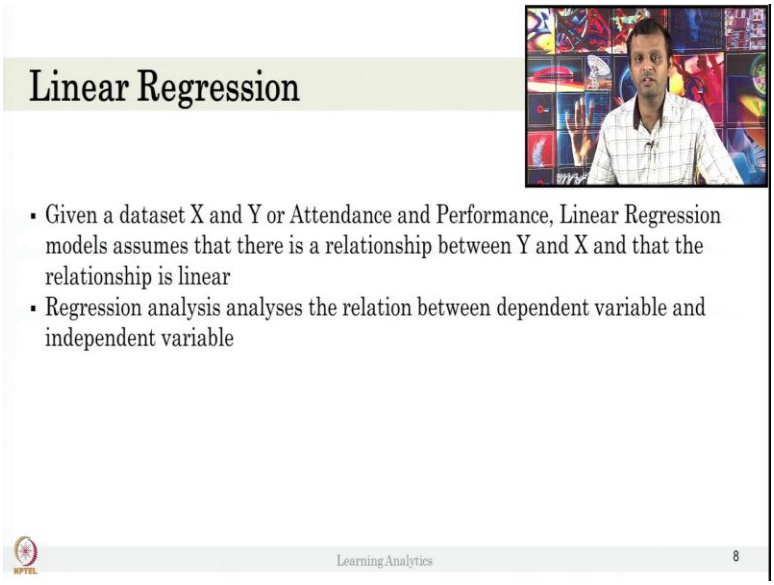
So, univariate regression and simple regulation are the same, there is no difference, its also have only one dependent and sngle independent. But, let us think about the multivariate equation, it is involves multiple dependent and multiple independent variables, it is not just y1, you might have yi, like you might predict yi y1 y2 y3 also possible.For this course, we will not talk about that, that is not needed. Let us think about simple regression.

(Refer Slide Time: 05:57)



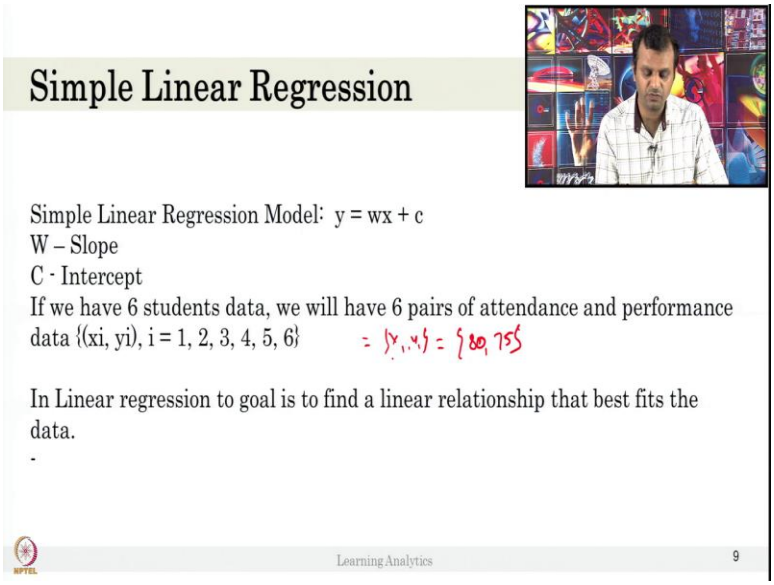
So, in a simple regression one variable is dependent that is to be predicted other variable is independent one , i.e. X and Y, the X is the independent and Y is the dependent variable. Assume that there is a linear relationship between X and Y, so this assumption is very very important, there is a linear relationship between X and Y, if there is no linear relationship simple cannot be used.

(Refer Slide Time: 06:23)



So, given a data set X and Y for example attendance and performance or engagement versus performance, linear regression model assumes that there is a relationship between X and Y. The regression model analysis the relationship between dependent variable and independent variable and try's to create a model from the historical data.

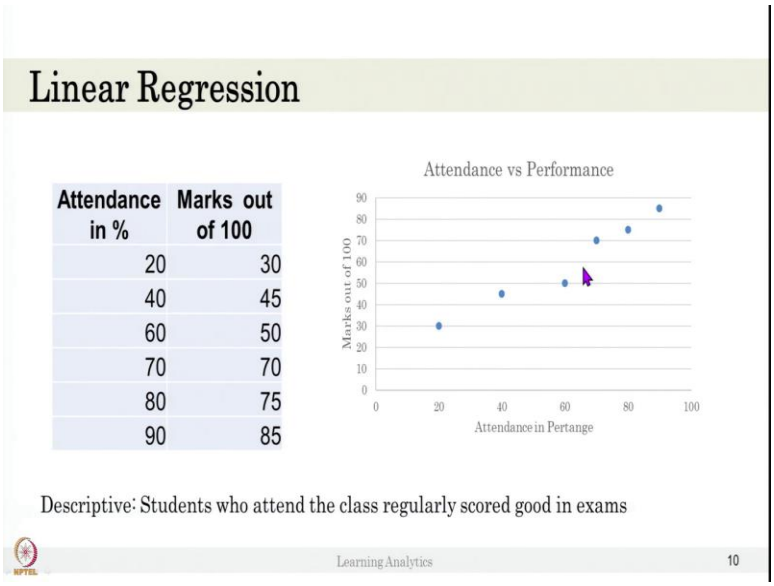
(Refer Slide Time: 06:41)



So, in the simple linear regression, y equal wx plus c, so you know the slope is the w and c is a intercept. We have 6 students data, which means you are 6 pairs of input data, 6 pairs of xi and yi, x1 y1, x2 y2, something like that, hope you understood what I am talking, so you will have a

one set of variable x1, y1, that can be say attendance of 80 percent final mark of 75 something like that. So, if we have, 6 pairs of x and y is available, that is the data. In linear regression the goal is to find a linear relationship between these two data that best fits this x and y that is it.

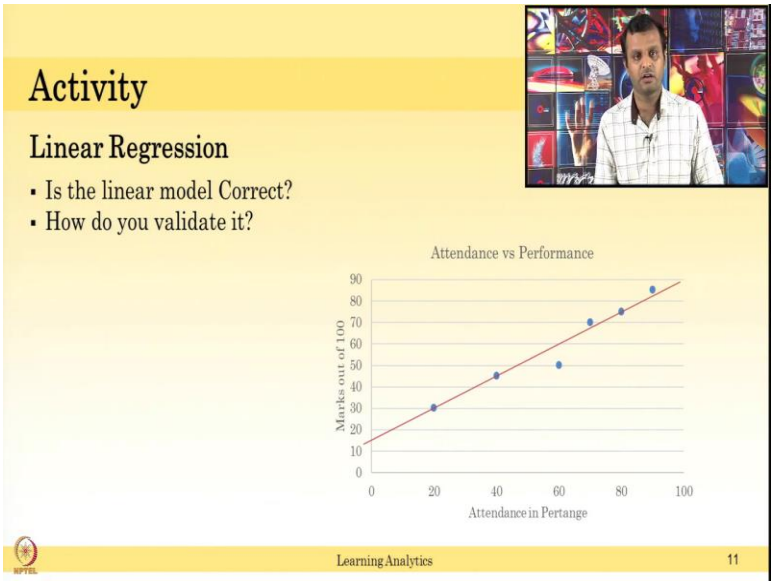
(Refer Slide Time: 07:34)



So, the very basic approaches is let us talk about the attendance, we have 6 student's attendance data and the marks out of the 100 for the same 6 student's, so we are plotting attendance in a percentage whereas marks out of a 100, 6 students attendance here, 6 students marks here. Let us see how this line fits.

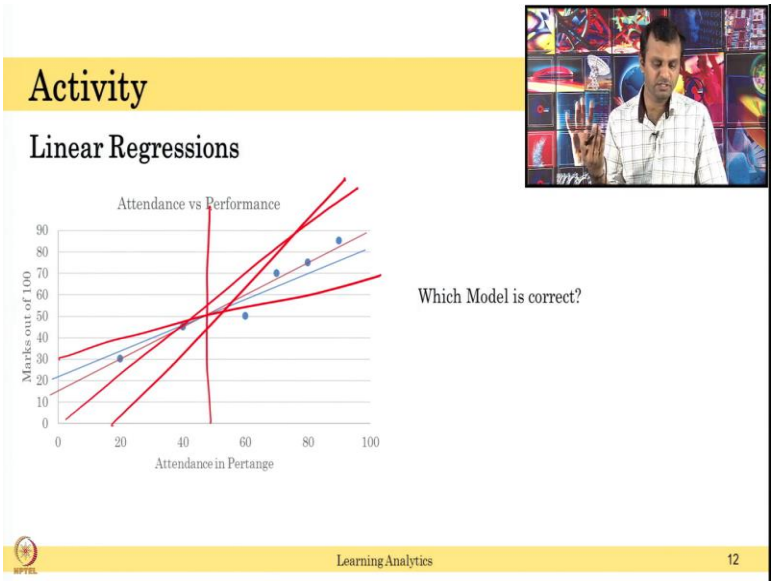
So, descriptive analytic tells that students who attend the class regularly, scored good in exams. So, by looking at this data and the plot, you might tell that there is a relationship. What is the relationship can be identified by diagnostic analytics? How to use that relationship to predict the future event is a predictive analytic. So, that is very simple example we are using it in the in this slide.

(Refer Slide Time: 08:26)



So, let us see I plotted a line here, this line is like this, so it fits almost all three data and can we say this linear model is correct? If this model is correct, how do you validate it? Can you say this model is correct? And if it is correct how do you validate it? Please pause this video and write down your answers after writing it down resume the video to continue.

(Refer Slide Time: 08:57)

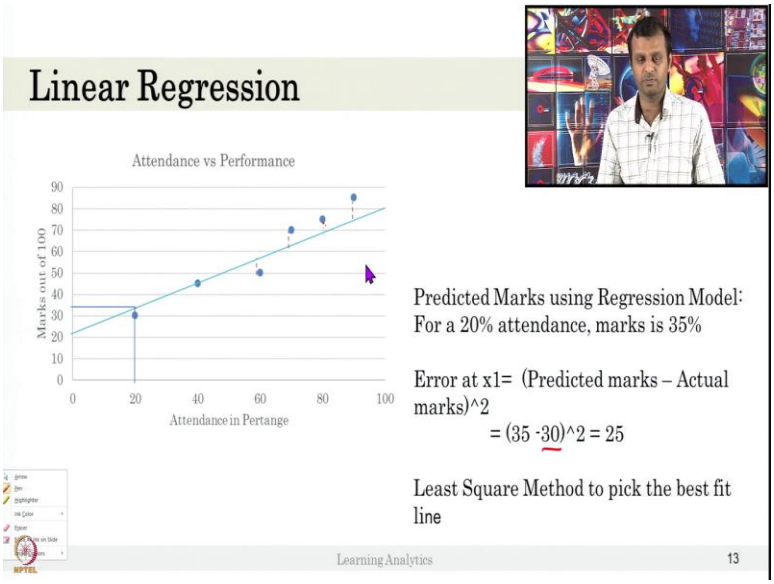


So, the linear regression, now we see two models are there, I can say one fitting all this line, one line in the blue colour it is actually bit low, but it might be better. Which model is correct? How

do you know which model is correct? There might be multiple not just two line. So, the different lines are possible, which line is good or something similarly something not at all related to it.

So, which line is good? How would you say this line is not good compared to other lines? How do you say this? This is the question. For that we have to look at the objective function, you have to pick the line which gives the very least objective function and we can see the objective function is mean square error. Let us see how to compute the mean square error.

(Refer Slide Time: 09:53)



The mean square error is identified by as a error means, the error of predicted marks minus actual marks square of that, let us see this example here. So, the actual attendance equal to 20 percentage actual mark obtained is 30, but line which is from linear equation model gives you value of 35.

So, we have to calculate the difference between this predicted versus actual, so the difference between this value 35 minus 30 is some 5, 5 squared is 25, so you want to use the least mean square method, so least mean square method is basically compute this difference between actual and predicted in each point sum it up and divided it by 6, where 6 is sample size.

So, compute that least compute that value for multiple lines, so not just one line, so you can have multiple lines, you can have one line here, one line there, so compute that least mean square for multiple lines and pick the one which is best. So, consider a system started with the say line like

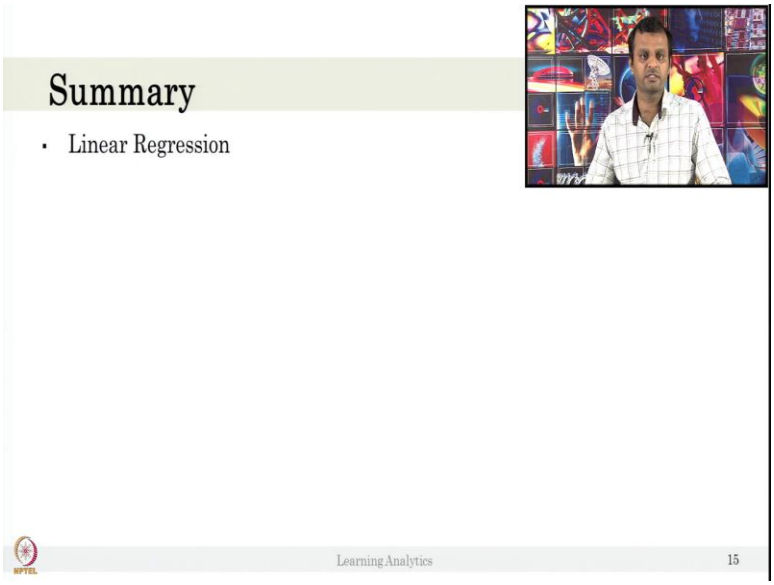
this and it found somewhere and it want to reduce or it want to go for the better model, how we go.

We can use different algorithms like gradient descent algorithm or some other algorithm, these are not discussed. If you are interested go ahead and watch the videos by Andrew NG on YouTube on introduction to machine learning, they explains it very clearly how the weight is learn to try in this linear regression model.

(Refer Slide Time: 12:09)

That particular model gives this is the better score. So, say 0.799x plus 12.402, 12.402 is actually intercept if you extend the line something like this it comes like that 12.402 that is where it cuts the 0.

(Refer Slide Time: 12:38)



So, in this video we discussed what is linear regression and how to pick the best fit model. Thank you.