Probability Methods in Civil Engineering

Prof. Dr. RajibMaity

Department of Civil Engineering

Indian Institute of Technology, Kharagpur

Lecture No. # 40

Regression Analysis and Correlation (Contd.)

Hello and welcome to this lecture. Today'slecture is basically the continuation of thelast lecture; weare discussing in this lecture aboutregression analysis and correlation. Wehave seen in the last class there are some of thisbasiccorrelation, basic regression analysis. Wehave seenits mainly that linear regression andthose thingsafter thatwe arein today's class we are going to coverthat remaining part ofthis correlation analysis. Firstthing that we will go is that multivariateregression. Inthe last class, we have seen that in the simplelinear regression, what you have seen? There is oneinput variableand other one is thatindependent variable, that is yversus x.

So that if the x is yourinput variable and y is your target variable or your independent variable then we have seen that how we can estimate theparameters for thisregression model and we have done how, what are theoriginal variance for the y and then after the regression what is the conditional variance and all.

Basically, we have also seen that through this regression what we are tryingto do. How much, what is the extent of thevariantsof thattarget variable; that is, y is being reduced. So, herein this multipleregression what we do is that now this input variables are not 1, is more than 1. Inthat waywe have to use the information of all those input variables and we have todevelop aregression model for that target variable y. So, here the inputs are say x 1, x 2, x 3 andlike that up to; obviously, x m. So, there are, there could be some m variables, what should be the input?

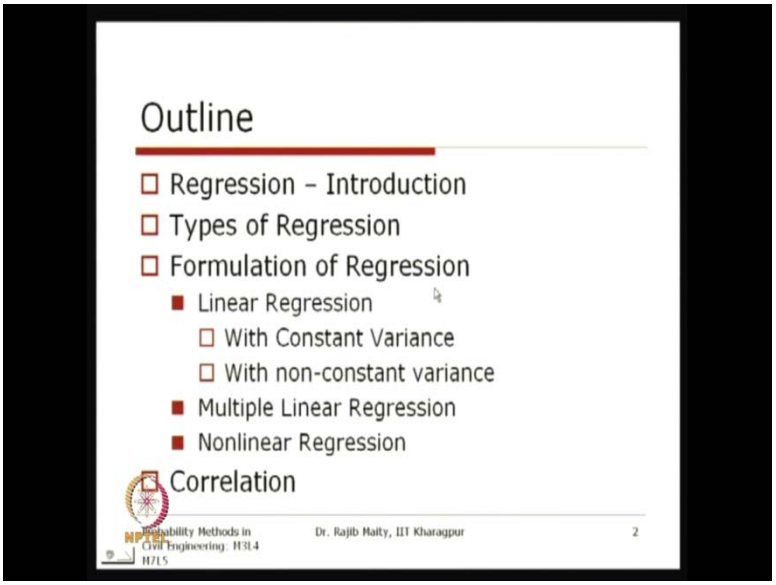
Now, if we just try to extend that analogy of thissimple linear regression to this multiplelinear regression, first we will take thatthen from this 1 dimensional to the 2 dimensional. Aswe are discussing in this last class that it is basically we are trying to take astraight line when it was asimplelinear regression. But, in case now, if I just extend it instead of one input if it is two input then, basically it is a 3 dimensional space.Thatyou

can imagine and through that 3 dimensionalscatter plots of those points because, one point now will correspond to the three entries. Oneis that from x 1 other one is x 2 and the target isy.

So, basically one point consists of these three a pair. So, that the threeentry, the three data point x 1, x 2 and y. So, in the 3 dimensional scatter plot basically we are trying to fit asurface; now depending on if it is a linear regression then that surface will be a plane surface and now that surface should be the best fit to that through those points. And, through for that one following the same principle of this simple linearregression, we have to find out that it should befitted in such a way that the sum of square error should be the minimum.

So, this is basicthattransition from the simple linear regression to the multiple linear regression and based on this we will see what are the theories involved in this. So, we arecontinuing with thatregression analysis and correlation.

(Refer Slide Time: 03:51)

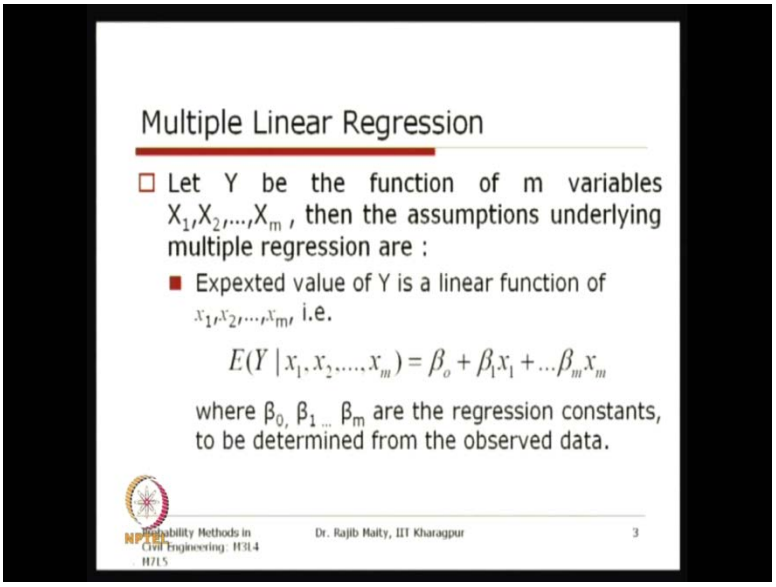


As we have seen in thislast class, that in thisdifferent types ofregressionthatwe have discussed and there in the last lecture we havecovered the linearregression and here we will seehere today's class we will start with this multiplelinear regression.

After we complete this one,we will see whatis this non-linear regression and this non-linear regression can also be for both the cases. It may befor the simple linear regression

and also for the multiple linear regression and then we willgo through this correlation and you know that this correlation we havediscussed earlier also. But, here in the context of this regression analysis,we will once againsee thisaspect of this correlation. Basically, through this measure we aretrying toidentify what is the, how perfectthe model that we have selected. So, that we will discuss under thiscorrelation.

(Refer Slide Time: 05:01)



Well, tostart with for thismultiplelinearregression, as I was justtelling that there will be onetarget variable, we generally call is that independent variable, sorry, that dependent variable this is our target variable y and there are more than one independent variable which are X1, X2, X m.

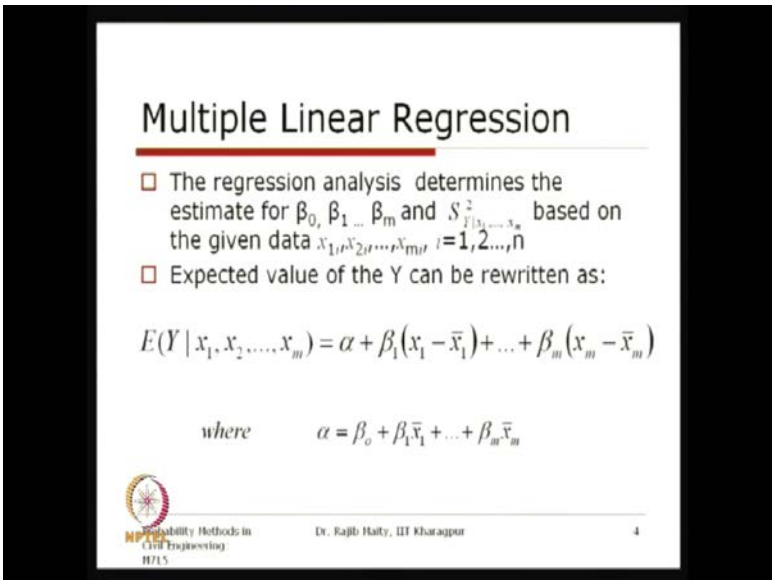
So, let Y be the function of m variables X1, X2, up to X m; then the assumptions underlyingthe multiple regression are again following the same principle that we have discussed for the simple linear regression in the in the last lecture that we are trying to find out what is the expected value of this y given the input of this X1, X2, X3 up to X m.

So, this isthe way that we are expressing thatexpectation of this target variable Y when the specific values of the input variables are given and this is expressed through thisthrough this linear regression as we are now referring to this linear regression and thatthe linear regression is that beta naught plus beta 1X1 plus beta 2X2 up to plus beta m X m.

If you recall, that in thatlast one we are having only one input and we areusing this beta naught andbeta 1X and there are only two parameters for this regression was there; beta naught and this one coefficient with that input variable. So, as you can see here there are m different inputs. So, this beta naught beta 1, beta 2, beta 3 up to beta m - these are our regression constantsand this is to be determined based on thedata that is available to us.You now can see that for this X1 if there are suppose that n numbers ofthat groupdata is available. Then, we are having that n numbers of X1 n numbers of X2 n numbers of X m.

So, through those points through those n points we have to fit one plane surface. Obviously, when we mention we are towe are referring to the context of these2 input variable in a 3 dimensional case and obviously, that concept can be extended to the higher dimension.For example, here it will be them dimensionalspace that you can imagine. So, these constants are to be determinedthat is thebasic underlying thing in this multiple linearregression as compared to the simple linear regression where the number of input was only one.

(Refer Slide Time: 07:53)



Thus, this regression analysisdetermines the estimate forbeta naught beta 1 up to beta m and that S y square that is the variance of this y given this X1, X2 up to X m based on the given data X1 i, X2 i up to X m i and i varies from 123 up to n. So, this i that the substitute that is used here that is basicallyrepresents the number of data that is available

to us and this 123 up to m represents this. So, this m represents that how many inputs that we are having now this S y square given X1,X2,X3 up to X m is that the conditional variants of the targetvariable that is y.

So, this conditional variants should reducewith respect to this unconditional variants which is that S y square and how much is this reduction that we canrelate through therelate through that. So, more the reduction is betterthe model andat the end of this lecture as I was discussing, that at the end of this lecture,we will see that how that information is related to thatcorrelation coefficient.

So, after somerearrangement of this equation that we have seen here that isexpectation of y given this X 1, X 2, X m equals to beta naught plus beta 1X1 plusbeta 2X2 plus beta m x m this we can rearrange to get the another form of such of the same equation which is equals to the alpha plusbeta 1 into X1 minus X1 bar plus beta 2 into X2 minus X2 bar like that up tobeta m into X m minus X m bar.

So, this X1 bar,X2 bar or X m bar is the mean of that particular input of that particular variable. So, as I was telling that Xi X1 is that first input variable and this I can vary from one to m. So, we are having n data and this that mean of thatinput variable is represented this X1 bar. Basically, this alpha what you can see now is basically one adjustedconstant. Onceagain, including that beta naught and if you see fromsee that, alpha can be expressed like this; thatalpha equals to beta naught plusbeta 1X1 plus beta 2X2 bar up to beta m X m bar. So, this once we know this data basically this X1 bar X2 bar X m bar are known and. So, once we get the estimate of this beta 1 beta 2beta3 up to beta m and alpha with the help of that we can estimate that beta.

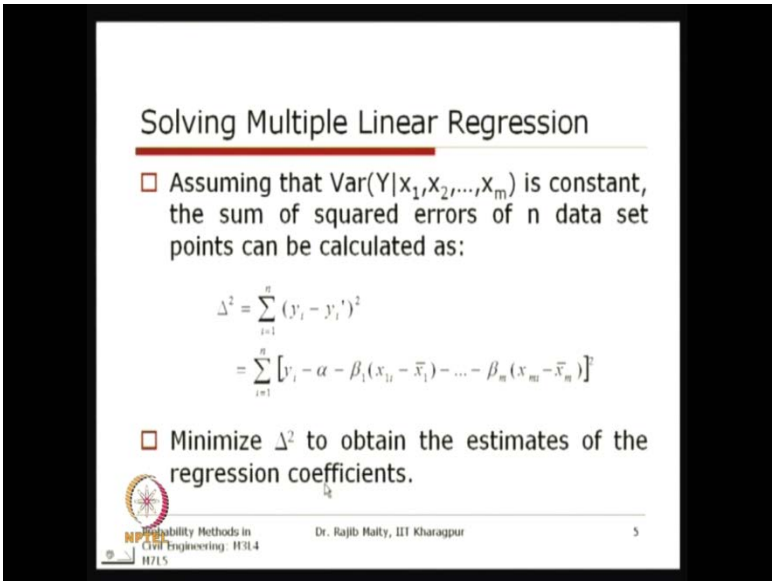
So, here in this expression what we cantarget is that we will first estimate these parameters alphabeta 1, beta 2, beta 3, beta m and with the help of this we will get what is the beta naught and we will get the final form of this regression equation like this.

So, now again we will use that sameprinciple that we usethere in the simple linear regression there basically as I was telling we are fitting a straight line and we are trying to minimize that error and that error means what is the error with respect to the modeledtargetvariable and what is the observe. Herealso, what we will do we will estimate these parameters in such a way. So that, thatestimate of thatof that variable Y and the observed Y their difference should beminimum. So, thisdifference now what

isactually observed and what isestimated fromthismodel is basically is your error and that error we should make it square and once we make that square and sum them up.

So, that is thesum of square error and with respect to that we willtake the partialderivative with respect to all thisconstant that we are suppose to determine and that if we equate to 0 in the sense that we are minimizing thaterror sum of square error and we will get some simultaneous equation and we can solve thisthing that we will seenow.

(Refer Slide Time: 12:35)



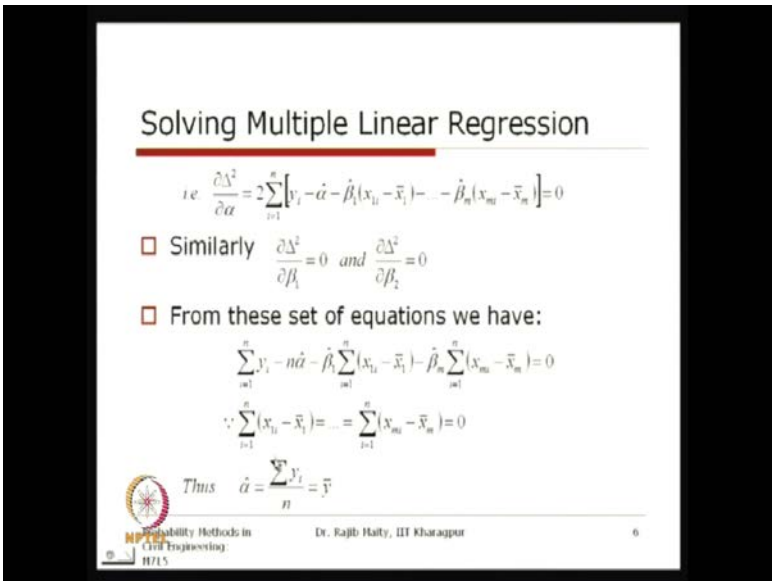
So, now here againif we recall that again from this simple linear regression there are two cases. Oneis that the variants of Y with respect tothat input variable whether that isconstant or that can also vary. So, that in the in case of this simple linear regression we have shown one case that where the variantscan even vary with respectto the range of the input variable x.

Similarly, here also with respect to which zone we are talking about with respect to thatthe input variables. Ifthe variants of this target variable that is Y is constant irrespective of this which zone that we are talking about, the combination of this input variable that is y that is X 1, X 2, X m, thenwe cansay thatis either constant that variants is either constant or that can evenvary. But, if it varies then that form that is how it is varying over this zone; so, that function should beknown. So, now, if we assume that first case thatis theconditional variants is constant then the sum of square error of the end data set points can be calculated as this.

So, this y i is our actuallyobserved target variable and y i prime is thatmodeledvariable that is we are getting from this regression equation. So, difference between them square them up and thensum up for all the individual observation that is n numbers of observations are there.

So, this quantity is giving that sum of square error now if we replace this one thisy i prime then it will come that this alpha minus beta 1 into x1 i minus x 1 bar minus beta 2 into x 2 i minus x 2 bar minus like this up to beta m into x m i minus x m bar and fullquantity square. So, this will give you that sum ofsquare error. Now, we have to minimize thisdelta square to obtain the estimate for the regression coefficient.

(Refer Slide Time: 15:00)

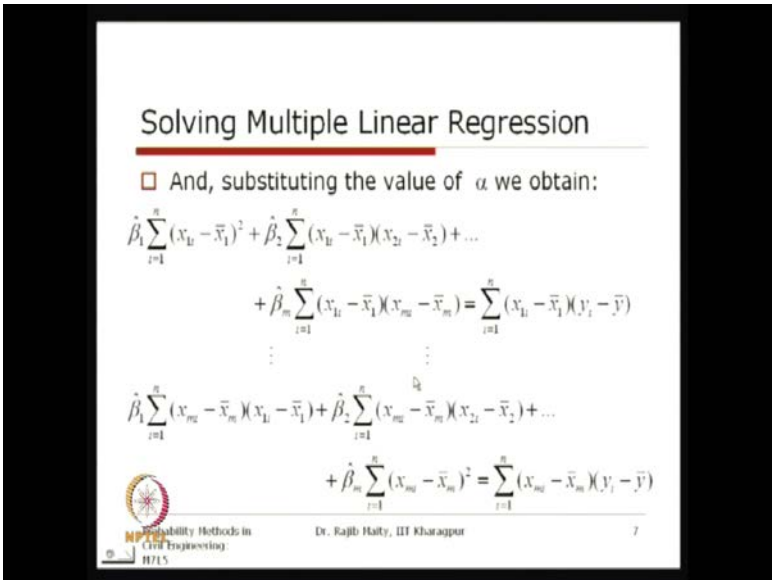


So, to minimize this one as we have seen in thissimplelinear regression also that with respect to all theseconstant we have to make it equal to 0 similarly that for this setdeltathis delta square dou alpha is equals to 0 similarly for this beta 1, beta 2, beta 3 up to beta m like thatall these partial derivatives should be equals to 0.

So, if you take this first one thatdou delta squaredou alpha and then we will we can take this onethis partial derivative and that if we equate to 0 then the form comes like this where we can if we take this summation inside and we can see that this x, x 1 i minus x 1 bar. Basically, if we take this difference and sum them up; obviously, here power is 1. So, if we take that, sum them up, this will become 0.

So, like that for this x 2, x 3 up to x m all these quantities will become 0 as it iswritten here. So, for all these quantities it will become 0. Now, if we just put this expression, these values here then, it will reduce to this form like this that alpha cap is equals to summation of y i; obviously, i from 1 to n divided by n. So, this is basically the mean of thatobserved target variable Y. So, this estimate of this alpha from this way we can see that it is the mean of Y now the remainingthere are remainingn equations are there which are the partial derivative with respect to all n betas basically this beta 1 beta 2 up to beta m and there also we will get that alpha and this one this expression the estimate of this alpha if we just put it back.

(Refer Slide Time: 16:48)



So, we canget a set of equations like this that thisbeta 1 estimate of this beta 1 this hat means here the estimate of this beta 1 multiplied byx 1 i minus x 1 barwhole square plus beta 2 barthis x 1 x 1 i minus x 1 bar into x 2 i minus x 2 bar like that all these plus and up to this beta m which is equals to that particular variable x 1 i minus x 1 bar multiplied by y i minus y 1 bar.

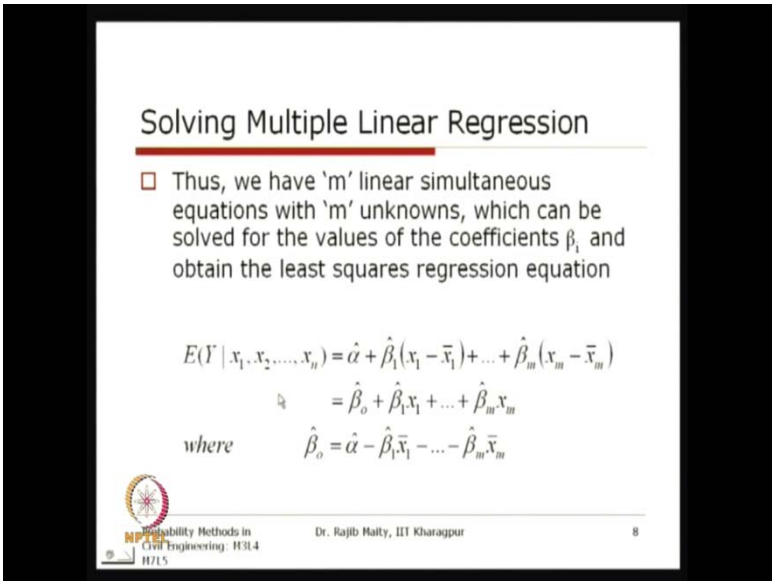
So, this summation for these all in observation and so, this is the first equation. Likethat, we can get all other equations and up to that m-th equation will be like this that beta 1 cap into x m i minus x m bar into x i minusx i bar like that for this beta 2 andat beta m it is x m i minusx m bar whole square which is equals to x m i minus x m bar into x y x i sorry y i minus y bar multiplication summation for 1to n. One thing thatyou can see here

when we are taking the partial derivative with respect tobeta 1, that is the first equation, this quantity is becoming square. Basically, this quantity is that x 1 i minus x 1 bar is basically getting multiplied for all theseleft hand side of this equation for all othervariables.

So, these things x, x 1 i minus x 1 bar this one is there for all these entry and here also right side also you can see that this is the targetthis is related to the target variable Y and this one is that related to that for which constant we are taking the partial derivative here it is for beta 1. So, we can say that this quantity is this. Like that for this last one which is the m-th equation the partial derivative taken with respect to the beta m. So, you can see that what we have seen for the first expression is here that is x m i minus x m bar square. So, this quantity is multiplied with this all otherexpressionin the left hand side and also on the right hand side withwhich is this function is related to the target variable which is also multiplied by this x m i minus x m bar.

So, similarly you can see that for other variables also. So, if it is for this beta 2 then the first quantity will be that x 1 i minus x 1 bar multiplied by x 2 i minus x 2 bar and the second quantity will be x 2 i minus x 2 bar whole square and like that right hand side will be x 2 i minus x 2 bar multiplied by y i minus y bar. So, in this way these are the simultaneous m equations are there and there are m unknowns.

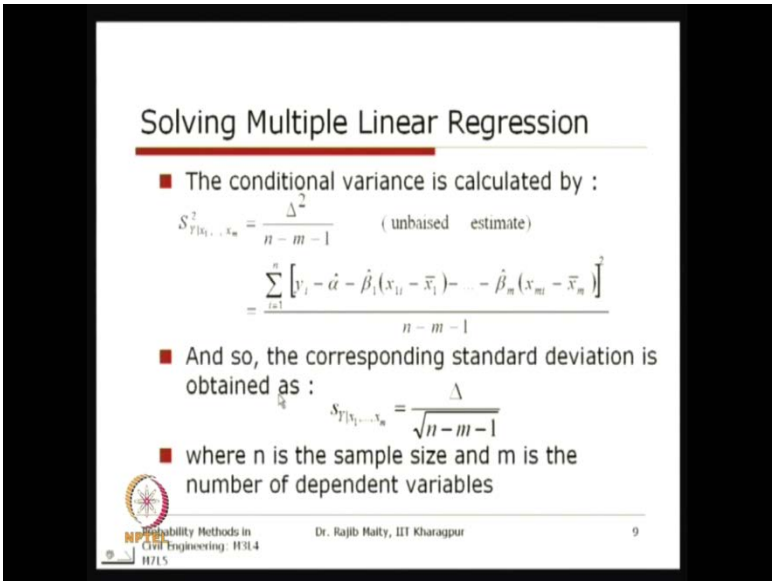
(Refer Slide Time: 19:44)



Thus, we have the m linear simultaneous equation with m unknownswhich can be solved for thevalues of the coefficient beta i. So, beta 1, beta 2, up to beta m and obtain the least squares regressionequation. Once we get this estimate finally, the expression that linear regression expression that we are getting isthat alpha cap plus beta 1 cap into x 1 minusx 1 bar plus up to this beta m hat means that the estimate of beta m into x m minus x m bar and now if we write in terms of this first equation that is the beta naught hat plusbeta 1 hat x 1 plus beta 2 hat x 2 plus beta m hat x m.

This beta naught is now that betathe beta naught hat that is the estimate of this beta naught equals to alpha hat estimate of this alpha minus beta 1 hat x 1 bar minus beta 2 hat x 2 bar minus beta m hat x m bar. So, this is the final expression for this multiple linear regression that we will get.

(Refer Slide Time: 20:57)



Theconditionalvariantsgiven that all those input variables will be that delta square this is the sum of square error divided by n minus m minus 1 which you can see that this deltasquare is the summation of this y i minus this alpha hat minus beta 1 cap. So, this is the sum of square error divided by n minus m minus 1.

So, now you recall that when there isoneonly oneinput was there in case of the simple linearregression if you recall that equationfor this conditionalvariants was delta square divided by n minus 2. So, there m was only 1. So, one input was there. So, if you put m equals to 1 here you will get the same expression that is delta square divided by n minus

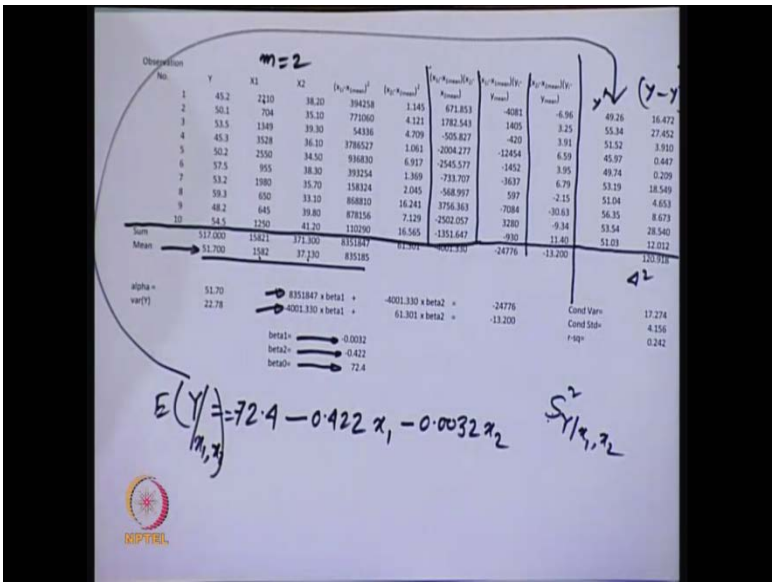
2. So, ifm increases, here their m is more than 1. You have to put that value and this n is this numberofdata points that is available. So, this is one unbiased estimatewhyyou we have already discuss earlier that unbiased estimate. So, that number of degrees of freedom is lost; that should bereduced.

So, here you see that n minus m. So, there are m estimates are there for this x 1 bar, x 2 bar up to x m bar this isthe estimate that we arethere arethose estimates are there. So, m degrees of freedom is lost here again one more parameter here the alpha hat is there that is that we are calculating.

So, it is that basically m plus 1 degrees of freedom is lost and in case of simple linear regression, there are two constants were there that is two estimates. So, 2 degrees of freedom waslost there. So, it wasdelta square divided by n minus 2 in that simple linear regression. So, this is the unbiased estimate of the conditional variants of y.

Similarly, ifthat correspondingstandard deviation is the positive square root you know. Thisis delta by square root of n minus m minus 1 where n is the sample size and m is the number of dependent variables. So, now we will take up oneproblemhere. So, you can see that there aremanyapplications can happen in the civil engineeringproblem.

(Refer Slide Time: 23:26)



If you just seehere for this particularproblem here, this could beany say for example, that we are talking about what should be thetemperaturewith respect to its altitude and

latitudegenerally the altitude increases. So, we know that the temperature will decrease andif the latitude also increases the temperature generally decreases.

So, suppose that type ofdata if we just takehere. So, this is the number of observation that 10 numbers of observations are given and the this is the target variable of this Y. So, you can see that it is 45.25.1 like this. So, these are the ten observations that you cansee here and these are the inputs of this see here we have taken the two inputs only x 1 and x 2. So, here m is equals to your 2. I can write that m is equals to 2 here. So, this is your x 1 and this is yourx 2. So, this is up to this you can see that this is thatdata that we aregetting.

Now, to get that estimate if you want to know that. So, these three columns that you can see is the input. So, first thing that you have tocalculate is that this could be used in a general trade sheet; just to explain that how these things can happen. First what we cancalculate, that this x 1, x 1 i minus x 1 bar that mean that we calculated here. So, thatsquare up for this one. So, basically thisx 1 minus this mean is calculatedhere. So, this minus this that square will be give you this value andsimilarly, for this allsuchvalues we can calculate this one.

Similarly, this is for thex 2 minus x 2 bar that square that you can calculate and then this is yourx 1 minus x 1 mean multiplied by x 2 minusx 2 minus that x 2 mean and their multiplication; it is basically is that column end. So, this entry minus this multiplied by this entry minus this mean. So, this is your meanrow that you can see here and this is the summationand there are ten observations are there. So, this 2 to 10 minus this value multiplied by this 38.2 minus this 37.13 this is the mean if you multiply this we will get this one.

Similarly, this column is forthat x 1 minus x 1 mean multiplied by y, yiminus y i mean. So, if yousee this onehere also basically, this minus this mean multiplied by this minus this mean we will get these values. So, hereagain this column if you see that then this y 2 minus y 2 bar multiplied by y i minus y bar. So, this x 2 minus this x 2 mean into thaty minus y bar that multiplication if you take we will get this value similarly for all these ten observations we have calculatedand the last row that you can see is their summation.

So, up to this we can just calculate first directly based on whatever the data that we are having and then you know that estimate of this alpha is equals to your thaty mean. So,

this is directly the 51.7 that we have seen for the mean for this Y now the variants of Yalso you can calculate whatever the we have seen this thethat Y that we can calculate thevariants of thatfrom this sample estimate that we discuss earlier you will get this 22.78.

Now, this expression if you refer to thisequation here that is that; just here you have toput that m equals to 2. So, this value minus this square multiplied by this beta 2 intothis one this value will be equal to this one. So, all these quantities have calculated now here. Using this information, these quantities that just now what we have calculated we will justset. So, here there will be two simultaneous equations. So, these two equations are written here. So, this is your that firstquantity multiplied by beta 1 plus this quantity multiplied by beta 2 is equals to the right hand side.

Similarly, this is the first equation; this is the second equation that we get and if we solve thesetwo equations and there are two unknowns beta 1 and beta 2. So, you will get that this is your beta 1 and this is your beta 2 beta 1 is minus 0.0032 and beta 2 is minus 0.422. So, with this estimate and that alpha also we know. So, that beta naught we can calculate that is beta naught as you have seen here that beta naught is equals to youralpha cap minus beta 1 hat x 1 bar minus beta 2 hat x 2 bar.

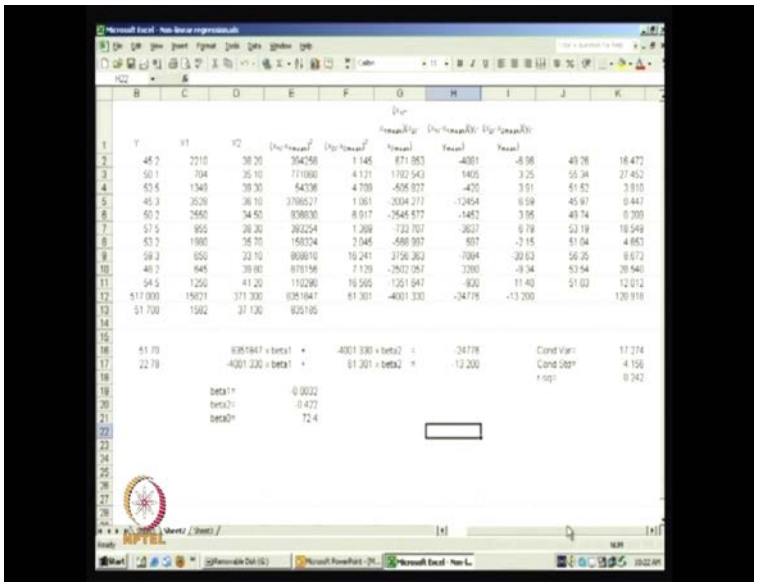
So, like if you use that one. So, beta 1 minus thismean of this twothat alpha minus beta 1 estimate multiplied by the mean of x 1 minus beta 2 estimate into that x 2 bar then we will get what is yourbeta naught. So, once we get these three information that is your this is your beta naught this is beta 1 and this is beta 2 then we are getting this full expression like this that Y is equals to your 72.4 which is your beta naught minus 0.422 this is the estimate for this beta 1 multiplied by x 1 minus 0.0032 x 2 is thatexpression.

Now, if we use this expression and use this what is this inputthen what we are getting this is the estimate of this Y basically you can say thatis the expectation of Y given x 1 commax 2 equals to this. So, this quantity is basically is calculated in this column. So, which is the estimate for this Y? So, I can write that y prime and these are the error. So, this Y minus Y, Y prime; these are the error and thatsquare. So, if I just square for all these 10 observations and sum them up this is basically is your that capital delta square that we are getting here.

So, once we get that capital delta square then we can calculate what is the conditional variants. Thisconditional variants means as we have seen that S y is squaregiven that x 1 comma x 2. So, this if we calculate we will get this 17.274 and the conditional standard deviation is thispositive square root of this one which is 4.156 and ther square value thatwe get is that 1 minus.

So, how much is this reduction? So, you can see that these variants of this Y that is unconditional variants was 22.78 and the conditional variant is 17.274. So, that we canseehere and thisthat r square is the. So, that 22, sorry 24.2 percent has is the reduction.

(Refer Slide Time: 31:36)



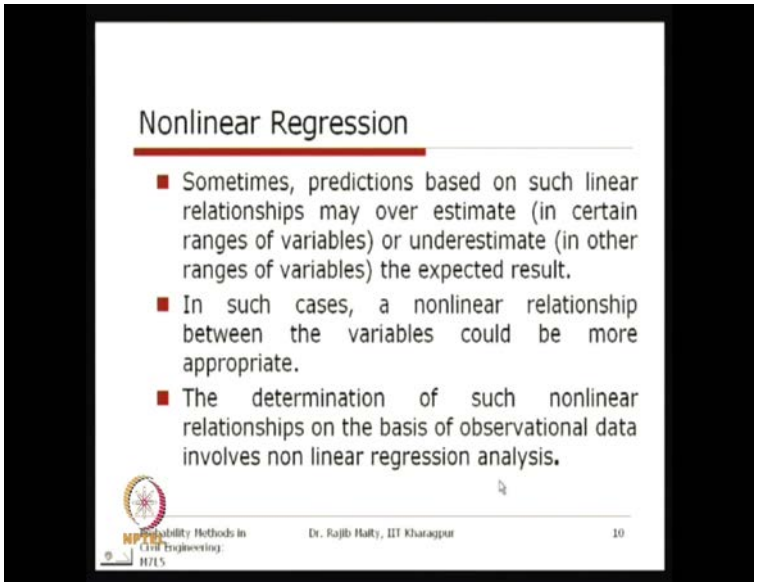
If you want to see this one in this larger font, you can refer to thisexcel file which you can see here moreclearly whatever the calculation that we have done.

So, with this wecan say here once again thatnow, when should I say that thismodel that what we have whatever we have we got that isstrong enough or not? That we will see that in terms of this correlation coefficient that we are going to discuss now and again this will be,this is basically a part of this hypothesis testing.

Once we estimate one parameter and that parameter whether that is significant or not that can be tested through thishypothesis testing that we havecoveredearlier. So, through that

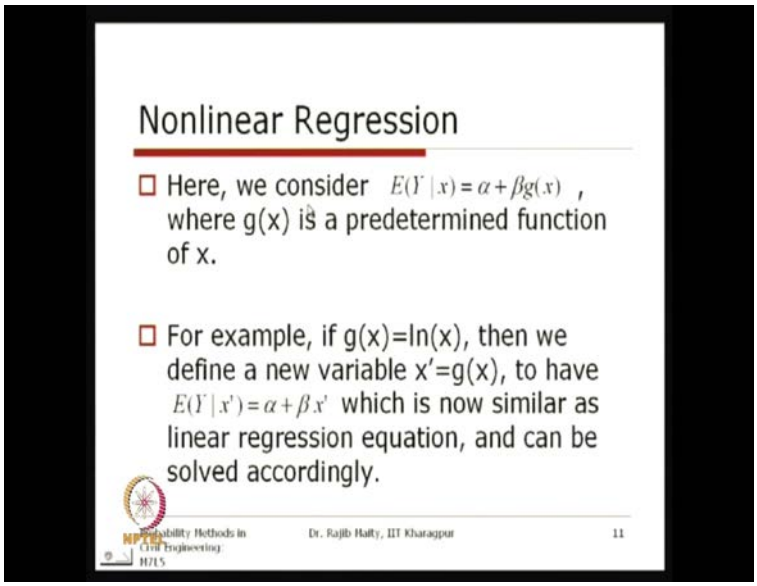
hypothesis testing we can see that how significant that estimate is and the hypothesis testing was covered in the earlierlecture.

(Refer Slide Time: 32:41)



So, here we will now we will go to that non-linear regression and this non-linear regression is essential when we see that when that linear in case of this simple linear regression. Wegenerally get one straight line and for thisif there are two inputs we get a straight surface. But, if that straight line or the surface which is linear in nature may not express thevariability fully, what we get sometimes is that the predictions based on such linear relationship may over estimate in certain ranges of this variable or underestimate in other ranges of thisvariable of this expected result. In such cases, a non-linear relationship between the variables could be more appropriate thedetermination of such non-linear relationship on the basis of the observational data involves a non-linear regression analysis.

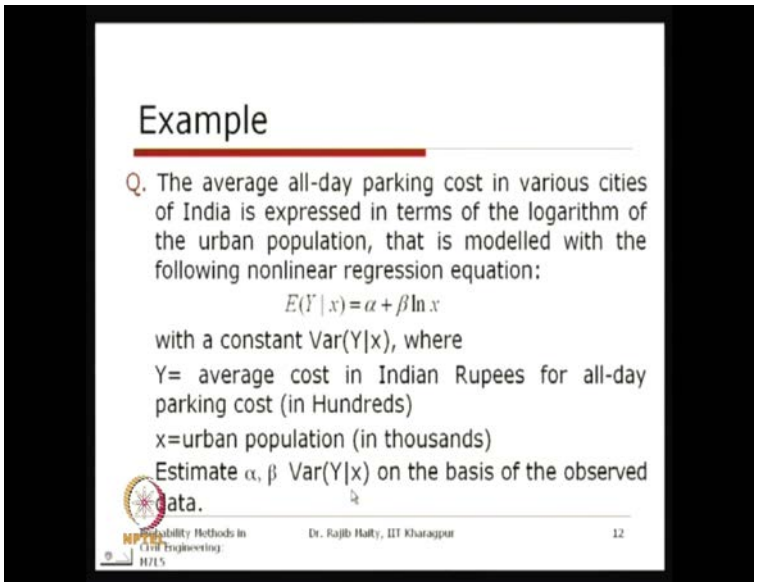
(Refer Slide Time: 33:38)



So, basically in this non-linear regression what we do is that the expectation of y on condition x is equals to a function like this alpha plus beta g x. So, this g x is a predeterminedfunction of this x. What we can do is that whatever the input x is there we will transfer;we will get a another newvariable through this g x and thatthat variable I can use with respect to this y and follow again thateither the simple linear regression ormultiple linearregression because, once we have converted it then you can see thisform of this equation is a linearregression form.

Say for example, that if that g x is equals to thatlog natural of this x then we can define a new variable x prime of g xto have thatexpectation of y given x prime. Now, it is the convertedvariable is equals to alpha plus beta x prime which is now similar as the linear regression equation and can be solved accordingly.

(Refer Slide Time: 34:43)

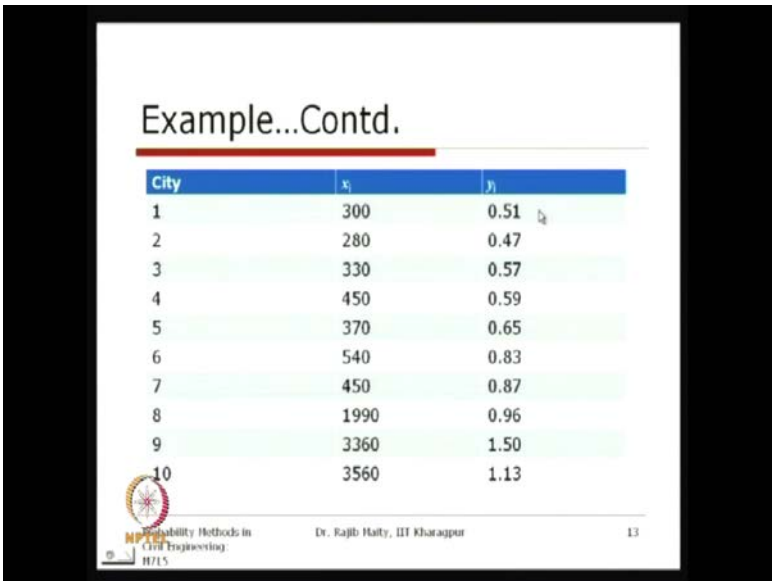


We will take up one example. Theaverage all day parking cost in various cities of India is expressed in terms of the logarithm of the urban population that ismodeled with the following non-linear regression equation.The expectation of Y given x is equals to your alpha plusbeta l n x with a constant variant Y given x where Y is the average cost in Indian rupees for all day parking cost in hundreds and x is the urban population in thousands.

Thisrelationship issometimes what happens if we just plot the data? Thatis say,for example, here the Y and x if we plot it through a scattered plot that time that is nature can be visible whether a linear or a non-linear expression should bemore appropriate or not. So, these are some initial guess. So, based on that, if we see that thislog transform of that x might be the betterestimate for this case, that is why the proposed equation is that alpha plus beta of log natural of thatof x. So, we have to estimate that alpha and beta.

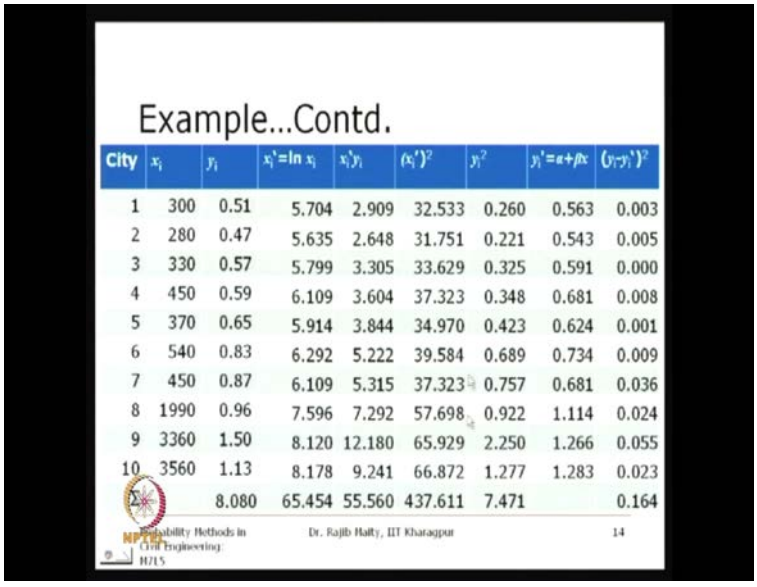
As we have told that we have to firsttransform thatvalues of this urban population through this log naturalfunction and then we will get a newset of thisnew variable, new observation in place of this x and then we can follow whatever we have seen in the linearregression equation.

(Refer Slide Time: 36:27)



So, here there are 10 cities; thisexpression that is the x i thatin thousands that is what is the population that is shown here and this is the y i in hundreds; what is the parking cost for all day parking?

(Refer Slide Time: 36:49)

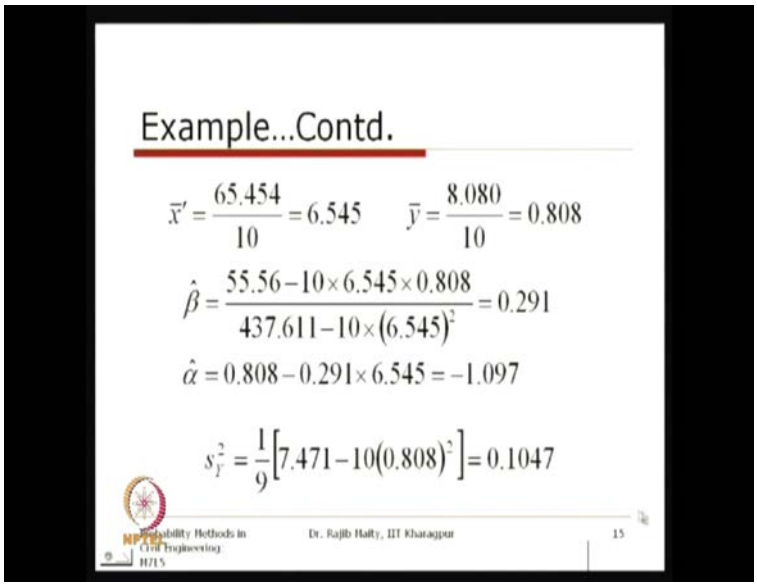


So, with these two data what we can first do is that, this x i that is the input that is converted through this log natural and we get this expression. So, using this one as theinput, we have to model this y i. So, all thosequantities that we require for this least

square estimate that isthat should be estimated only in place of this x i we should use this x i prime.

So, x i prime multiplied by y i then x i prime square then y i squarethis we canwe can calculate. So, up to this column whatever the data that is available, we can calculate and we can take theirindividual summation also. Now, with the help of this information this summation - the summation of x i prime y i summation of x i prime square summation of y i square;we can get that.

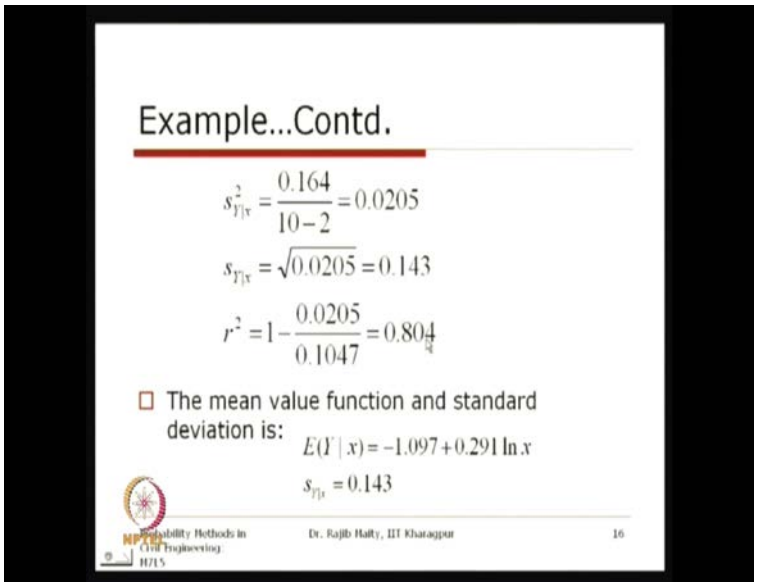
(Refer Slide Time: 37:42)



What is their mean of this xprimewhich is 6.545?ybar is 0.808. So, this estimate of this beta hat as if you refer to thisexpression of this least squareestimates then we can get that estimate of this beta hat will be 0.291 and the alpha hat will beminus point oh sorry minus 1.097 and thevariants of this y which is unconditional which is your 0.1047 mow using this alpha hat and beta hat. What we can calculate? Thatthismodeledvalues, we can get for this y i. So, this alpha and beta whatever we have estimated now and now we will use this inputs thisx i prime as this input as this x and we will get what is the model estimate of this y i.

So, that we will get and after we get this one then we can calculate what is their square errors. So, this 0.51 minus 0.563 square will give you this value similarly for all 10 observations which you have been calculated and sumthem up to get that sum of square error which is 0.164.

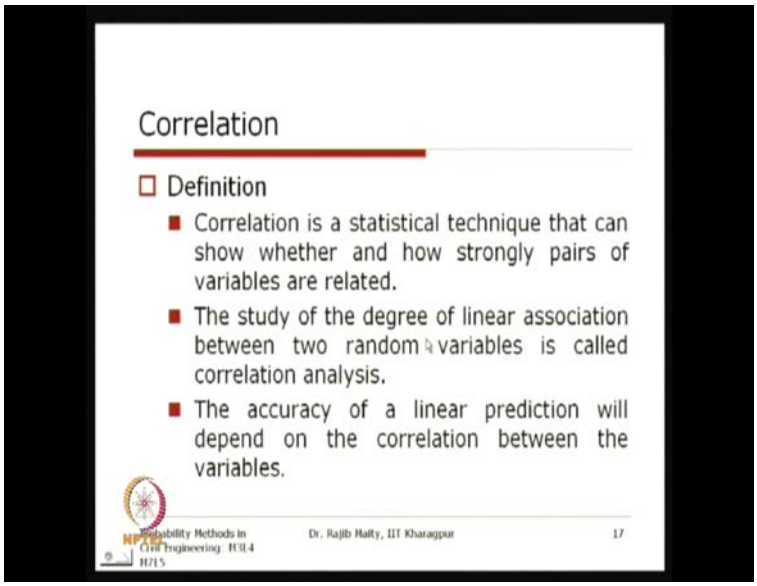
(Refer Slide Time: 39:07)



Now, using that 0.164 we know that this is conditionalvariants of that target variable Y is that sum of square error divided by n minus 2 and here you know that two means thatin number of input variable is one. So, 1 plus 1 it is 2 that isfor just now we have seen for this multiple linear regression that it is thatsum of square error divided by n minus m minus 1. So, m plus 1 degree of freedom is lost; here this 10 minus 2. So, if we calculate this one. So, these conditional variants will be 0.0205. So, this conditional standard deviationpositive square root of that which is 0.143 and that percentage of that reduction of this variants. So, this 1 minus what you got for this conditional 0.0205 and what was the unconditionalwhich is 0.1047 which is equals to 0.804.

Finally, theequation - that final equation that we get the mean valuemean value function and the standard deviation is that expectation of Y given x equals to minus 1.097 which is the estimate for thisalpha plus 0.291 estimate for the beta into log natural of that x. And, conditional variants given the x is constant, sorry, conditional standard deviationgiven that x is constantwhich is equals to 0.143.As we have seen and thisr square that you got is that 0.804. So, you can say thathere that 80.4 percent of this variability has been explainedthrough thismodel.

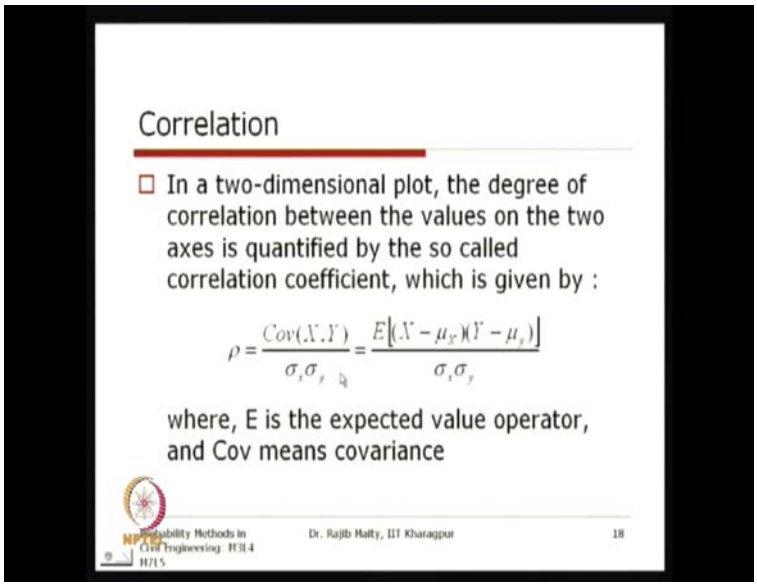
(Refer Slide Time: 40:54)



Well, now, we will just take that correlation part that we havediscussed earlier and you know that thecorrelation is a statistical technique that can show whether and how strongly the pairs of the variables are related. The study of the degree of linear association between two random variables is called this correlation analysis and the accuracy of a linear prediction will depend on the correlation between the variables.

Now, we can in this regression context, what we can say is that, if we say that what we have modeled and what we have observed is this two are linearly associated and that linear association is stronger enough then, we can say that yesthat just now we have seen that in terms of this percentage reduction of these variants in through this r square which iswe can say that. So, that muchvariability is can be explained through thatdeveloped model. So, like that this accuracy of the linear prediction will depend on thiscorrelation between thosevariables.

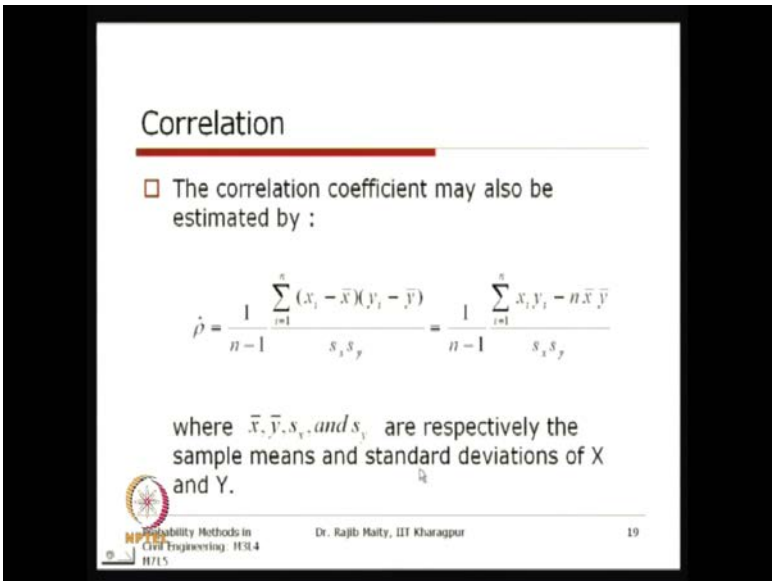
(Refer Slide Time: 42:03)



In a 2 dimensional plot now, if we just take that only oneinput here in case of the simple linear regression the degree of correlation between the values on two axis is quantified by so called correlation coefficient which is given by this equation.We have discussed earlier that is correlation coefficient is the covariance between x and y divided by standard deviation of x multiplied by standard deviation of y and the co-variants of x.Youknow that is x minus the is a expectation of x minusmu x into y minus mu y; where this e is the expected value of this operator and C o v is thatoperator, means the covariance between these two; basically the sum. In that, after we develop this regression model we have to see that what is the correlation coefficient between that what we have observed Y and what we have modeled through.

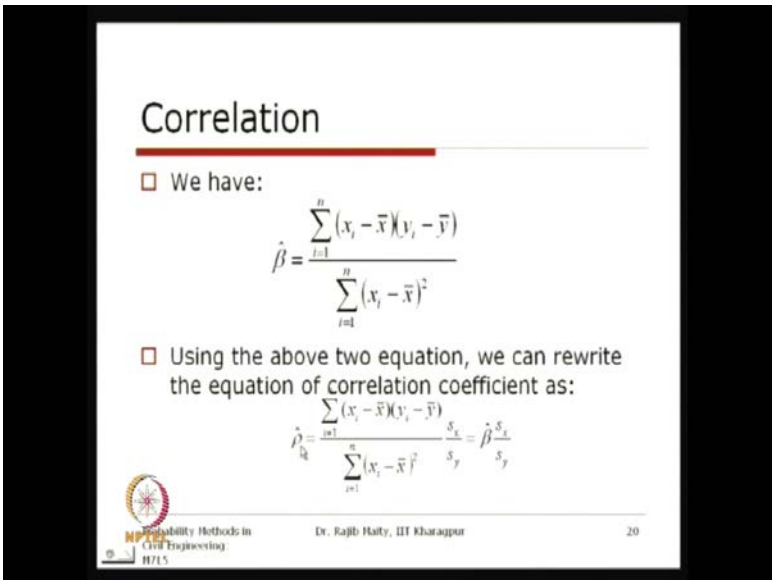
So, basically here even though we are expressing this one, just to relate our earlier discussion x and y basically, we have to see it for this y i and that y estimates that is the y hat from that regression expression.

(Refer Slide Time: 43:11)



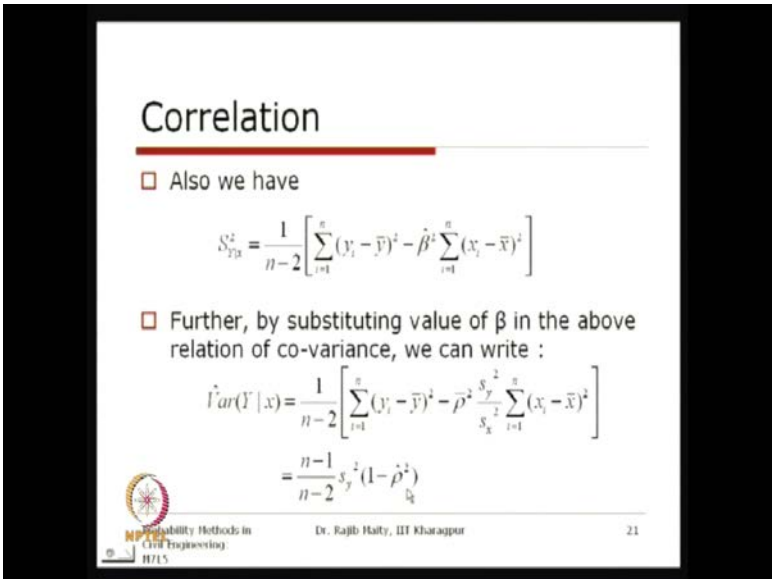
Now, this correlation coefficient may also be estimated by this rho hat; now this estimation that we are getting 1 by n minus 1 i equals to 1to n x i minus x bar y i minus y bar divided by S x S y. So, this one you can see thatthis expression can bewritten as this i equals to 1 to x I into x i y i minus n into x bar into y bar. Theseare the means of this x and y whatever we have observed in this data divided by that S xinto S y where thesethings x bar y bar Sx and S y are the sample means and standard deviation of x and y respectively.

(Refer Slide Time: 43:57)



So, we have seen also that this in the simple linear regression that the estimate of this beta is equal; is having this form of this. If you just put this one in whatever in the expression of this rho then, we can get this expression; that this rho hat is equals to that beta hat multiplied by this ratio of this Sx and S y. So, that ratio between the standard deviation of x andstandard deviation of x and standard deviation of ymultiplied by this beta hat will give you thatestimate of this correlationcoefficient.

(Refer Slide Time: 44:39)



Now, also what we have seen that the conditional variants of ygiven that x isequals to 1 by n minus 2 that multiplied by i equals to 1 to n y i minus y barsquare minusbeta hat square i equals to 1to ninto x i minus x y bar square.

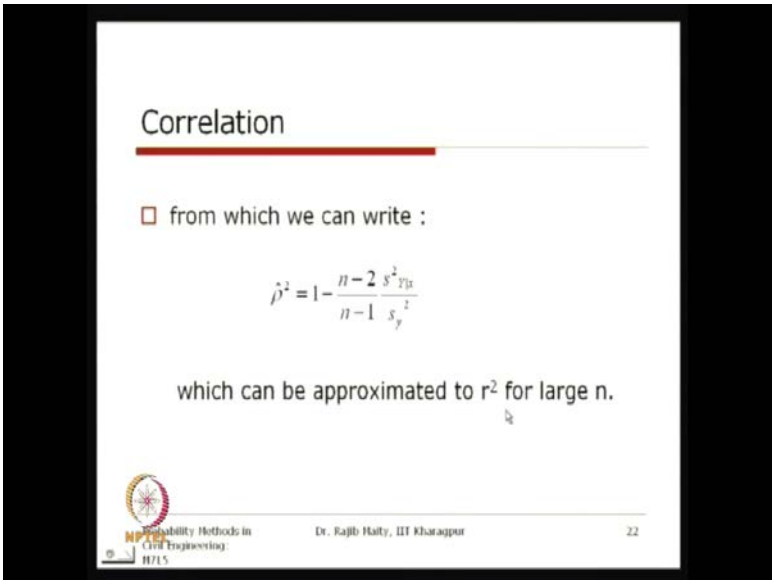
This expression that we have seen in thatsimple linearregression to express what is theconditional variant. So, if we just put thatestimate of this beta hat in terms of theircorrelation coefficient and theirvariance ofthose y and thus and x then thisconditional variants that we can write that 1 by n minus 2 this expression and in place of thisbeta hat square, we can write that thisrho hat, sorry, this will be rho hat the estimate of this correlation coefficient that rho square multiplied by this S y square and S x square.

These are the variants of y and this is the variants of x. So, if we just express thisone then we can write that this is n minus 1 divided by n minus 2 s y square into 1 minus rho hatsquare. So, this one thispart that we can take and we can express that what is coming

is that S y square will come. So, this S x and this one we canrelate to this S x square variants that you know that this is this divided by n minus 1 will give you the estimate of this S x square. So, that canbeabsorbed here and final expression that we are getting is that conditional variants of y given this x will be equals to n minus 1 by n minus 2 into s y square into 1 minusrho hat square.

Now, if wecan say that this estimate if we say that this nequals to 1, sorry, if we say that this n is very large. So, that means, thatwhen we say that there arelarge numbers of observation is available. So, for n when n is very large then we can say that thisquantity can be equated to unity and that we can say thatthis will be equals to your S y square into 1 minus rho hat square.

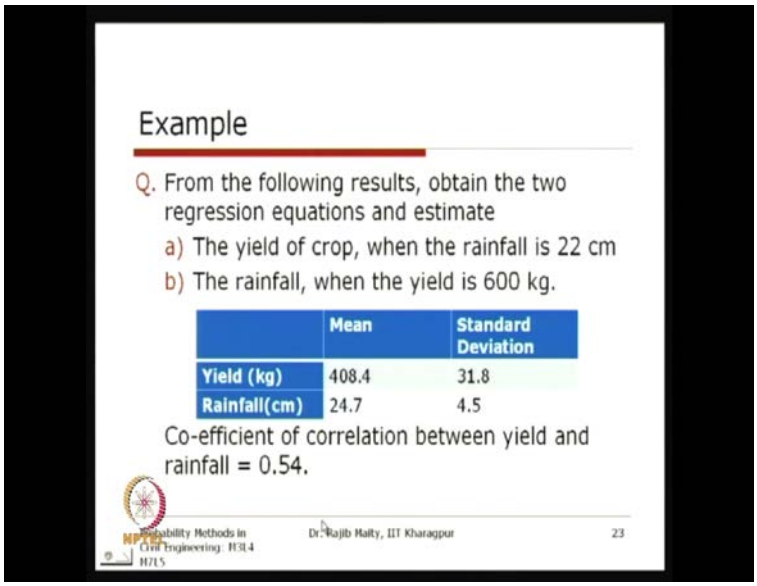
(Refer Slide Time: 47:39)



This is the final expression that we get which can be approximated. So, this one expression for this n equals to large or for any n if we justconsider this factor to be multipliedthat which can be approximated to this r square for this large n.

What we have seen earlier thatthis r square that is we have explained in terms of this percentage reduction of these variants which is equals to 1 minusthe conditional variants divided by conditional divided by unconditional variants that was ther square percentage of the reduction or that there is a percentage. Howmuch is explained through that regressionmodel. So, here we can see that for if the n is large than this quantity can be approximated to this r square.

(Refer Slide Time: 48:29)



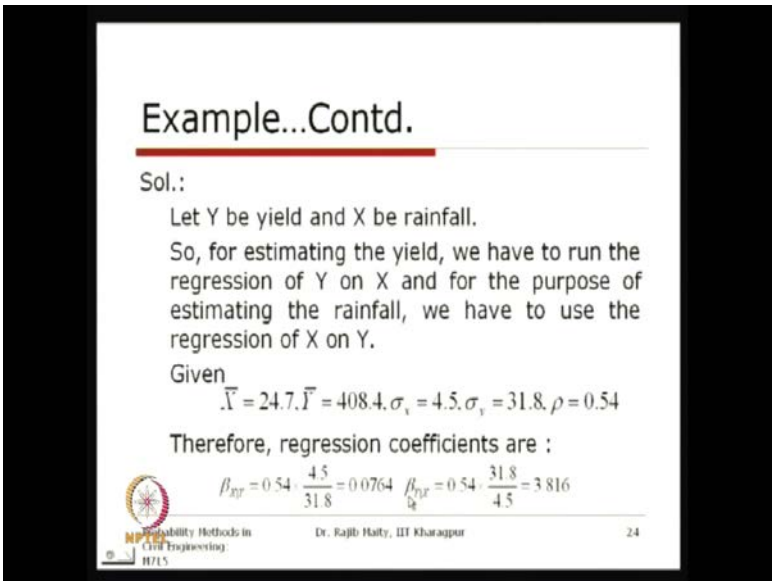
So, what we have seen so far is that, if we know thatthe correlation coefficient between the variables that we are modeling that is what is yourtarget variable and what is your input variable if you know, that what is their correlation coefficient between then and if we knowthat, what is their respective mean then, basically what we can do? Wecan develop,we can get the estimate of thoseparametersof the regression.

So, we can basically develop theirregression equation. One such example we will just see now. Thisexampleis on that; from the following results, obtain the two regression equations and estimate the yield of crop when the rainfall is 22 centimeter andthe estimate the rainfall when the yield is 600 kg.

Basically, this is the relation between this yieldof the irrigation and what is the rainfall in terms of this depth of the rainfall in centimeter and the mean of this yield? The mean yield is 408.4 k g and this mean rainfall is 24.7 centimeter.

So, the standard deviation of this yield is 31.8 k g and standard deviation of this rainfall is 4.5 centimeter and the correlation coefficientbetween the yield and rainfall is point 54. So, this informationare available to us. So, if you know this one then we have to develop the two regression equations, that is the what is theexpression for this yield given the rainfall and what is theexpression for the rainfall given by yield. So, this both this equation we can usewith the help of thistheirinformation of this correlation coefficient.

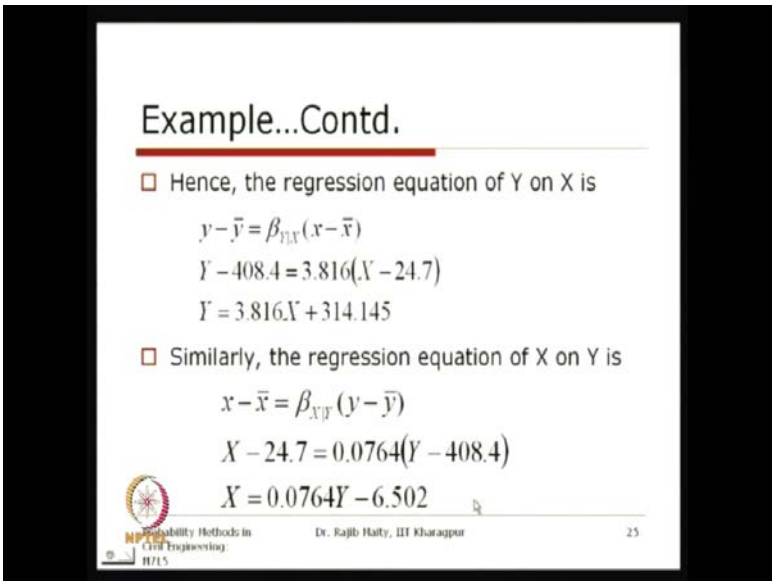
(Refer Slide Time: 50:23)



Let that Y be yield and X be the rainfall. So, for estimating the yield we have to run the regression of Y on X and for the purpose of the estimating the rainfall we have to use the regression of X on Y and the information that you know is that mean of this X mean of this Y standard deviation of X standard deviation of Y and their correlation coefficient between them.

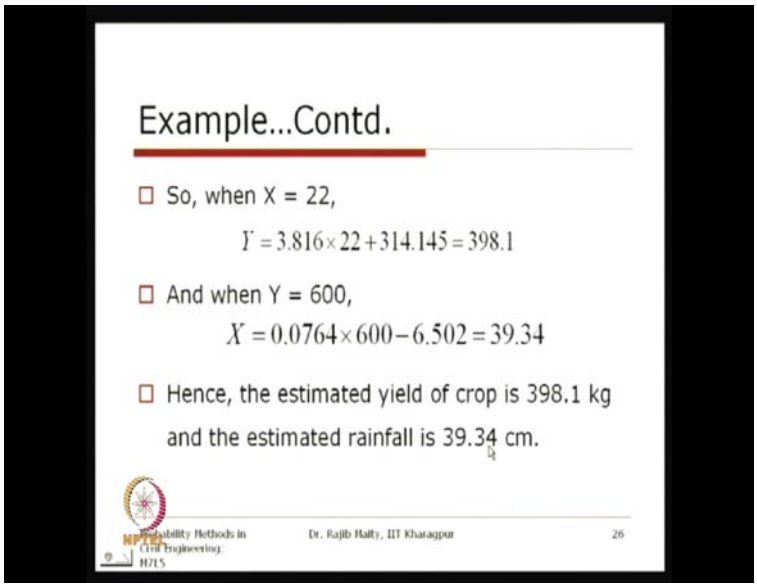
So, this beta X given Y that is when we arewhen we regressX on Y. So, this is that correlationcoefficient divided by their ratio of their standard deviation. So, what we get that 0.0764 and on the other handwhen we regress that Y on X then that beta coefficient will be 3.816.

(Refer Slide Time: 51:17)



So, theregression equation for this Y on X will be y minus y bar equals to beta y. So, the coefficient when we are regressing Y on X that is thenoted like this that beta y slash x multiplied by x minus x bar. So, that we have estimated 3.816. So, after rearranging this we can get the equation like y is equals to 3.816 x plus 3, sorry, 314.145. So, this is the regression equation for the Y on X similarly we can regress the equation of this X on Y which finally, we get that x is equals to 0.0764 y minus 6.502.

(Refer Slide Time: 52:26)



With this one if we get, if we use this expression and then the question was given that when x is equals to 22, what is the yield? Thatis when the rainfall equals to 22 centimeter what is the yield? So, putting that in this expression we get that Y is equals to your 398.1 and when thatyield is equals to 600 then the estimate of this rainfall is your 39.34. So, the estimated yield of this crop is 398.1 k g and the estimated rainfall is 39.34 centimeter. So, this is whenthis yield is when rainfall is 22 centimeter and this rainfall is estimated like this when the yield is 600 kg.

So, in thislecture or including this last lecture we havediscussedthe regressiondifferent regression techniqueincluding their simple linear regression, multiple linear regressions then non-linear regression and that in terms of this correlation. How we can estimate that? Wehave discussed. So, in this entire module what we whatever we have seen in thisprobability and style statistics. Wehavestarted with the sample statistics then we have covered thishypothesis testing. Then, how to test what the data follows? Whatdistribution through this probability paper and differenttestdifferent statistical test? To test that, what is the distribution of this of the parameter and finally, we have covered theregression analysis and correlation, thank you.