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

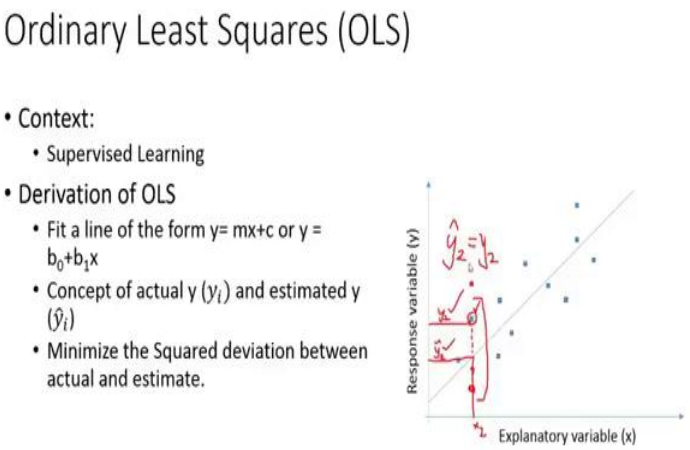
Module - 04

Lecture - 19

Ordinary Least Squares Regression

Hello and welcome to our second lecture on a Regression. In the previous lecture we provided you with motivation for why linear regression could be a very useful data analytic tool. And today we are going to take the ordinary least square regression, which is one type of regression and actually step through the process and in some sense, derive the formulas or the math that enables you to convert the data to performing a regression analysis and the context in which, we are going to do that is, we are going to do a simple regression, which just means that this single input variable only involved and we are going to step through a mechanics doing that.

(Refer Slide Time: 01:04)



So, what is the broader context of this exercise, so we introduced, we give a motivation for linear regression in the previous class. And since then, you should have had a few classes by Professor Ravindran talking about machine learning in general and also a

module on supervised learning and we purposefully choose the straddle regression before and after supervised learning. Partly because, it is important to realize that you know regression linear regression the whole process or any other form of regression is a supervised learning tool. You know supervised learning being a more an umbrella term, definitely encompasses a regression and regression based approaches.

And this is despite the fact that for instance regression is something that has existed many, many, many years before you know even the terms machine learning or supervised learning or artificial intelligence was even taught off. So, you know the context that you often learn regression could be quite different, where you learn it from statistics course, whereas in a machine learning course the emphasis sometimes might be another tools or not I mean depends on where, what the focus is.

But, the important thing is to acknowledge us, while regression sometimes stand alone in your statistics text books, not sharing pages with the other machine learning techniques. The regression, linear regression is just as much as supervised learning tool or anything else or any other supervised learning tool. Another source of confusion that I just wanted to clarify before proceeding is, supervised learning techniques tend to get broadly classified as regressions type problem versus classification problems and there, what people I meaning is quite different from what we are learning as regression at this stage.

Out there, what people I am talking about and it is just a definition is, when they say it is a regression problem in supervised learning, all they are saying is that the output variable is a continuous quantitative variable, whereas when they say they dealing with a classification problem, they are saying that your output variable is a discrete or categorical variable. And you know within those two broad classes you have many techniques and some techniques comfortably handle both types of data.

But, that is sometime gets confusing people saying, it is a regression problem does not mean I am doing regression no or there people, what people are meaning is that the output variable is continuous quantitative. Having said that let us proceed with, what we wanted to do today, which is deriving the ordinary least squares regression.

So, the goal out here is to fit a line essentially of the form y is equal to m x plus c. So, that is the form you might be heard of more frequently, what we are going to use in this

class and in most classes is y is equal to beta naught plus beta 1 x and that is you can readily see those are both the same, I just replace m and the c with two other terms.

So, the coefficient is beta 1, the intercept is beta naught, so a line once someone comes and tell you the values of beta naught and beta 1 or m or m and c whatever you prefer, but someone comes and gives you those two values, then you can define a line. If someone says draw a line, you can draw different lines you can draw line like this, you can draw line like this, you can draw line like this these are all lines, now these are all as straight as you see them lines.

But, once someone comes and gives you the exact beta naught, there is the intercept and the slope that is a very specific line, only one line will have that exact beta naught and beta 1. So, that those two terms are what define the line and to give you some intuition beta naught is nothing but, where the line intersects the y axis. So, if you wanted different lines with the same beta naught, but different beta 1s, then you can think of many lines that go like this, go like this, that go like this, these are all lines and just keep in mind I am trying to draw as straight as possible.

So, these are all lines that essentially have the same intercept beta naught and different slopes beta 1s. Similarly, you could have different lines that have the same slope, but different intercepts, so that would look a little bit like this. So, these lines all at least in, what at least in terms, in theory have the same slope, but different intercepts. Now, if but once you defined a slope and intercept there is only one line that has that, so that is the idea and what we are trying to do out here is saying, what should that slope and intercept be; such that you feel like that is going through lot of your data points.

Now, I have said that in a fairly big way, but I am going to define that more formally. To define that more formally you want to have a concept of the actual data point, this is the actual data points, all those squares are the actual data point and what the estimated value of those data points are. So, this data points has value a particular value, so we will call that y 1 and this data point has another y 2 and so on.

And, so we call those the actual values as y i and for each of these, now if I chose to fit a particular line that I feel is like going through this data, I am going to have some predicted values. So, what I will do is I will fit this line that is I will put this line here and I will say, my prediction of this y 2 is y 1 is nothing but, for that value x 1, where is my

line. So, I push this value x 1 up to the line, what value am I getting a y and this is my predicted y 1 and it is represented usually with the small hat that you put on top.

And this same process for y 2 I will I will try to write a this actually this line is not perfectly correct, so let me just erase that. Essentially, what I will do is this is my y 2 I will draw a line there this is y 2, but my prediction for y 2 is here. So, I am going to put a dash line here and this is x 2 and my prediction is y hat of two out here and you guys can see what I have done here. So, I have basically said look this is value of y 2 and it corresponds to some x 2 and I am going to take x 2 and see, where my line goes through in terms y values.

So, this is in some sense my actual value and this is in some sense what I would wind up predicting for y 2, because I have tried to kind a fit some line through a data and you might ask a question. So, if this is x 2, then why do not we then just predict y 2 in the sense y is not y hat of two equal to y 2 and the answer is fairly simple you do not want to predict the exact data point because you are getting a sample we discussed how this is not a population.

So, the population here for y 2 would be a for the same x 2 I had entire universe of possible y is and we know that for this same x value if you what take a other sample that might not fall exactly on this data point. The next one could wind up falling somewhere here, the next one could wind up falling somewhere here, the next one could wind up. So, we do not have the entire population of possible y is at the x values at the input variables x 2 and, so what we wind up doing instead is not predicting exactly on top of that value that you got.

But, instead trying to fit this line acknowledging that there is going to be some noise above and below and you might do better of predicting at this point, where x intersects with the line and that is your predicted y 2. This is, so that you do not wind up getting fooled in some sense by just some amount of noise or uncertainty there is above and beyond the exact that the trend that your sighting the line in some sense indicative of the trend that is, which is in general when x seems to go up y seems to go up and that is, what the line is showing at least this particular line with the positive slope.

And you want to capture that, so that tomorrow someone say, what will, what do you expect when x is equal to this value you go to the line rather than you going actual in

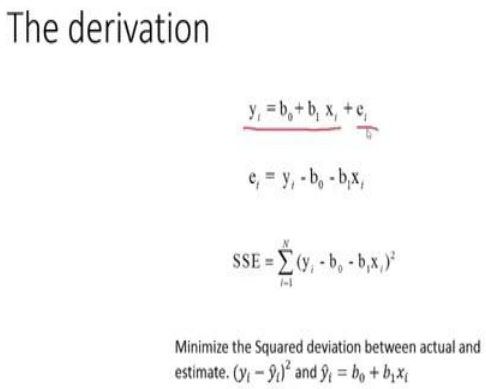
individual data point. So, let us see an idea you know the concept of actual y i and y i hat and the goal in terms of what we are trying to do is that we trying to minimize the squared deviation between the actual and the estimate. So, we are actually saying this is measure which is y i minus y i hat and sometimes y i minus y i hat is going to be positive this case this case is positive, because actual its greater than the estimate.

In some case it is going to be negative, but you take all these positive thing and negative numbers for each number and square it, then they all become positive. And then, you sum it this is the sum of the square deviation between the actual and the estimate and it is a measure of how close the line is to its data points and what we are going to try and do with an ordinary least squares regression is figure out that line. And, how do you define a line you define it with beta naught and beta 1 once you fix with the beta naught and beta 1 the line gets fixed.

So, we are going to try and figure out the goal of this exercise to figure out that beta naught and that beta 1, which defines the line, which results in minimizing the squared deviation between actual and estimate. Because, may be this line with another beta naught and beta 1 is not is very far away from your points. So, the square deviation between actual and estimate is going to be huge or take another example this line, which is like this is also not going to work very well.

Because, look at the kind of deviation that you have between actual values and estimated values. So, this line with it is beta naught and beta 1 this line in red with its beta naught and beta 1 might again not do to well. So, what whatever we trying to do is we are trying to figure out that line when I am saying we are trying to figure out I am saying we are going to figure out that beta naught and that beta 1, which is what represents the line. So, we are trying to figure out that line, which minimizes deviation between actual and estimate, so that is kind of make sense good.

(Refer Slide Time: 14:26)



So, how do we go about doing this the way we go about doing this is the process of the derivation. We start by saying this is the functional model we have we have the model which says that y i is nothing but, beta naught plus beta 1 x i, which is the line that we are creating we do not know beta naught and beta 1 is yet. But, if you had beta naught and beta 1 your line would nothing but, beta naught plus beta 1 plus some amount of error just going back.

For instance to this, what we are saying is each y i, which is nothing, but this value this is y 1 is nothing but, is equal to, where you can get to in the line this is equal to this distance and this distance can be defined as beta naught plus beta 1, x i, because this value is x 1. So, this distance y 1 is nothing but, this distance beta naught plus beta 1 x plus some amount of deviation, which I am going to call as error this is the deviation between actual and estimate that y i minus y i had that distance I am calling as the error.

So, ultimately y i is nothing but, beta naught plus beta 1 x 1 plus e i and I shall really say that beta naught plus beta 1 x i plus e i, which is what I have done out here, so said y i is nothing but, the model plus the noise. So, will call that model or you can call that y i hat and the noise, now all I do is just rearrange the terms such that e i is on one side and we have said that our goal is to minimize. So, this is the deviation between y i and y i hat and our goal is to minimize the square of the deviations for each data point.

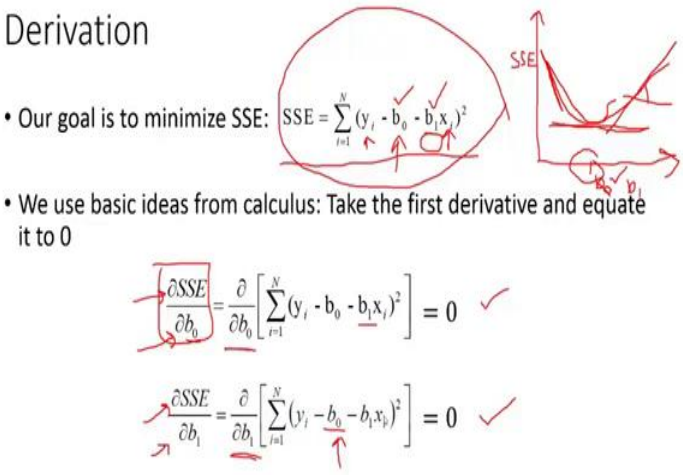
So, I am go through I equals 1 through n I am going through each data points 1 through n all the way and for each data point I am trying to look at the deviation between actual and the model and this is your estimate or you can think it as y i hat. So, this is what you are estimating and this is what is actual value you are taking the difference between them and squaring it and it is the you get the two minus signs because you can think of it is minus and put the beta naught plus beta 1 x i into the brackets, then you open the brackets minus comes in front of both terms.

So, you are ultimately just taking the summation of the square the square term is here the square of the deviation between actual and estimate. And; that is, what we are going to call as a sum of squares error and that is what we are going to try and minimize you essentially want to minimize the squared deviation between actual and estimate. And you can also kind of think of it this is one way of getting into sum of squares you can also think of this definition, which is I started by saying I want to minimize the actual minus estimate square and we know that the estimate is nothing but, so y i hat is beta naught plus beta 1 x i.

See notice the difference y i is beta naught plus beta 1 x i plus the error term, where as y i hat, which is the estimate of y i is just beta naught plus beta 1 x i this basically defines the actual and this defines the estimate. So, you can just plug in this beta naught plus beta 1 x i and I am using the word beta, but really these terms are still b. So, b naught plus b 1 x i will be more accurate. So, b naught plus b 1 x i and when you plug and expand that this is exactly what you get the same notion of sum of squared error, which is cycle through each data point and look at difference between estimate and actual.

So, our goal in determining the beta naught and beta 1 see the y i and x i are data points that you collected from the field x i represents the input variable y i represents the output variable. So, you have 10, 20 or a 100 or 1000 of x and y pairs, so for a particular x there was a particular y and there are I such there are n such x and y pairs and I is just the index that represents a particular combination. So, x and y are actual data points b naught and b 1 is what we are going to determine and the way we are going to determine that is by finding, what values of b naught and b 1 minimize this function. So, that the exercise that we are embarking upon.

(Refer Slide Time: 20:07)



So, how do we do that, like I said a goal is to minimize this term and the way we going to do that is to take a very basic idea from calculus, which is that you take the first derivatives of this term and equate the that first derivative to 0, why do we do that it is a very basic idea from calculus, which is when you if you take on different values of beta naught and beta 1. And right now, these are the two variables of interest y i and x i are actual data the idea that if you fix one let us say you fix beta 1 and you keep on changing beta naught.

For a given beta 1 there exist a beta naught, where this error will be the lowest and, so for a fixed beta 1 if this was the variable beta naught I am plotting the beta naught here and I am plotting the sum of squared error this is SSE out here. The core idea is that you are going to get as you keep on changing beta naught for a fixed beta 1 you might get a function that looks, let me make more smooth I will just try again am I get beta naught you might get function smooth function like this, which basically says like this there exist a particular beta naught value, where the sum of squares is minimum.

Now, how do you go about finding that, it is a simple idea if you take slope of this function the slope of this function at different point at the point at, which the sum of squares error is lowest that is slope is equal to 0. So, the idea is that the slope is nothing but, the tangent to this function just like the slopes always are and this is, what is

considered a positive slope a flat line is considered as 0 slope and this is negative slope on the left hand side.

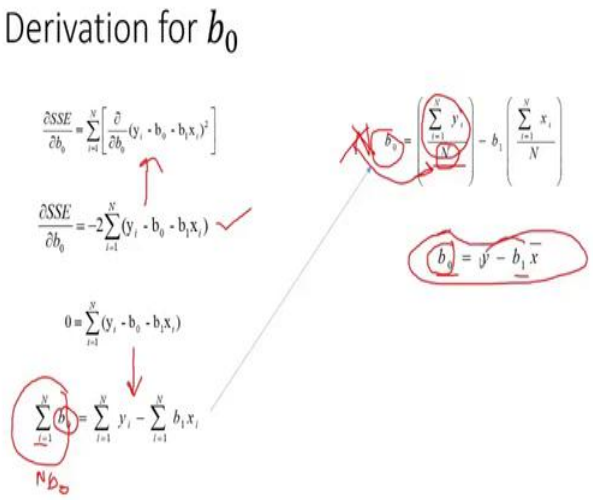
So, the idea is that if I take the first derivative, which is nothing but, the derivative of a of a particular function is nothing but, the slope of the function and if I take the derivative in equate it to 0 and I should be able to find out that value of beta naught, which gives me the lowest value, now remember I of course, said beta naught for a given beta 1. So, am I get that in the form of in the form of beta 1, but then what I can do is then I can do this same exercise that I just did for beta 1 I can say for a given beta naught as I keep changing beta 1, what is that value of beta naught that minimize it.

So, essentially you wind up having a concept two equations with two unknowns the two unknowns are beta two and beta 1 the two equations are, what we get are what you get when you derive with respect to beta naught and what you get when you derive with respect to beta 1, so b naught and b 1 again. So, what you get when you derive with respect to b naught and what you get when you derive with respect to b 1, then you equate that to 0.

And then, you have two equations with two unknowns; that is just a simple form of simultaneous equation for you to solve it. And as you can see what you have done in each of these two is to take the first derivative and these are partial derivatives, because you are clearly deriving with respect to beta naught but, you also have an another variable in this equation beta 1 same here these are partial derivatives. Because you deriving with respect to b 1 and you have a another variable which is b not in the second equation.

So, but the core idea is this, which is you take two partial derivatives of sum of squares error with respect to b one b naught and b 1 and solve for the values of b naught and b 1; such that you will be able to get that b naught and b1, which minimize this sum of squared derivation in a sense. So, we have explain the a principle lets actually go to the steps that how we will do it. Now, first we are going to take care of first equation that we saw, which is this equation, so let just call this one and this is two, now we are going just to do the process for 1.

(Refer Slide Time: 24:26)



You take the first derivative and all, what you might notice, now is that I have essentially brought this differentiation in because it is a sum of that we would looking at the first derivative of a sum of terms, which should logically be the same thing as sum of the first derivatives of those terms and what I have do here is I am differentiating with respect to the beta naught. So, you can do this in many ways you can just basically take this square and just expand it and say what is y i minus beta naught minus b 1 x i times.

Because, it is a square times the same term again and then, you get many terms and then break them up or you can use something fairly simple called the chain rule, which is just that this is the function of b naught. So, I will first take this square of the function and do you know the usual idea that I difference the first difference of x square the derivative of x square is 2 x. So, the two moves out and the derivative of minus beta naught is minus 1, so the minus also comes out.

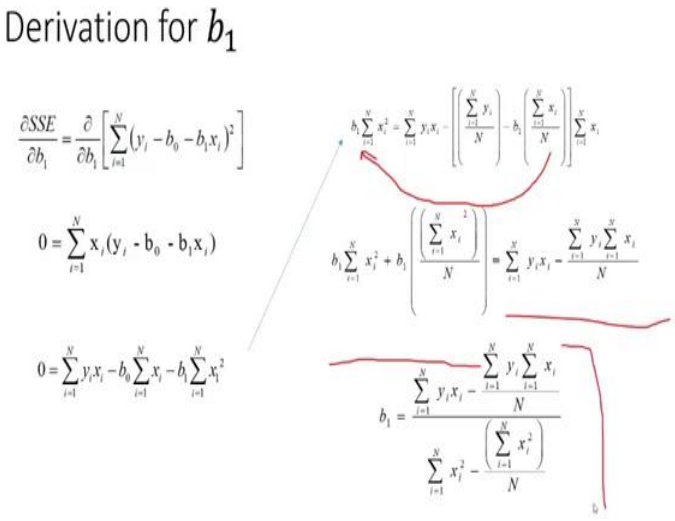
And, so essentially use can use whichever the approach in differentiation that you like, but this is the answer to this step, now all I all I am going to do is equate to 0 and then, I am going start solving it. So, what I do is there are many separate term here again, so summation of a minus b minus c you can basically say it is summation of a minus summation of b minus summation of c I have done that and I have reshuffled the terms in this equation such that beta naught comes to one side, because we are interested in beta naught.

Now, beta naught is essentially a constant it mean it is a variable in this equation, now but it is take on one value it is not like x i, which takes on different value depending the value, what is the value of i is . So, if I am doing the summation from i equals 1 to n each x i will be a different value, but beta naught is not a function of i it is a same beta naught for whatever value of i you pick. So, it is essentially out here is all you are doing is here adding n such b naught and that is nothing but, n times b naught.

So, what will do in the next step is to just isolate b naught and therefore, if you notice we took that n that was coming up on this side and we moved it as the denominator. So, that is, what you are seeing here in terms of moving from previous step to the step. And finally, we realize the sum of y i divided by n, which is the number of times y different y is nothing but, y bar is nothing but, the sample mean the sample mean is nothing but, the sum of your data points divided by the number of data points, so this is the easiest representation.

Again we just simplified equation 1 of a 2 equation combination with two unknown variables hence, b naught is described as a function of b 1. Now, we are going to solve for b 1 by substituting values of b naught with the term on the hand side, so let us do that.

(Refer Slide Time: 28:22)



Here is the derivation of b 1 again a same idea, which is you are doing partial derivative over the summation and again you can take this term in side, which should be fine. And

again you can use the chain rule for deriving it or just basically expand this whole set of terms expanded by the square and you can do it that is your convenience. But, one additional thing is it, which won't look exactly at b naught, because the coefficient for b naught was just this minus 1, where is a coefficient for b 1 is you have the minus, but you also have x I, so minus x i.

So, the answer is also going to look it different, the result of this derivation is this value and essentially you have an x i in brackets y i, but this derivation again should be fairly straight forward once you do this derivation you get this and you again go through the process of breaking this down or simplifying it. Again you had a summation over the entire set you can break up into many summation, so that what we have done here in the next step.

Now, reshuffle things such that the b I the b 1 comes to one side, so that is what we have done here b 1 come to this side and on this side you would have had only these two terms. But, the problem is, so this b 1 I just shifted to this side, which is what you are seeing here, but the look the right hand side is looking different and the reason for that, because in the right hand side only these two terms should have been there.

But, again look it is the function of b naught and we know, now from the previous exercise we know that b naught is equal to y bar minus b 1 times x bar that is was the conclusion and actually show you that value that is just erase we conclude that b naught is equal to y bar minus b 1 time x bar and let me also erase this and that is exactly what we are substituting, what we are doing is we are going and substituting this b naught with that term.

So, yes we have shifted this to this side and that is how you get the left hand of the equation when the right hand side in addition to this y i x i we substituted the value of b not with another term. Again note this summation y i divided by n is nothing but, y bar summation x i divided by n that the two term are here and here are nothing but, y bar and x bar, so we said y bar minus b 1 x bar and of course, this x i out here just stays out here.

So, that should give an idea, where the expansion is and, now again what we are trying to do is we trying to keep all b ones to one side. So, this guy out here also gets shifted here and that is what you are having on the left hand side of the equation and the right hand side this step gets unaffected. And finally, you can just simplify this is just basic

algebraic simplification to get to the final form of b 1. So, what would you do when you given a whole bunch of x in y and you need to fit a line through it you use this formula essentially to get slope b 1.

And if you look there is nothing in this that has been not in it is just a function of y's x's and n and x I squares, but again that is from x I and n is nothing but, total number of data points and you can use this and get b 1, which is the slope of the equation. And then, you can go and substitute b 1 out here and you know x bar and y bar from the data and get b naught and as you know once you have b naught and b 1 you have a line on your hands. And we essentially use this is the process of ordinary least squares regression where you take a bunch of data x's, y's, x and y pair of inputs and outputs and fit a line through that such that you minimizing the square deviation between the line and the line represents your estimated values y for a given x and the actual data point y i hope that was clear and that is your ordinary least square derivation.

Thank you.