Computational Techniques

Prof. Dr. Niket Kaisare

Department of Chemical Engineering

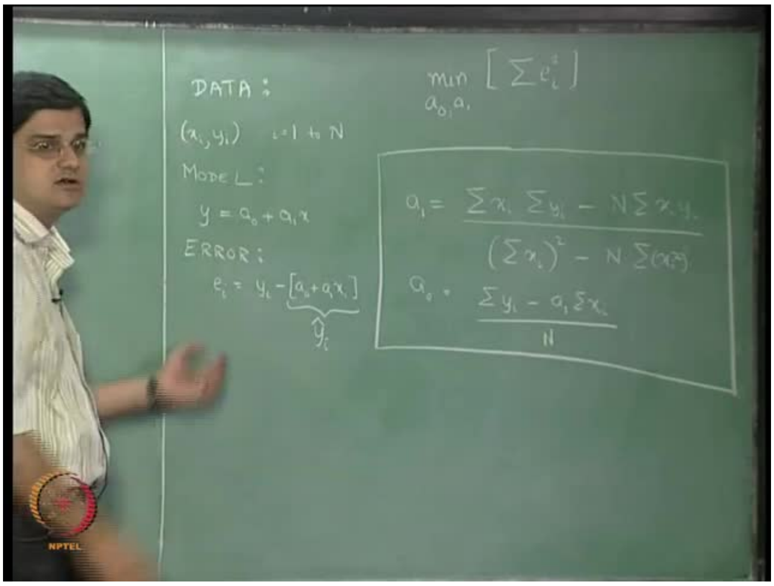
Indian institute of Technology, Madras

Lecture No. # 02

Regression and Interpolation

Hello and welcome to the second lecture of module 5, for our computational techniques course. What we were doing in module 5 - the first part of the module 5 - is to look at various methods for doing regression.

(Refer Slide Time: 00:38)



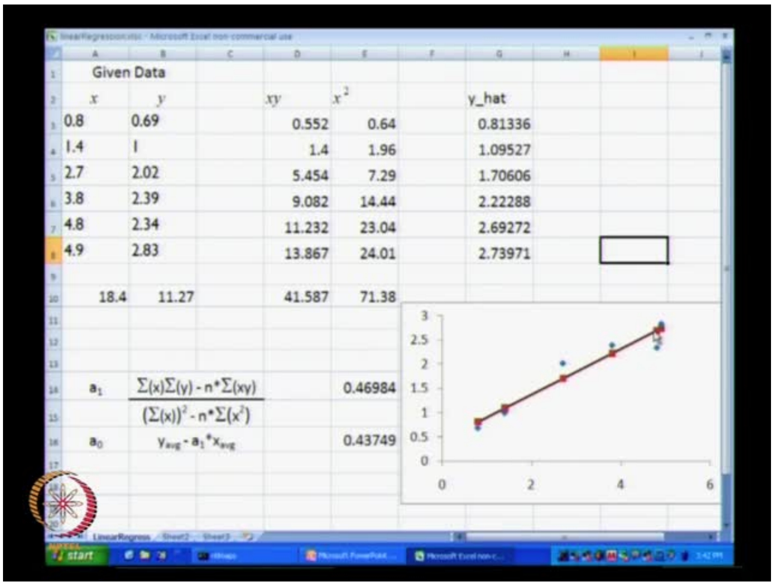
We started off with linear regression - linear regression with one independent and one dependent variable. So, we wanted to get a straight line of the form y equal to a 0 plus a 1 x, where a 1 is the slope of the curve and a 0 is just the y intersect; that is the model, that we wanted to fit to a dataset given to us, x i y i from i going from 1 to n; what this representation means, is essentially we have data x 1 y 1, x 2 y 2, x 3 y 3, and so on up to x n y n.

In this particular data, we have defined our x 1 as the independent variable and y 1 as the dependent variable. The error, we have defined as the actual value of the data, y i minus the value that we get from the model by substituting a 0, a 1 and x i value over here; this particular model prediction by..; so, this is our model; when we substitute the value of x i in this model, the value of y that we get is nothing but y i hat; so, y i minus y i hat term is the error term e i. So, the objective in linear regression or for that matter in any regression, is going to be to minimize the sum of square of errors. How do we define error is essentially going to be important and we will look at that, when we come to advance parts of regression. But for now, we will just minimize the sum of squares of errors between the true value y and the model production y i hat.

In order to minimize the sum of square of errors and to find a 0 and a 1 that minimize the sum of square of errors, what we did was we took this s e; we call this s e to represent sum of square of errors and we differentiated it with respect to a 0 and substituted that equal to 0 and then differentiated with respect to a 1 and then substituted that equal to 0; that gave us two equations and two unknowns; we solved those two equations and two unknowns and we got the value of a 1. The value of a 1 that we got was summation x i multiplied by summation y i minus n times summation of x i y i divided by this term minus this term; that was the value of a 1 that we got and this is the value of a 0 that we obtained. So, this is pretty much what we did in the first lecture of module 5.

What we will start off today in this particular lecture is, take this example, the example of the six data, that we looked at in the first module and get the value of a 0 and a 1 to fit that particular data; that is what we will do first and then we will go on to regression - linear regression in multiple variables and then look at functional forms and so on and so forth.

(Refer Slide Time: 03:40)



So, let us now go and look at Microsoft Excel. So, what I have done over here is the data that we had previously; so, this particular data, I have just taken that in Microsoft Excel; so, this is our x, this data is our y, and on the board we had derived the value of a 1 and the value of a 1 that we had derived was basically, summation x summation y minus n times summation x y divided by summation x squared minus n times summation x the whole squared minus n times summation x squared. And likewise, we had derived basically the equation for a 0 as well, a 0 was y average minus a 1 times x average; so, this is the data that is given.

So, what I will do is, I will just get the summation of that data; so, I will write that equal to sum and so what I have done is I have used the cursor keys to take the blinking rectangle over here and now I press the shift key and use the cursor key to take the rectangle upwards and then complete the brackets; so, closing the brackets we will get the summation, that we see over here. Adjust, move the a 0 down, so that we get more space for doing over stuff. Now, in order to get summation of y, we will just drag this particular cell over here and if we press F2 key, we will be able to see that, this particular guy is nothing but summation of y I; what else do we need, we need summation of x summation of y, that we have already obtained; we need summation of x y and summation of x squared.

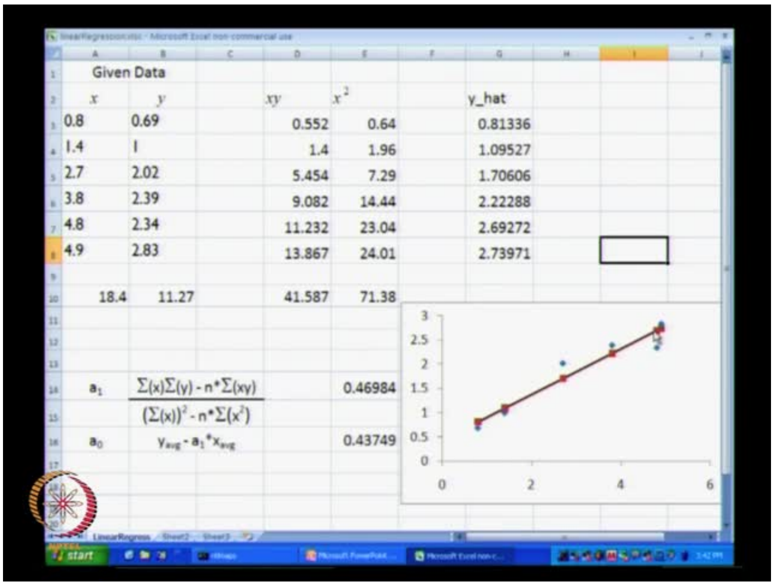
So, what we will write over here is x y and we will just increase the font, change the font to Times New Roman, and this becomes our x y; and likewise, we need x squared and we will just use a super script, so write it just looks good.

So, we will calculate x y, which is nothing but the product; so x y is nothing but equal to x star y that is going to be our x y and we will just drag this for the entire column and our x squared is going to be equal to x carried 2 x - this particular cell the carried sign is the power sign and 2 that is going to be our x square - and then we can drag this for the entire column or we can just double click it and it will get dragged itself; and then we go on to find out summation of x y and summation of x squared; for that we will just copy this and then Control C and Control V over here and we can just press F2 to verify what that is that is nothing but summation of this particular column, which is summation of x y and this is nothing but summation of x squared.

So, our value of a 1 is going to be equal to summation of x multiplied by summation of y that is going to be equal to - I will start the brackets for the numerator - summation of x multiplied by summation of y minus n - now, n value is 6 - minus 6 multiplied by summation of x y, which is this value over here and i click over here and this is our numerator, slash to represent division and our denominator is summation of x the whole squared, summation of x is nothing but this guy, this guy whole squared minus n times, which is 6 times summation of x squared and summation of x squared is this and we close the bracket and that becomes our denominator and this is nothing but our a 1. a 0 is going to be y average minus a 1 times x average or this can also be written as summation of y minus a 1 times summation of x whole divided by n; so, y average is nothing but summation of y divided by 6 minus a 1, which is this value, multiplied by summation of x divided by 6 and that is going to be our a 0.

So, what we did just to recap, what we did is, we were given the data x y, that is over here. What we did in this particular case is, found out the summation of x, found summation of y, then computed x multiplied by y over here for each of these cases.

(Refer Slide Time: 03:40)



So, if you just take any cell and press F2, you will be able to see the formula for that cell; formula for that cell is, the blue colored cell multiplied by the green colored cell, which is nothing but x i multiplied by y i or x 4 multiplied by y 4 in this particular case; this is nothing but x 4 squared, we can see that when we press F2 key, this is f 4 squared.

And finally after calculating x y for all of the six data points and x squared for all of the six data points, we get summation of x y and summation of x squared; we press F2, we will see that is what we get over here and likewise this is summation of x squared is what we get over here and then we can substitute this; so, what i am going to highlight over here is our numerator, our numeretor is summation of x which is this guy multiplied by summation of y, which is this guy, minus six times summation of x y minus six times summation of x y, which is this guy, and if you look at the cell number it is D10 that is our summation of x y. That is our numerator divided by the denominator and that is how we get our a 1 and likewise, we can just easily get our a 2 as well.

If we go back to our regression, that we obtained earlier, we had said y equal to 0.45 plus 0.47 x; essentially, if we are going to truncate these particular values only to two decimal points, that is the approximate value of our straight line we get as.

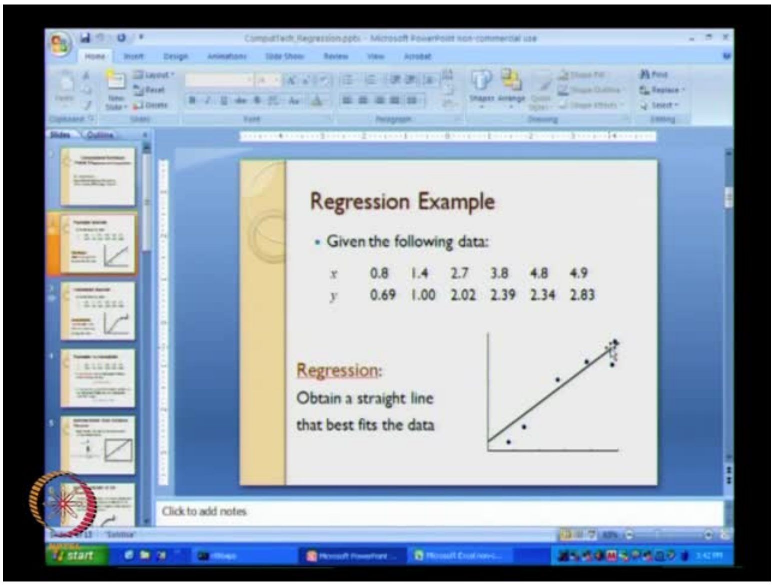
So, what we will do is, we will just plot this data x versus y; so, we will highlight all the data click on Insert tab on Scatter and then we will plot using this Scatter function and this is going to be our original data; this is our x versus y data and then we will get y hat, y hat we said was nothing but a 0 plus a 1 x. So, we will again go to full screen, y hat was equal to a 0 plus a 1 multiplied by x and this is the the first y hat that we get, what we actually need to do in order to drag and drop is, just we do not need a y hat at each and every point; but what I will do over here is, in any way just get this y hat at each point. So, I will give this delta, what this delta does in Excel is, when i drag this particular column downwards, this particular e and 14 will not change.

So, I will just demonstrate that to you. So, let us say, when i just drag this down and do F2 instead of E 16 and E 14 it has changed to E 17 and E 15, because of which we get the value as 0, because E 15 has nothing and E 17 has no value over there. But now instead when i include this dollar signs in Excel and then i drag this, just see what happens, if i do F2 over here, the value within the dollar sign has not changed, only the cell values over here have changed; That is what has happened if you look at this particular expression; it is dollar E dollar 16 plus dollar E dollar 14 multiplied by a 3 and the next expression is dollar E dollar 16 see this particular value has not change this, particular cell value also has not change, only thing that is changed is from a 3 we have gone to a 4.

So, that is what we will consistently we able to do for the entire cell. And now, what we need to do is, we need to plot now, I will plot this x versus y hat also in that data, we can do that by Select Data, click on Add the series name is going to be y hat series, x values are going to be these values, and series y values are going to be y hat values, and that is going to give us this red data, and we will basically do Format Data Series and add a solid line with red color over here.

So, this particular line represents our y hat; the blue data points over here represent our y; and the red data points over here represent our y hat. So, the red data points are y hat, the line is y equal to a 0 plus a 1 x and this distance is our E 3, this distance is E 2, E 1, E 4, E 5, and E 6

(Refer Slide Time: 17:07)

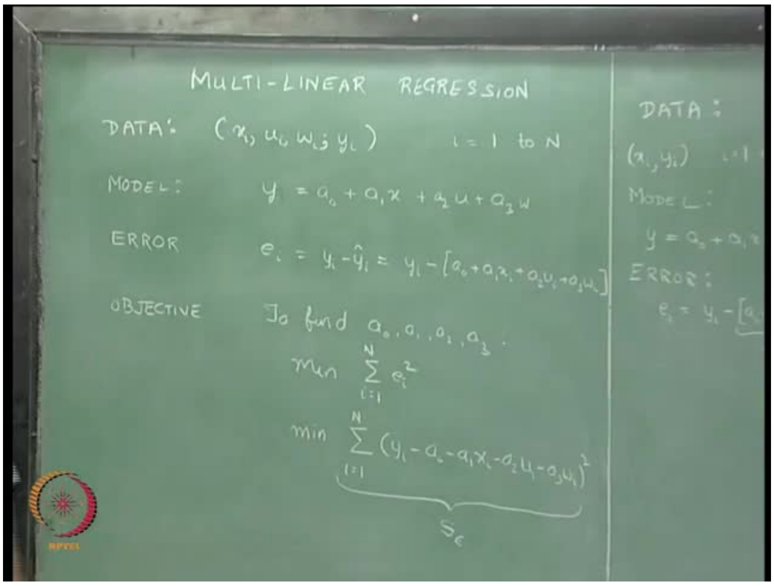


So this is the - best fit line - best fit straight line, that best fits the six data points that we had over here and I had already done that before and this was the best fit line that I had obtained.

So, this is how you are going to do your linear regression in a single variable

Now, let us go ahead and look at how to go ahead and do linear regression in multiple variables.

(Refer Slide Time: 17:26)



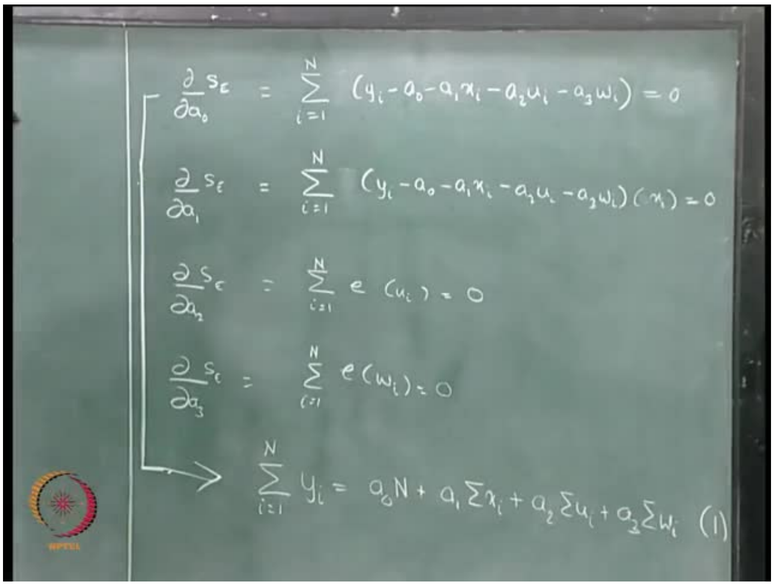
And what I will do is, just for notational simplicity rather than confusing ourselves with x 1, x 2, x 3, and so on I will just take the independent variables as x, u, and w and the dependent variable as y, just for the sake of simplicity; So, the data that we will have x, u, and w are our independent variables; y is our dependent variable; and the data we will represent in terms of x i, u i, w i; y i, where i goes from 1 to n.

The model we are interested in obtaining - the linear model - that we are interested in obtaining is going to be y equal to a 0 plus a 1 x plus a 2 u plus a 3 w. If we have additional variables - say v, m, p, q - it will be a 4 multiplied by v plus a 5 multiplied by m plus a 6 multiplied by q so on and so forth. There is no restriction on the number of variables, we might be able to have in our model, the number of variables that we have in our model is predicated by the functional form of how many terms does the independent variable y depend on. So, the model that we are interested in obtaining is this.

The error e i is going to be nothing but y i minus y i hat which is nothing but y i minus a 0 plus a 1 x i plus a 2 u i plus a 3 w I; this is going to be our error. And the aim in multi-linear regression or objective of multi-linear regression, such that, we minimize this sum of square of errors or in other words and this particular term, we will call this as S e, just as before.

So, in order to find the values of a 0, a 1, a 2, and a 3, what we need to do is, to follow the exact same procedure that we followed earlier; we need to differentiate S e with respect to a 0, differentiate S e with respect to a 1 with respect to a 2 and a 3 and equate each of them as 0.

(Refer Slide Time: 21:48)



So, what we have to do is, dou by dou a 0 of S e, dou a 1 of S e, S e, and dou by a 3 of S e. Now, keep in mind that, we take the summation sign, partial differentiation will go inside the summation sign, because the differential operator is a linear operator and d of f squared. So, let us say, there is a function f, then d by d x of f squared is going to be two times of f multiplied by d f by d x.

So, we will use essentially this property in order to simplify our life. So, what we will get over here is d by d of this particular term with respect to a 0 or a 1 or a 2 or a 3; it is going to be nothing but twice this particular term, that is twice f multiplied by d f by d x d f by d a 0 is minus 1 d f by d a 1 is minus x i d f by d a 2 is minus u i and d f by d a 3 is minus w I; so, that is what we are going to get over here.

So, we will have summation i equal to one to n outside - we will have two times f or in this particular case two times e I - two times y i minus a 0 minus a 1 x i minus a 2 u i minus a 3 w i, this multiplied by d f by d a 0 this particular guy is f; so, d f by d a 0 is nothing but minus one; we can delete the two and the minus one throughout and essentially, we will get the first equation an this equals 0.

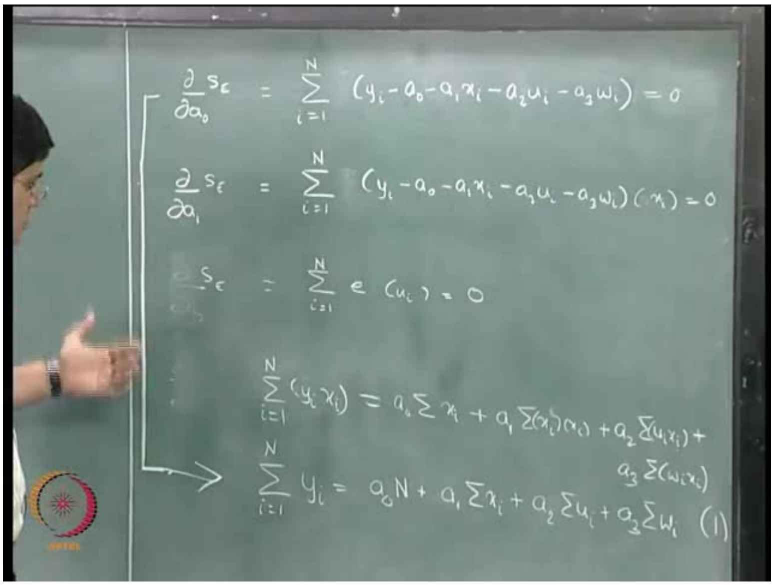
The second equation is going to be, summation i equal to one to n two times y i minus a 0 minus a 1 x i minus a 2 u i minus a 3 w i; this multiplied by partial differential of this term with respect to a 1, there is only 1 term with a 1 over here, and that is this term; so, this will be multiplied by minus x i equal to 0; again we can delete this 2 and we can delete this minus sign and this is going to be the expression.

Likewise, we will have the expression for this as, summation i equal to 1 to n e multiplied by - in this case - u i equal to 0; and in this case, we will have summation i equal to 1 to n e multiplied by w i equal to 0.

We can take all of these guys onto the right hand side so and essentially, what we will we end up getting is four equations in four unknowns. I will just write down the first equation and based on that you will be able to write down all the four equations also.

The first equation, if you look at this particular equation over there, we have taken all of these to the right hand side, so what we will get is, summation i equal to 1 to n y i is going to be equal to a 0 multiplied by summation of one, summation of one is nothing but n, a 0 multiplied by n plus a 1 multiplied by summation x i plus a 2 multiplied by summation u i plus a thee multiplied by summation w i.

(Refer Slide Time: 27:18)

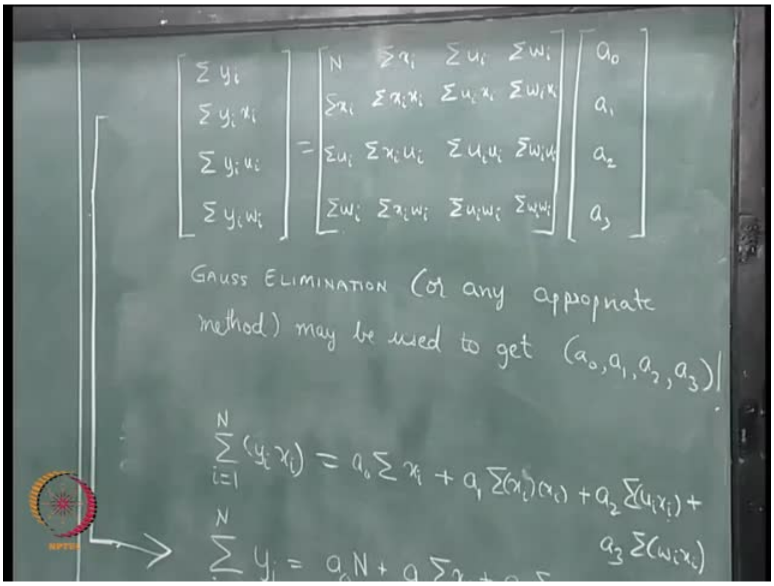


This is going to be our equation one. We will just do this equation also and that we will call that as equation 2 - our second equation I will just erase this particular equation from here.

Our second equation is going to be summation i equal to 1 to n y i multiplied by x i will come over here, summation x i y i is going to be equal to all these terms will be taken to the right hand side and we will have a 0 multiplied by x i summation from i equal to 1 to n, a 0 can come out of the summation sign; so, it is a 0 summation x i plus a 1 multiplied by summation x i squared or summation x i x i plus a 2 times summation u i x i plus a 3 times summation w i x i, that is our second equation.

So, our first equation is, y i equal to a 0 n plus a 1 summation x i plus a 2 summation u i plus a 3 summation w i, that is our first equation. Our second equation is, i equal to 1 to n summation x i y i equal to a 0 summation x i plus a 1 x i multiplied by x i u i multiplied by x i w i multiplied by x i. The third equation is going to be, y i multiplied by, I will just write it in the other sequence instead of x i y i, I will write it as y i x i, so just for consistence notation so i equal to 1 to n y i multiplied by u i is going to be equal to a 0 summation u i plus a 1 summation x i u i plus a 2 summation u i squared plus a 3 summation w i u i and then the fourth equation.

(Refer Slide Time: 29:33)

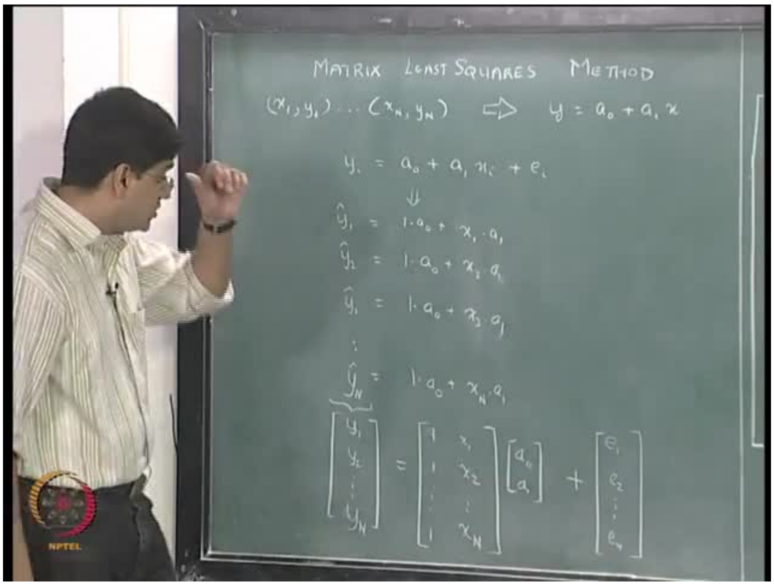


So, now, we have four equations in four unknowns and the four equation in four unknowns, we can write them down in the matrix form; the first equation is, summation y i, the second is summation y i x i, the third is summation y i u i, the fourth is summation y i w i, and you can continue this further if there are more independent variables, it is going to be equal to… and we have this a matrix multiplied by our a 0, a 1, a 2, a 3, and if you are not yet comfortable with the way we have obtained this particular matrix form, I suggest you can go back and look at the videos of module 2 and essentially the lectures 2 and 3 in module 2 are going to cover how we can put this multiple equations in forms of this sort.

So, we have summation x i sorry, we have - n summation - the first term is going to be n then we will have summation x i summation u i summation w i over here, we have summation x i summation x i squared…, next we will have summation u i summation x i u i summation u i u i summation w i u I; keep in mind, the pattern that is forming over here. This is summation of one which is n this is summation of x i summation of u i summation of w i this guy is summation of x i summation of x i x i x i u i x i w i u i u i x i u i u i u i w i w i w i x i w i u i w i w I; it follows a consistent pattern, which that pattern you can keep continuing if you have multiple variables and the same type of expression you will get in n equations and n unknowns also.

Now, we can use any of our methods such as, say Gauss elimination; so, we can use Gauss elimination or any other appropriate method to get a 0, a 1, a 2, a 3, or if we have even more independent variables a 4, a 5, a 6, and so on and so forth. Now, this is one way of doing the - linear - multi linear regression; another way of doing writing the multi linear regression is using the matrix method.

(Refer Slide Time: 33:47)

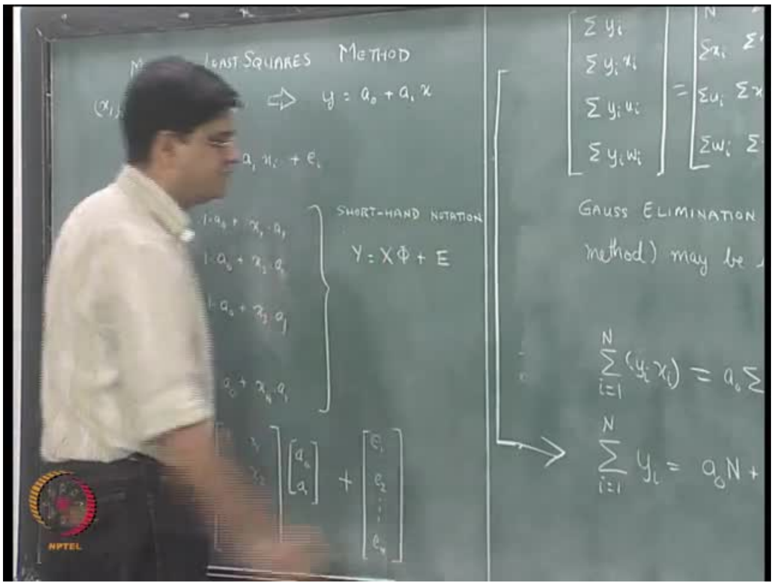


Let us look at the problem in just two variables x and y and then we will extend it to multiple variables. We have the - data x - data is x 1 y 1 and so on up to x n y n. we are interested in fitting the curve y equal to a 0 plus a 1 x. So, the true data is going to be y i equal to a 0 plus a 1 x i plus e i.

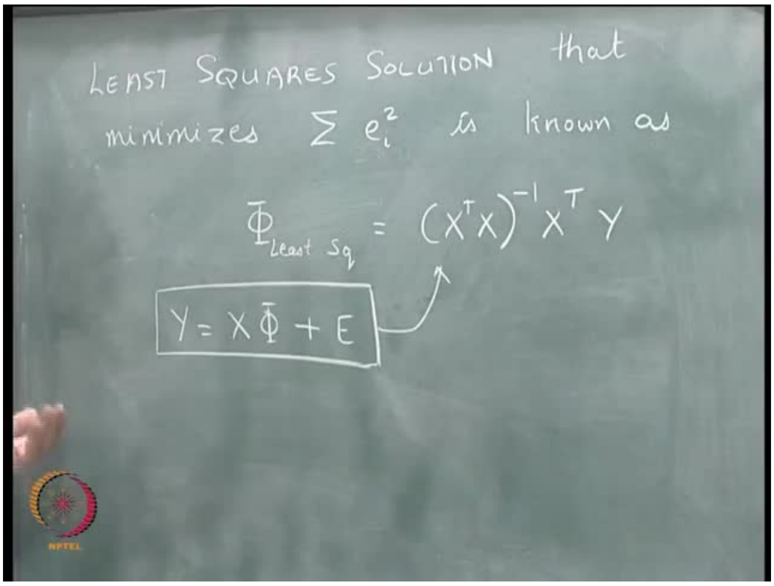
So, this we can write this as, y 1 equal to one times a 0 plus a 1 times x 1; y 2 is one times a 0 plus a 1 times x 2 or more appropriately we will put the hats over here; y 3 hat is one times a 0 plus a 1 times x 3 or more appropriately, i think to write it in a consistent form, is x 1 times a 1, x 2 times a 2, x 3 times a 3. It is the same term - instead of writing as - instead of writing this as a 1 x 1, a 1 x 2, a 1 x 3, I have written as x 1 a 1, x 2 a 1, x 3 a 1, and so on up to y n hat, is going to be equal to 1 times a 0 plus x n times a 1.

This n the matrix form, we will be able to write this as our matrix y 1, y 2, and so on up to y n; this is going to be equal to 1, 1 and so on up to 1; x 1, x 2, and so on up to x n multiplied by a 0, a 1. That so, y hat is equal to this, y is going to be equal to plus the error terms e 1, e 2 and so on up to e n. So, this is essentially what - we are - we are going to get and the short hand notation of writing all this is, y equal to capital x multiplied by phi plus e.

(Refer Slide Time: 37:01)



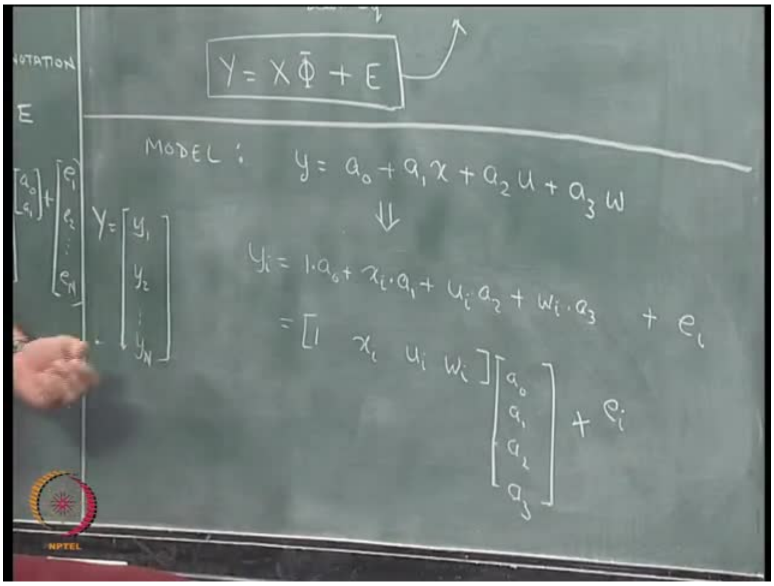
(Refer Slide Time: 37:27)



And the least squared solution, this is the least squared solution for the problem (Refer Slide Time 37:27); as we had written Y equal to X times phi plus E. For example, if we have two equations and two unknowns, then x transpose x transpose x is…, so x transpose x inverse can be written as x inverse multiplies by x transpose inverse x transpose inverse multiplied by x transpose is going to be identity and for a two equations and two unknowns will get this as x inverse multiplied by y. If you recall from the previous lectures on Newton-Raphson's method, we had said that, this particular type of an idea can be used in what is known as the Luenberger mark what method for finding out the solutions; this indeed if with with the same modification, as we talked earlier, becomes the Luenberger mark what method for finding out the least squares. These l m methods were usually not required in linear least squares problem, but in non-linear least squares problem, this l m method becomes very useful.

So, for an equation Y equal to X times phi plus error E; the least square solution that minimizes the sum of square error is given by this particular equation and that is independent of the size of x and the size of y.

(Refer Slide Time: 40:03)

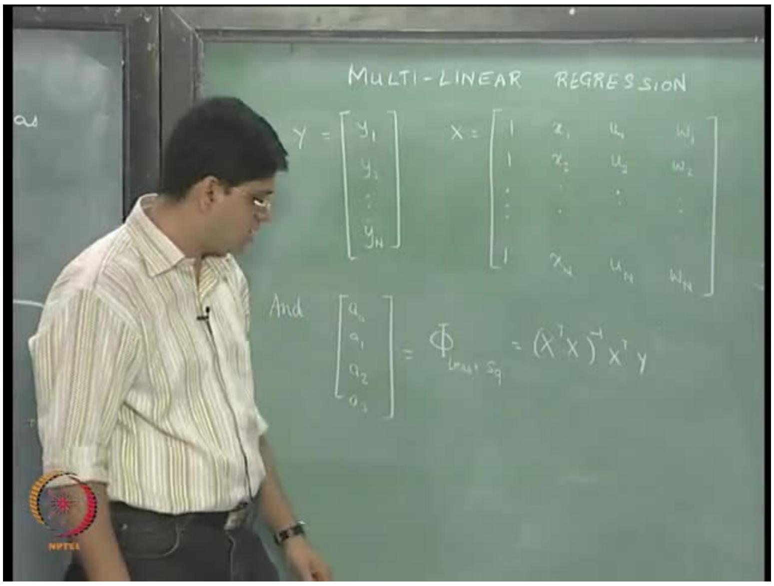


So, now, let us extend this particular problem to multiple independent variables to find y as a function of multiple independent variables; so, if the model that we were interested in is, y equal to a 0 plus a 1 x plus a 2 u plus a 3 w; if this is the model that we are interested in, then our data can be written as, y i equal to 1 multiplied by a 0 plus x i multiplied by a 1 plus u i multiplied by a 2 plus w i multiplied by a 3 plus the error e i; so, y i in other words can be written as, 1 x i u i w i multiplied by a 0 a 1 a 2 a 3 plus e i.

So, the least square solution, again for this particular type of a problem is going to be x transpose x inverse x transpose y, where our matrix y as before is nothing but y 1 y 2 and so on up to y n. Now, the question is how will the matrix exchange previously, what did we have in matrix x, we just had 1 and x 1 so the first column was 1 1 1 1 1 1 n number of times, the second column was x 1 x 2 x 3 up to x n.

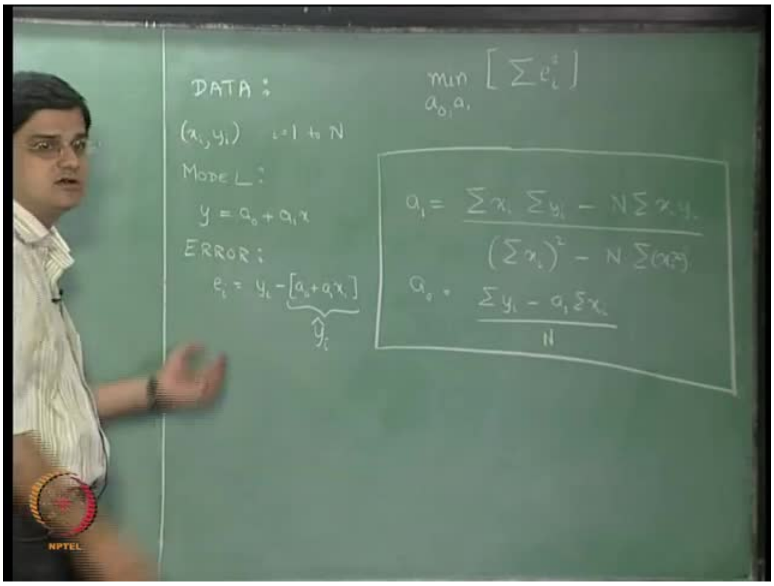
Now, with one x i u i w I, the overall matrix is going to be 1 1 1 1 n number of times in the first column, x 1 x 2 x 3 up to x n in the second column, u 1 u 2 u 3 up to u n in the third column, w 1 w 2 w 3 w up to w n in the fourth column.

(Refer Slide Time: 40:20)

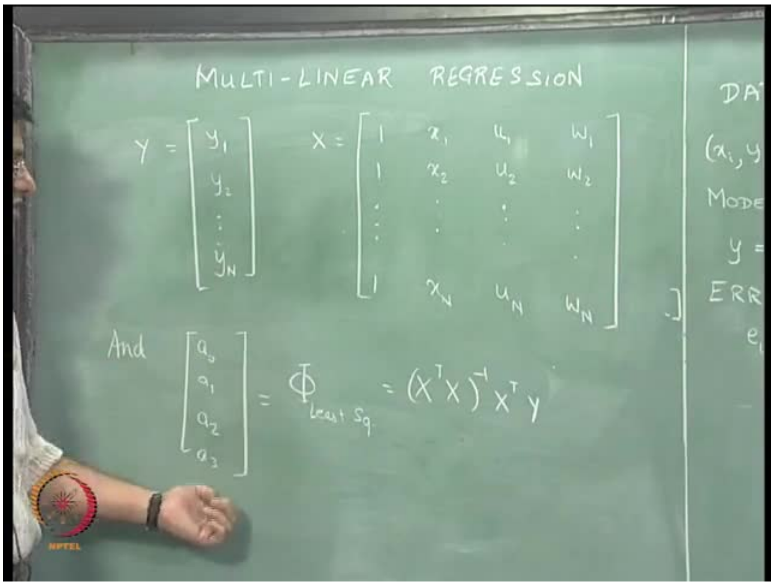


And that is the result of the multi linear regression. When we look at the entire problem using the matrix notation and the advantage of this matrix notation over any other method is that, this is an extremely general method irrespective of what the value of n is or how many number of independent variables that we have in our system; this method can be in a fairly straightforward way extended to this to any number of independent variables.

(Refer Slide Time: 44:31)



(Refer Slide Time: 44:40)

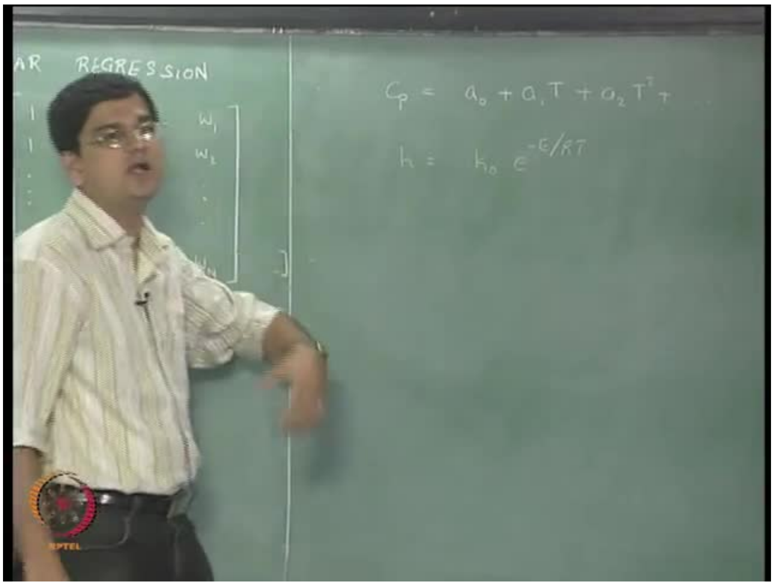


So, that is essentially what we have with respect to multi linear regression. What we have covered in today's lecture is, we started off with linear regression and recapping what we did in the previous lecture and this is what result we had derived in the previous lecture and what we said is an alternative way of doing this, is using the matrix notations and in the matrix notations the alternative way was with y represented as y 1, y 2, up to y n x represented as 1 1 1 and x 1, x 2, up to x n for linear regression in single variable and the phi least squares was obtained from this particular solution.

For the multi linear regression case, the x matrix in general will have a structure of this particular form, the y matrix will have the same structure as before and a 0 to a n we can get or a 0 to a m we can get as phi least squares, again using the exact same equation as before.

Before we went on to the multi linear regression using the vectorial notations what we did was, we covered the multi-linear regression, using the same method as minimization of the function summation of S e and the minimization of the summation of S e resulted in essentially m equation and m unknowns, for example, when we have the independent variable as x, u, and w. We will essentially get the four equation in four unknowns, the four unknowns are a 0, a 1, a 2, and a 3 and then we can solve this equation using any of the linear algebra techniques, that we have learnt in module 3 and we can get the solution of the various parameters, that we are interested. So, this is what we have covered in the first two lectures of module 5.

(Refer Slide Time: 48:42)



In the next lecture, what i am going to cover is, what is known as functional regression, which is essentially instead of having a multi linear regression in variables like x, u, and w, what if we have some kind of a functional form; an example of that as we saw earlier in the first lecture of this module was the specific heat, that can be written as some a 0 plus a 1 t plus a 2 t squared plus so on, that is one example.

The other example, we had seen earlier is, what if the rate of reaction is written as k 0 times e to the power minus e by r t; how we can convert it into linear case? We will not just tackle this particular example, but we will consider a couple more examples where you can convert this particular functional form into linear regression or into linear regression in multiple variables. And once we reduce it to a convenient multi regression form, we can then use any of the multi regression tools that we talked about today either using the matrix method or using the method that we talked say twenty minutes earlier, using either of these methods we can get the value of a 0, a 1, a 2, and so on.

That is what we are going to cover in the next lecture.

Thank you and see you later.