Introduction to Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module - 03

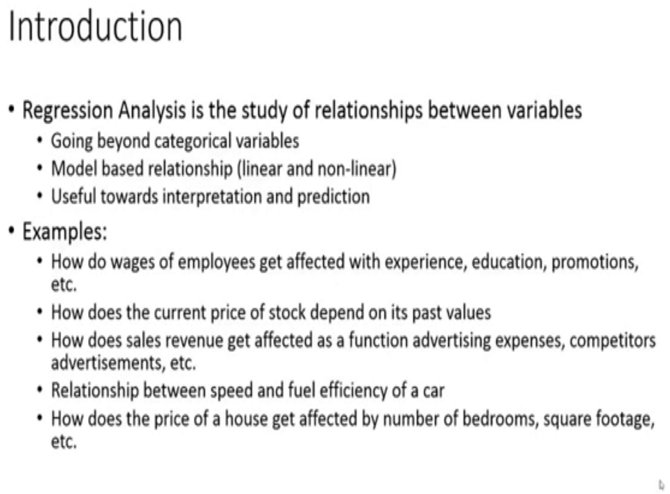
Lecture - 15

Short Introduction to Regression

Hello and welcome to our class that today on Regression. Today’s class is going to be a very short introductory level treatment of the topic regression analysis and the idea here was to a kind of position just towards the end of your modules and descriptive in inferential statistics, so that you are at point by you know enough to appreciate, how regression uses these concept.

But, we will also be revisiting regression down the road, where we will actually talk about the mechanics about, how you implement it and that is something we will do, right after you get an introduction to machine learning and more specifically to supervise learning.

(Refer Slide Time: 00:52)



Introduction

- Regression Analysis is the study of relationships between variables
  - Going beyond categorical variables
  - Model based relationship (linear and non-linear)
  - Useful towards interpretation and prediction
- Examples:
  - How do wages of employees get affected with experience, education, promotions, etc.
  - How does the current price of stock depend on its past values
  - How does sales revenue get affected as a function advertising expenses, competitors advertisements, etc.
  - Relationship between speed and fuel efficiency of a car
  - How does the price of a house get affected by number of bedrooms, square footage, etc.

So, jumping into, what a regression is. Essentially a regression analysis and I say regression analysis here, because the word regression itself in different fields it is used differently and in fact, even in statistics there is a concept called regression towards a mean, which has a lot to do with regression analysis, but it is also quite a different from, how we understand, how we use regression analysis today.

So, again the idea is to you know it use the word regression analysis. So, regression analysis essentially is the study of relationships between variables, more specifically it looks at understanding the relationship between 1 or more input variables and their relationship to an output variable. And the exact nomenclature is sometimes different, we use the words I am using the words input and output, you might also come across the term response to describe the output or dependent variable to describe the output and for the input variable you might come across the terms explanatory variable or independent variable for the input.

So, the idea is to study this relationship between some set of inputs or one input variable to this output variable that you have in mind. And, so right from the minimum you are looking at two variables and the whole idea is, have you come across this kind of studying relationships between variables, so what is regression doing newly. And the idea is yes, you have, you studied, when we started looking for instance two sample tests, two sample t tests for instance.

Right there, you for the first time you encountered scenarios, where there were two variables involved to jog your memory. An example of the two sample t test that we looked at was, for instance blood pressure and the idea of the average blood pressure was lower when given calcium supplements versus when given placebos. Placebos are just sugar pills that kind of it look like, the original medicine.

So, the idea was there were two variables involved; there was your notion over output variable or response variable, which was your blood pressure. It is a quantitative variable and you had another variable and this other variable was, you can call it calcium. And this other variable calcium, it is a variable, because it can take on two states. It can take on the state yes, meaning you have given the actual calcium supplement or you can take on the state no, which I am using as proxy to mean that you were given instead the placebo.

So, you now have two variables and you can think of your, this variable calcium is often called the treatment. But, you can think of it as the input variable, which takes on two possible states and you are looking at essentially that to see, if there is any difference in blood pleasure for people, who given calcium versus people, who have given placebo. Right there, you are looking at a relationship between two variables, you are looking at the relationship between the variable calcium and the variable blood pressure.

Now, another example that we saw on the two sample case was boys and girls in 10th standard in public schools to see, if the average height of boys was equal to the average height of girls. Again, here your output variable is the height and your input variable is gender, whether you are a boy or a girl. You are looking to see, if there is a relationship between gender and height, yes the answer is; obviously, there is, but may be in 10th standard is one higher than the other in public schools.

So, you went even behind that, for instance with something like an ANOVA, Analysis of Variances that we discussed in the last class. You go beyond just having two categories of your input variable, you can have multiple categories. We saw the example, where we looked at states as the input variable, we looked at height of 10th standards, to make it simple I just say boys. But, you looked at the heights of Tamil Nadu, Karnataka, Maharashtra, there you looking at the relationship between state as your input variable and height as the output variable. Is there any relationship? Is there any difference between boys of different states and their heights?

And finally, we also looked at something like test of independents, where here you would looking at two categorical variables and you were looking to see looking to see there was a relationship between them. But, so the reason for going into all of these and kind of revising some of the topics is to say that, you have studied the relationship between variables. But, for the first time you are going to go beyond dealing only with categorical variables. Regression analysis is especially powerful, because it creates a relationship between two variables, but this relationship can, both these variables can be continuous and quantitative and therefore, the scope of application for regression really becomes much greater.

So, to give you an extension for instance of some of the previous examples we have looked at from the top of my head, one thing I can think of is we will looking at the relationship between blood pressure and you would given calcium plus or placebo, so there is just two possible states. Now, what would happen if I gave, if I had 20 different types of terms, which started on one end from the placebo, which had 0 calcium and went all the way to the other end, which has 100 percent calcium.

And in between, the 20 different types of pulse, I am not saying there are 20 pulse there are 20 different types of pulse. The different types of pulse had varying degrees of, varying percentages of calcium. So, now, you have at least 20 states and those number of states can, if especially if we are randomly sampling, that can be infinite number of states, but you just have a range. So, it is a continuous quantitative variable between 0 to 100 percent of calcium and your output variables still remains your blood pleasure.

So, you can actually create a relationship between these two variables, it can be a mathematical relationship, it can be a graphical relationship, but it goes beyond just 2 states or 3 states or n states. It has the potential to go to infinite number of states of the input variable or variables and, so it can describe you know, richer things. The idea again behind regression analysis is to not just unlike most of what we studied with inferential statistics. It is, the goal is not just to establish that there is a relationship, but it is to quantify that relationship as well.

So, a lot of emphasis in a regression is given to the model itself, that mathematically equation that describes this relationship between the input and the output variable and that can be either linear on non-linear relationship. That, the relationship between amount the calcium that is given and the drop in blood pressure if there is 1 can be a linear or non-linear relationship and you can extend this to all the other examples. Now, as with most supervised learning techniques and that is something that we briefly discussed in the course overview.

But, it is also something that you will hear a lot more often the upcoming lecture. The goal of a regression can be to fold, it can be towards prediction, the idea that you now create this linear relationship. And, so I can now use it to predict, what the drop on blood pressure would be for a given amount of calcium supplement and that particular amount of calcium supplement might not have even been something that was given in the original data set from which I built the line.

But, the fact that I used this data to create this relationship in the form of a mathematical equation, which represents a line, which represents the relationship between calcium supplements, the amount of calcium given to the drop and blood pressure allows me. This, the fact that I have created this allows me to use this tomorrow to predict, how much the drop and blood pressure would be. If someone came and told me, I am thinking of giving this amount of calcium supplement, it allows me to make predictions.

The other thing that it also does is, given that you are now able to create a mathematical formulation, it gives you an understanding of the word. It gives you an understanding of, how far how much extra calcium that I gave, how much of a blood pressure drop will I see. So, it is an understanding of how these two variables are actually related and that is, what we essentially call as one prediction, which is one goal, which is the kind of predict and the other is interpretations, which is the kind of interpret the world that we level to through the lens of this linear relationship.

So, let us just talk about just to give you some feel for where a regression analysis is useful of what kinds of contacts. Here are some examples that we have listed. For instance, so how do wages, how do salaries of employees get affected with variables, such as experience, education, promotions, etcetera. So, the idea here is that salary is the dependent variable, it is the output variable, it is the response variable and variable such as experience, which might be measured in number of years, education again measured in terms of number of years after high schools or something and promotions, again a numerical value, how do these variables affect the output variable.

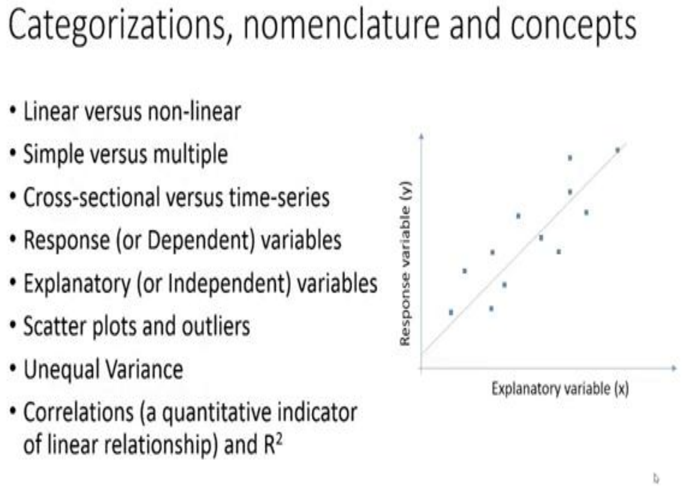
Another example is how does the current price of a stock or a share depend on it is past values and perhaps, also on values of the market indices or other stocks. A circle difference here that you might notice and we will circle back on this is, in the case of the stock for instance you will notice that the output variable is the price of the stock, but the input variable is also the price of the same stock on previous days. So, how do it is current price depend on it is past values? So, here the past values become the input variable and the current value becomes the output variable.

So, we call that time series and we briefly revisit that concept as well, but regression are also used in that context. So, the next example is, how does sales revenue get affected as functions of advertising expenses and comparative advertisements. So, the more money through into advertising you might have belief that the sales goes up, but what is their exact relationship. Can I look at past data on different advertisements that have different campaigns, advertising campaigns that are made for different products and what my competitors did and how that affected my sales revenue and create this relationship?

And also something else from the mechanical engineering, house would be relationship between speed of a car fuel efficiency within a certain range, when you are driving on a certain gear, what is a relationship between the speed, which you drive in a fuel efficiency and you could take a sample of 10 cars, make them drive at different speed or you know, take the same car, make it drive it at a certain speed and get data and try to fit a line in this data to understand the relationship.

And finally, another example could be how does the price of a house get affected by the number of bedrooms, the square footage of the house, you know it distance from the center of the city and so on and so forth. Again some of these input variables need not neatly fit into being continuous quantitative variables, but the idea is that a regression is not constrained by that. A typical application of regression would use continuous variables, but the whole selling point associated with a regression is that, it is a technique that also capable of dealing with continuous quantitative variables, whereas a typical ANOVA or a t test or any of the tests that you seen so far are built around the idea of using a categorical input variable and that is, what they meant to do.

(Refer Slide Time: 14:19)



Categorizations, nomenclature and concepts

- Linear versus non-linear
- Simple versus multiple
- Cross-sectional versus time-series
- Response (or Dependent) variables
- Explanatory (or Independent) variables
- Scatter plots and outliers
- Unequal Variance
- Correlations (a quantitative indicator of linear relationship) and R^2

So, let us talk about some more concepts associated with regression. Just for reference, I have shown you a typical graph, graphical representation of regression, where the x axis is your explanatory variable, your y axis is your response or output variable. And each square is a data point, by square I mean each of these small points, it is a data point and the idea is loosely the idea is to kind of fit a line through this data point and that is what we would call linear regression. If you are trying to fit a line through the data point, we would call it linear regression.

The word actually linear regression could also be used, when you are doing a linear combination of variables. But, some of these variables themselves represent a non-linear transmission. I think, the simple way to put this is whether you are looking to have a linear relationship between the input and output variables or a non-linear relationship. You could still use the core concept of regression, but you should be a little careful in terms of what gets called linear and what gets called non-linear.

Because, often the camp of linear regression could include variables, the input variables themselves which are non-linear transformations and, so it would still be called a linear regression, but it will not exactly look like the graph that you are looking at. You also have the categorization of simple versus multiple regression. The idea behind simple regression just means that there is one input variable, whereas in multiple regression you have more than one input variables.

So; obviously, in the graph that you see in the slide, you looking at a simple regression because there is one explanatory variable, which is the x axis, if you have more than one that becomes hard to represent on a two dimensional slide, you would have to use a 3D model for two variables and then, after that it becomes much harder. But, the idea is that if you have one input variable it is simple regression, if you have multiple input variables more than one input variable it is multiple regression.

We then come to the next concept that we briefly discuss which is cross sectional versus time series. The idea behind cross sectional is that, it is not a function of time. Your data is collected across the board and that temporal, the time component of basically means that it is not either function of a time or it is not a function of it is previous values.

So, you can put all the variables in the basket and think that they were all created at the same time or at least that you do not care, whether they were created a different points of time. You are still going to treat them it is just different data points and look at your relationship and that would be cross section, whereas time series is the idea that previous values in time affects subsequent values.

So, your modeling techniques themselves become a little different and classic example of the time series was the stock example that we spoke in the previous slide, where previous price of the share influence tomorrows price, where as the equivalent of the cross section analysis of that would be to completely ignore the stock price of this to completely ignore the previous days or previous time periods stock price and just say can I predict what this stock is this stock is value is going to based on other stocks or other indices.

So, can I use the, in nifty index can I use another stock to kind of say based on the stock this is what the stock should be this is what the output the response variable stock value should be. Next we just come to the idea that this is notion of response variable or it is called dependent variable and the input variable is called explanatory variable or independent this is terminology that you will keep hearing and, so it is kind of important.

So, the next important point is that we learnt of about something called scatter plots during the descriptive statistics face and regression analysis in many ways essentially captures mathematically, what does scatter plot tries do graphically. So, scatter plot tries to graphically show you that relationship between two quantitative variables, where is regression analysis tries to go beyond that and tried to actually fit a line to this or fit a model to be more general fit a functional from to this data and there for that is the improvement on scatter plot that regression analysis does.

But, often even before getting into regression analysis scatter plot could be very useful graphical window into what you should expect to even see if you should try to fit a linear equation or non-linear equation, because just visually you might be able to say you know this does not make sense to fit a linear equation, because that is not the co relationship. Another big advantages scatter plot is that if there these outliers in by outliers we mean essentially there are data points, which look like they are essentially just errors or something.

So, you let us say you had this data point that was in some where completely unrelated that could come from an error in that it completely rue in a regression analysis along, but if you see it in a scatter plot you might choose this say I want to ignore this data point that is idea a behind scatter plots and outliers. The next concept is an unequal variance we going to talk about this little bit, but the idea could be that they could be a very strong linear relationship or non-linear relationship.

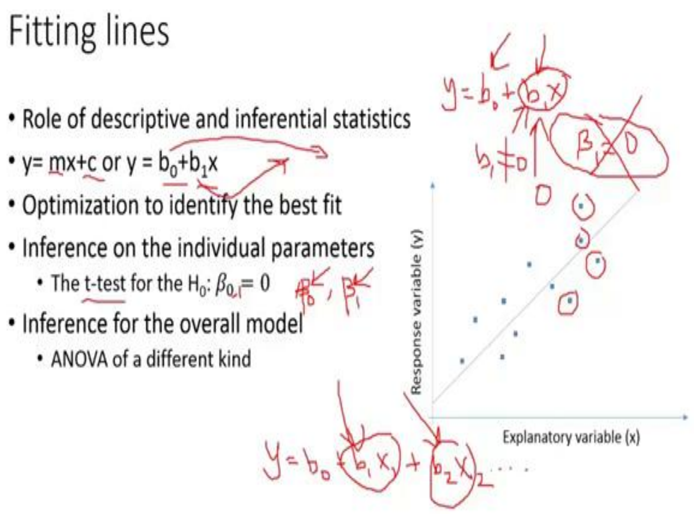
But, the variability at different points of x or your input could be different you could have something that looked like this in terms of the data. So, this is the line that your fitting, so there is le this central line is still the line that you are fitting and it is a linear line. But, the data points in the lower end, so in this side of x the data point a closer to the line, where as once you go to this side of x once you go to the side of data points of far more spread out.

And that influences, how the line gets created in there is in techniques that available to kind of counter this problem it is a problems, because again visually you my just see this speaker phone kind of shape and say still I am just going of fit a line in the center, but this higher variability on one side and lower variability kind of makes does not always result in the correct line being fit and you need some kind of transmission to understand the relationship better.

So, the last concept associated with the slide in what I want to talk you about is with co relationship. So, co relationship you can think of is a more advanced form of summary of descriptive statistics we dint speak about in detail or in the descriptive statistics part. But, the idea is that when you have two variables just like a scatter plot can be used to graphically describe these two variables, correlations is a single number that describes that a linear relationship.

And its essentially a quantitative indicator of that linear relationship, what you will discover is that in a regression analysis you will also be coming up with certain numbers that quantify this linear relationship and if you have not already heard if this you will come across this something called r square and that has a direct quantitative mapping correlations. So, it is not a new concept essentially r square is nothing but, correlation square, but you will come across it under the term r square, where as you might of heard is of the word correlations more locally used or used more in the concept context of descriptive statistics.

(Refer Slide Time: 22:54)



Fitting lines

- Role of descriptive and inferential statistics
- y= mx+c or y = b_0+b_1x
- Optimization to identify the best fit
- Inference on the individual parameters
  - The t-test for the H_0: $\beta_{0,1}$= 0
- Inference for the overall model
  - ANOVA of a different kind

$y = b_0 + b_1x$
$b_1 \neq 0$
$\beta_1 \neq 0$
β_0, β_1
Response variable (y)
Explanatory variable (x)
$y = b_0 + b_1x_1 + b_2x_2 \ldots$

Now, let us just briefly talk about the exact roll of descriptive and inferential statistics. If we start first with descriptive statistics the whole idea there was either graphically or to quantitatively summarize the data that you see. And one way of summarizing that one essentially summarizing that you do of the data in regression is to capture the linear relationship, which is there in the form of the actual data into line.

And, so you essentially summarize this relationship in the form of an equation and many of you might know that if there is if the simple line in two dimensions you have this formula y is equal to m x plus c you might of come across that in high school in some form or the other, where c essentially represent the intercept of the line on the y axis and m represents the slope.

So, you can think of ms that angel m represents that slope and you can in regression often we use the terms beta and here I am just using b not in b 1, because I am representing the sample based of the sample that we have we have some estimate of b not and b one. But, the idea is that this is some form of summarizing the data, because you taking the data and summarizing it to the single equation.

So, what is the concept between behind, how you do it and the idea is there is some form of optimization it is the form of optimization, because you have a set of y s and x s. So, you are given a series y s and x s and that is your data and for a each y and x you essentially plot it on this graph you than go had an you take line and this is that line that we are going to play with and you choose some criteria to fit the line through the data points.

So, you might say that my goal is to fit a line through the data points and you know I do not want to draw a line out here that has nothing to do this data. So, what kind of a line as something to do the data I am its say well I want the line to split the number of data points equally above and the below it. So, this will do be overlap the, but we get the point, which is the you might want you might want that line to split the data points another idea could be its say I want the line to go through the two extreme points.

So, there are two extreme points I want two extreme points to be connected by a line and you can have various other criteria and another very common one could be to say I more advance one could be to say I want to minimize the distance between each point to the line and that incidentally happens to be what we call is the ordinary least squares. But, we will get into that later, but that could be one criteria, another criteria could very well either I want to minimize the perpendicular distance to each points to the line.

So, you could have various criteria and I know you can choose some criteria and you can say I want to achieve that criteria and that is how a want to fit a line. And, so at the root of doing that is this series of techniques and optimization where you say I want to do some form of optimizations. So, you might ask the question, what are you relay optimizing and the answer is your really optimizing this objective of minimizing some metric and let us stick to the metric of minimizing this distance this sum of this distance.

So, how are you optimizing that, what it can you change what you can change is this line and what I am going do is I am going to change the two parameters associated with this line and the two parameters are m and c. So, c is, where this line intersects on the y axis, so I am going to try and move this line above and below keeping that is slope this same. And try to see, where should this line b such that I minimize the sum of these distances those red lines such you see my goal is to minimize the sum of these distances.

And that is an optimization to see that is my objective function my objectives is the minimize the some of the distances. And I am going do that by changing two variables I am going to change the variable c by moving this line above and below and I am going to change the variable m by rotating this line, so by rotating this line, so this way and this way. So, simultaneously I am trying to change two variables and thereby find that line by changing my m and c I find that line, which minimizes this deviation the deviation if each data point to that line.

And there are different ways of doing there I am just giving you one objective that am I choose to optimize. But, it the route to way you need to realize that ultimately the process of regression fitting a regression line is process of optimizing and you might different criteria ordinary least squares, which is one type of regression uses the criteria of taking this distance of each point to the line and squaring it and trying to minimize the some of those squares.

Because, each point will have some distance to the line and then, I can take that distance square it and then, I can take go to the next point in do this same and then, you can sum all those squares and least squares tries to minimize that objective function. But, that need not be the only objective function there can be many other objective functions, but at the idea behind putting line through many data points is to say I am interested in maximizing and minimizing some goal associated with the process of fitting the line.

And I am going to do that by changing the two variables set I can, which is MNC or in other words you are b not in b 1. So, that is essentially the summary statistics the process of describing the data through line and that is how that line gets created. Now, where is the concept of inference the concept of inference again comes from the score idea there ultimately these data points are samples they do not represent the population.

So, any b naught and b 1 that you calculate are coming from this samples, so there essentially the sample statistics. So, but just like x bar in some sense is the sample mean, which is a point estimate of new, which is the population mean these this b naught and b 1 essentially represent beta naught and beta 1. So, you can you essentially have the population parameters beta naught and beta 1 yes these are the population parameters let us just beta naught in beta 1.

Now, the idea is that, which statistically inference what you doing is your trying to see if either of these terms beta naught or beta 1 are actually equal to 0. Because, if they are then, they do not have any business in being a part of this equation this equation that you have out here now the this equation that you have out here that is an equation that you started with before you even fit anything. Now, you a just in the process of finding out to b naught and b 1 through some optimization procedure, now imagine a situation, where x has nothing to do with y x has nothing to do with y.

Now, you calculate the sample of x s in the sample of in y s if you had infinite number of sample you might arrive at the conclusion that b 1 is thoroughly equal to 0 therefore, this terms itself becomes 0 on should not be there in this equation. But, you still only have a sample and the sample is 5 data points 10 data points 20 data points. So, it is perfectly possible that even though true beta 1 is equal to 0 that is the null hypothesis even if that is true you it is possible that you just take a sample of y coma x s and you get some beta b 1, which is not equal to 0.

So, b 1 winds up not being equal to 0 even though beta 1 is equal to 0. So, you might be erroneously thinking that x has this relationship to y, where is in reality beta 1 is 0. So, the whole idea is to do a statistical test to see to test the hypothesis the beta 1 is 0. So, the null hypothesis would be that beta not is 0 and you can think of it as also beta 1 is 0 and it turns out that under certain assumptions of normality and even when those assumptions are violated to some extents central limit theorem it terms out that the distributions of the betas of these coefficients winds up being t distribution, so like t distribution.

And, so you can use essentially t test to test the hypothesis that each of these coefficients are actually could the 0 not and the idea is that if you wind up you need to reject the null hypothesis and reject the hypothesis that beta 1 is equal to 0 and only then can you actually have the term in the model and this become really useful when you have you know multiple input variables. So, you might have something that is is beta naught class beta 1 x 1 plus beta 2 x 2 and so on, and it can go on you can have many variables.

So, which of these terms actually get to stay in the modeling, which do not is not just the function of this magnitude of beta 1 and beta 2, because that is the magnitude becomes really function of it becomes also a function as magnitude x 1 x 2 in the units and so on. But, what really determine whether these things stay in the equation or not is the inferential statistic test that you do for each of their population parameters, which is for beta 1 for beta 2 and so on. And in, so for is you can reject the null hypothesis that these are equal to 0, then you can leave them in the model.

But, if you can if you if you cannot the reject the null hypothesis than these do not these terms do not have any business being in the model. So, that is the idea behind doing statistical inference on the individual parameters you will also notice there in a typical regression analysis or regression analysis output independent of the software that you use there is you will find an inference for the overall model. So, in a side from the inference such you see for beta 1 beta 2 separately.

For given model you will have an inference and that inference again uses the concept of an ANOVA and the way it does that is by looking at if you remember the ANOVA in the past you have the concept of mean squares between and mean squares error, where the mean square between was trying to quantify the effect of each treatment or each state of the input variable and mean square error was trying to quantify the inherent noise that was there even with in each state.

Here; however, you do not have finite number of means square between, because they could be infinite states of the input variable x. So, you have a single term that tries to capture how much of the variability is being captured by this linear or non-linear model that you built, how much of variation in x is being captured by this equation, how much of the variation in, how much of the variation in y is being captured by this model versus how much of the variation in the data points is not captured by the model.

So, 1 becomes the mean square of the model essentially and the others becomes the mean square error, which is just the general, which is some notion of general noise that is there in the system and essentially the ANOVA out here. Again will be using that test the same ways use the f test when we where describing the ANOVA in the previous class for categorical variables. And there by the rejecting that null hypothesis would mean that this model explains variability in y, where as if you fail to reject it the idea is that you cannot say that this model essentially you cannot say that this model is any different than a few word to just randomly choose y s completely ANOVA, where what the model was.

So, no variability in y is being explain by different points in x. So, an essentially when you think of that nothing but, the concept of straight line its essentially concept of something were not describing the variability in y at different points of x. So, various points of x you looking at the same position same position y I hope that made that gave you an introduction to the concept of regression and like I mentioned after we introduce the idea machine learning and supervised learning we will be revisiting regression to look at more end up treatment of how you actually do it and give you an understanding for how the co derivation are made.

Thank you.