INDIAN INSTITUTE
OF
TECHNOLOGY
KHARAGPUR

NPTEL
National Programme
on
Technology Enhanced Learning

Probability Methods in Civil Engineering

Prof. Rajib Maity

Department of Civil Engineering
IIT Kharagpur

Lecture – 26

Topic

Regression Analyses and Correlation
(Contd.)

Hello and welcome to this lecture today's lecture is basically the continuation of the last lecture we are discussing in this lecture about regression analysis and correlation we have seen in the last class there are some of this basic correlation basic regression analysis we have seen it is mainly that linear regression and those things after that we are in today's class we are going to cover that remaining part of this correlation analysis. First thing that we will go is that multivariate regression.

In the last class we have seen that in the simple linear regression what you have seen there are one input variable and other one is that independent variable that is y verses x. So that if the x is your input variable and y is your target variable or your independent variable then we have seen that how we can estimate the parameters for this regression model and we have done how what are the original variance for the y and then after the regression what is the conditional variance and all.

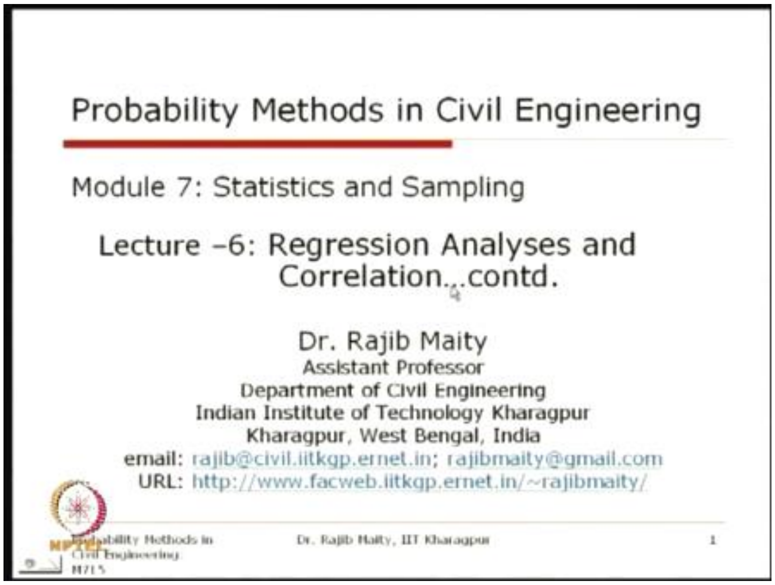
Basically we have also seen that through this regression what we are trying to do how much, what is the extent of the variants of that target variable that is, y is being reduced. So, here in this multiple regression what we do is that now this input variables are not1, is more than 1. So in that way we have to use the information of all those input variables and we have to develop a regression model for that target variable y.

So, here the inputs are same x_1, x_2, x_3 and like 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 this simple linear regression to this multiple linear regression first we will take that then from this 1 dimensional to the 2 dimensional as we are discussing in this last class that it is basically we are trying to take a straight line when it was a simple linear 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.

That you can imagine and through that 3 dimensional scatter plots of those points because, one point now will correspond to the three entries one is that from x_1 other one is x_2 and the target is y. So basically one point consists of these three a pair so that the three entry the three data point x_1, x_2 and y. So, in the 3 dimensional scatter plot basically we are trying to fit it a surface 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.

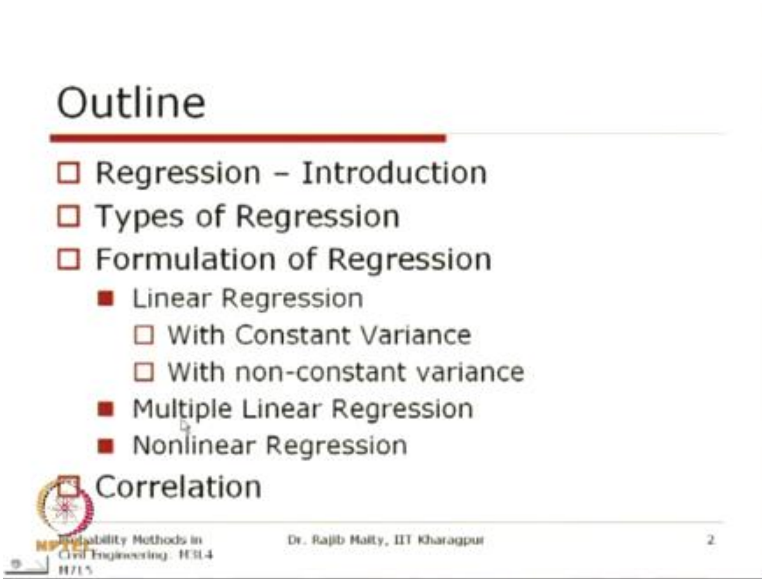
So and through for that one following the same principle of this simple linear regression we have to find out that it should be fitted in such a way that the sum of square error should be the minimum. So this is basic that transition from the simple linear regression to the multiple linear regression and based on this we will see what are the theories involved in this.

(Refer Slide Time: 03:47)



So, we are continuing with that regression analysis and correlation.

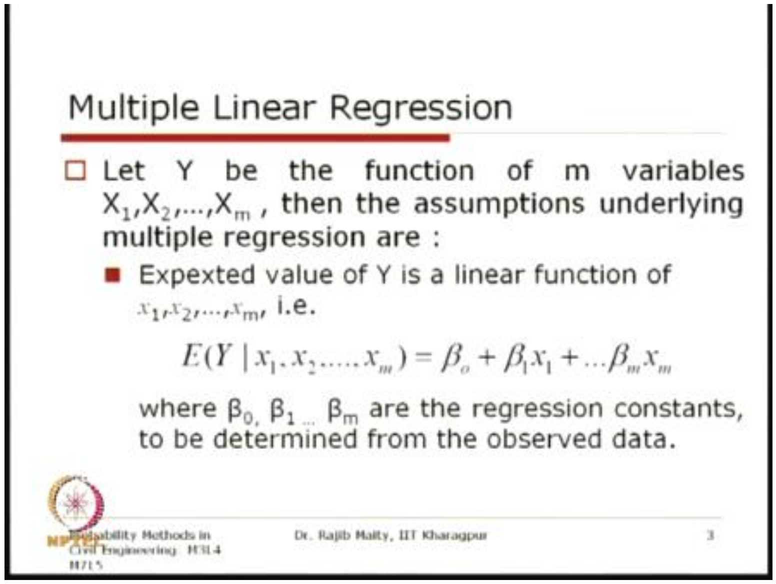
(Refer Slide Time: 03:52)



And as we have seen in this last class that in this different types of regression that we have discussed and there in the last lecture we have covered the linear regression and here we will see here today's class we will start with this multiple linear regression so and after we complete this one we will see what is this non-linear regression and this non-linear regression can also be for both the cases. It may be for the simple linear regression and also for the multiple linear regression and then we will go through this correlation and you know that this correlation we have discussed earlier also.

But, here in the context of this regression analysis we will once again see this aspect of this correlation basically through this measure we are trying to identify what is the how perfect the model that we have selected so, that we will discuss under this correlation.

(Refer Slide Time: 05:01)



Well, to start with for this multiple linear regression as I was just telling that there will be one target 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 X_1, X_2, X_m. So, let Y be the function of m variables X_1,X_2, up to X_m then the assumptions underlying the multiple regression are again following the same principle that we have discussed for the simple linear regression in the in the last lecture.

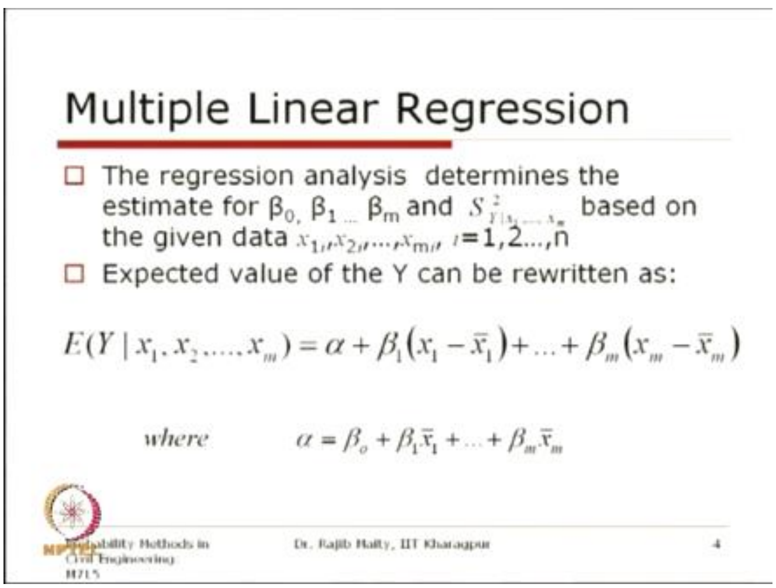
That we are trying to find out what is the expected value of this y given the input of this X_1, X_2, X_3 up to X_m. So this is the way that we are expressing that expectation of this target variable Y when the specific values of the input variables are given and this is expressed through this through this linear regression as we are now referring to this linear regression and that the linear regression is that $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ up to $+ \beta_m X_m$.

So this so if you recall, that in that last one we are having only one input and we are using this β_0 and $\beta_1 X$ and there are only two parameters for this regression was there β_0 and this one coefficient with that input variable. So, as you can see here there are m different inputs. So, this β_0, β_1, β_2, β3 up to β_m these are our regression constants and this is to be determined based on the data that is available to us and you now can see that for this X_1 if there are suppose that n numbers of that group data is available.

Then, we are having that n numbers of X_1 n numbers of X_2n 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 to we are referring to the context of these 2 input variable in a 3 dimensional case and obviously that concept can be extended to the higher dimension for example here it will be the m dimensional space that you can imagine.

So, these constants are to be determined that is the basic underlying thing in this multiple linear regression as compared to the simple linear regression where the number of input was only one.

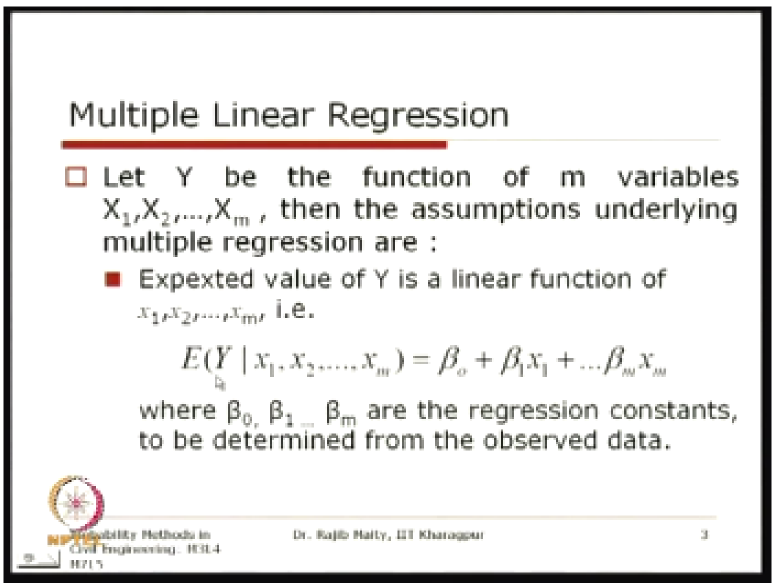
(Refer Slide Time: 07:54)



Thus, this regression analysis determines the estimate for β_0 β_1 up to β_m and that Sy^2 that is the variance of this y given this X_1, X_2 up to X_m based on the given data X_{1i}, X_{2i} up to X_{mi} and i varies from 1, 2, 3 up to n. So, this i that the substitute that is used here that is basically represents the number of data that is available to us and this 1, 2, 3 up to m represents this. So, this m represents that how many inputs that we are having now this Sy^2 given X_1, X_2, X_3 up to X_m is that the conditional variants of the target variable that is y.

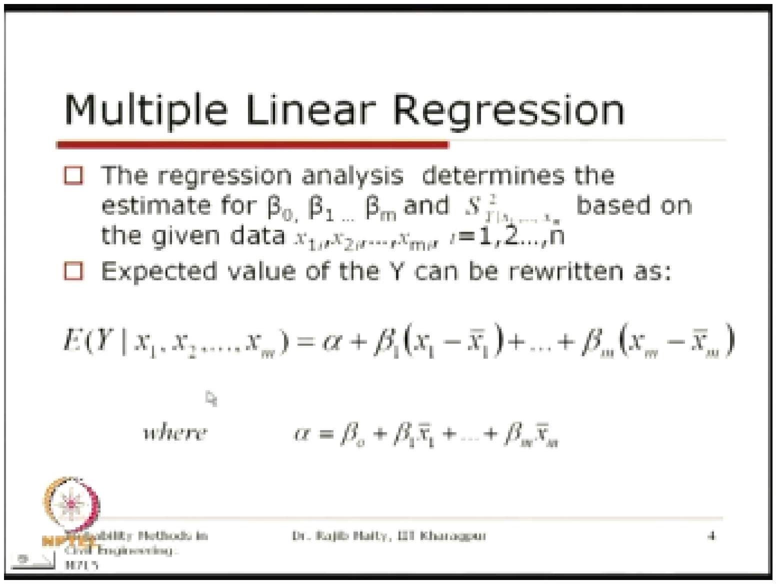
So, this conditional variants should reduce with respect to this unconditional variants which is that Sy^2 and how much is this reduction that we can relate through the relate through that. So, more the reduction is better the model and at the end of this lecture as I was discussing that at the end of this lecture we will see that how that information is related to that correlation coefficient. So after some re-arrangement of this equation.

(Refer Slide Time: 09:23)



That we have seen here that is expectation y given this X 1, X 2, X m equals to ß^{O} + ß 1 + ß 2 + ß x1 + ß xm.

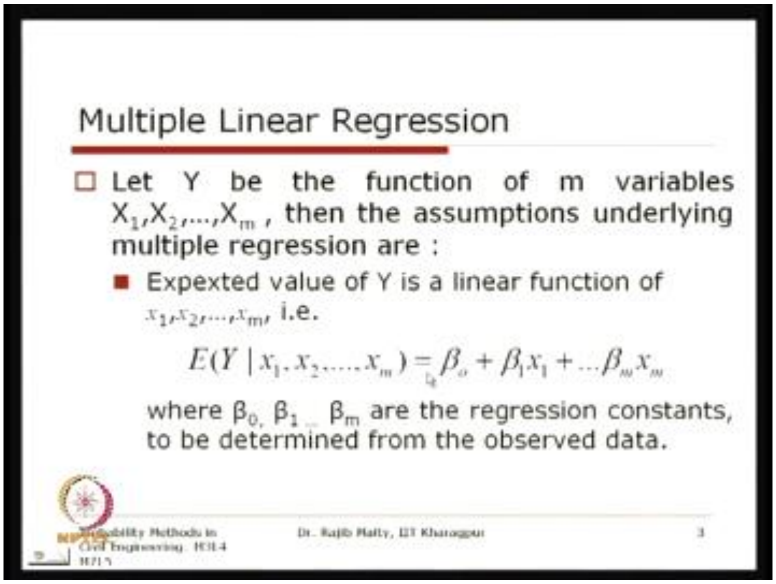
(Refer Slide Time: 09:33)



This we can rearrange to get the another form of such of the same equation which is equals to the $\alpha + \beta_1 \times X_1 - \bar{X}_1 + \beta_2 \times X_2 - \bar{X}_2$ like that up to $\beta_m \times X_m - \bar{X}_m$. So this $\bar{X}_1$, $\bar{X}_2$ or $\bar{X}_m$ is the mean of that particular input of that particular variable. So as I was telling that X_i, X_1 is that first input variable and this i can vary from 1 to m. So, we are having n data and this that mean of that input variable is represented this $\bar{X}_1$.

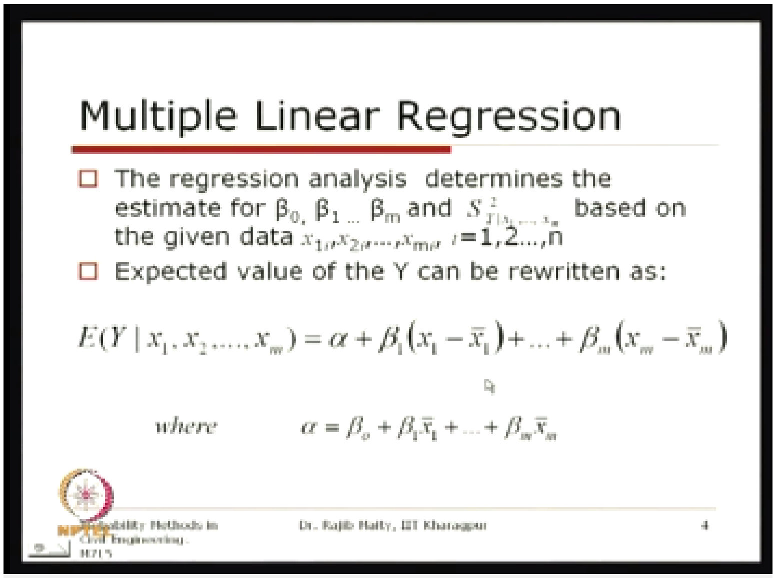
So basically this α what you can see now is basically one adjusted constant once again including that β^o and if you see from see that α can be expressed like this that $\alpha = \beta^o + \beta_1 X_1 + \beta_2 \bar{X}_2$ up to $\beta_m \bar{X}_m$. So this once we know this data basically this $\bar{X}_1$ $\bar{X}_2$ $\bar{X}_m$ are known and so once we get the estimate of this β_1 β_2 $\beta_3 \ldots \beta_m$ and α with the help of that we can estimate that β. So here in this expression what we can target is that we will first estimate these parameters α β_1, β_2, β_3, β_m. And with the help of this we will get what is the β^o.

(Refer Slide Time: 11:07)



And we will get the final form of this regression equation like this.

(Refer Slide Time: 11:12)

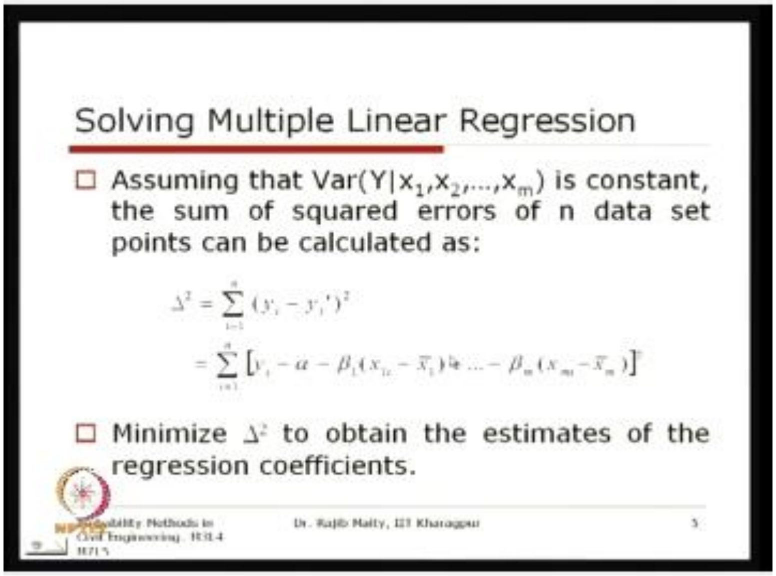


So now again we will use that same principle that we use there in the simple linear regression they are basically as I was telling we are fitting a straight line and we are trying to minimize that error and that error means what is the error with respect to the modeled target variable and what is the observe, so here also what we will do we will estimate these parameters in such a way so that estimate of that of that variable Y and the observed Y their difference should be minimum.

So this difference now what is actually observed and what is estimated from this model 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 the sum of square error and with respect to that we will take the partial derivative with respect to all this constant that we are suppose to determine and that if we equate to 0.

In the sense that we are minimizing that error sum of square error and we will get some simultaneous equation and we can solve this thing that we will see now.

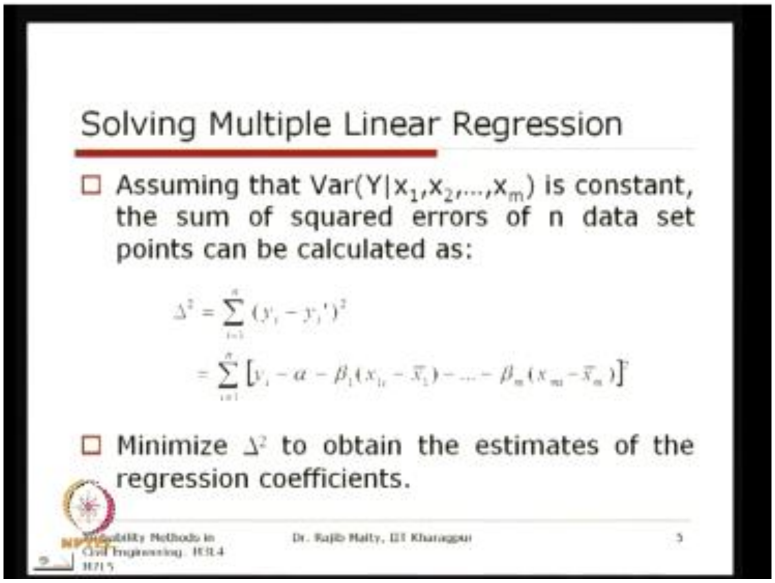
(Refer Slide Time: 12:35)



So now here again if we recall that again from this simple linear regression there are two cases one is that the variants of Y with respect to that input variable whether that is constant or that can also vary. So that in the in case of this simple linear regression we have shown one case that where the variants can even vary with respect to the range of the input variable x. Similarly here also with respect to which zone we are talking about with respect to that the input variables.

If the variants of this target variable that is Y is constant irrespective of this which zone that we are talking about the combination.

(Refer Slide Time: 13:21)

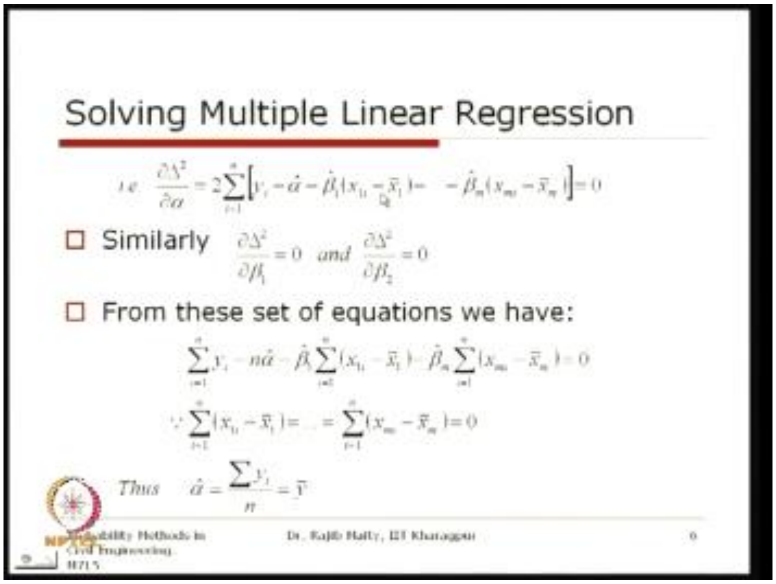


Of this input variable that is y that is X 1, X 2, X m then we can say that is either constant that variants is either constant or that can even vary. But if it varies then that form that is how it is varying over this zone so that function should be known. So now if we assume that first case that is the conditional variants is constant then the sum of square error of the end data set points can be calculated as this.

So this yi is our actually observed target variable and yi'is that modeled variable that is we are getting from this regression equation. So difference between them square them up and then sum 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 this yi' then it will come that this $\alpha - \beta_1 \times X_{1i} - \bar{X}_1 - \beta_2 \times X_{2i} - \bar{X}_2$ like this up to $\beta_m \times X_{mi} - \bar{X}_m$ and full quantity square.

So this will give you that sum of square error. Now we have to minimize this delta square to obtain the estimate for the regression coefficient.

(Refer Slide Time: 15:00)



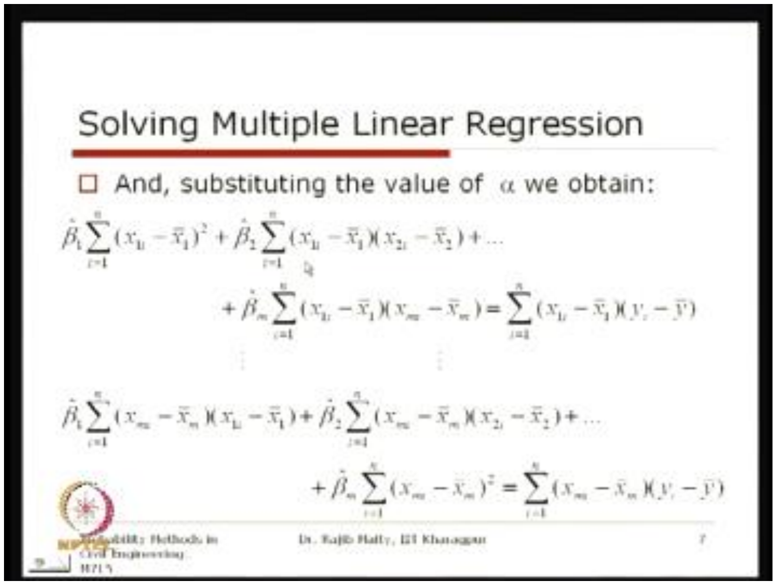
So to minimize this one as we have seen in this simple linear regression also that with respect to all these constant we have to make it equal to 0, similarly that for this set delta this delta square ∂ α = 0 similarly for this ß1, ß2, ß3 ….. ßm like that all these partial derivatives should be equals to 0.So if you take this first one that $\partial\delta^2$ $\partial\alpha$ and then we will we can take this one this 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 1i - $\bar{X}$1. Basically if we take this difference and sum them up obviously here power is1. So if we take that sum them up this will become 0.So like that for this x 2, x 3…xm all these quantities will become 0 as it is written here. So for all these quantities it will become 0. So now if we just put this expression these values here.

Then it will reduce to this form like this that α cap is equals to summation of yi obviously i from 1 to n/n. So this is basically the mean of that observed target variable Y. So this estimate of this α from this way we can see that it is the mean of Y, now the remaining there are remaining n equations are there which are the partial derivative with respect to all nß basically this ß1, ß 2….. ß m.

And there also we will get that α and this one this expression the estimate of this α if we just put it back.

(Refer Slide Time: 16:49)



So we can get a set of equations like this that this ß1 estimate of this ß1 this hat means here the estimate of this (ß1 x X1i - $\bar{X}$1)^2 + ß2 bar this x1 x1i - $\bar{X}$1into Xy - $\bar{X}$2 like that all these plus and up to this ß m which is equals to that particular variable x1i - $\bar{X}$1 x Yi – Y1 bar. So this summation for these all in observation and so this is the first equation like that we can get all other equations and up to that m^{th} equation will be like this that ß1 cap into xmi - $\bar{X}$m x Xi - $\bar{X}$I like that for this ß2.

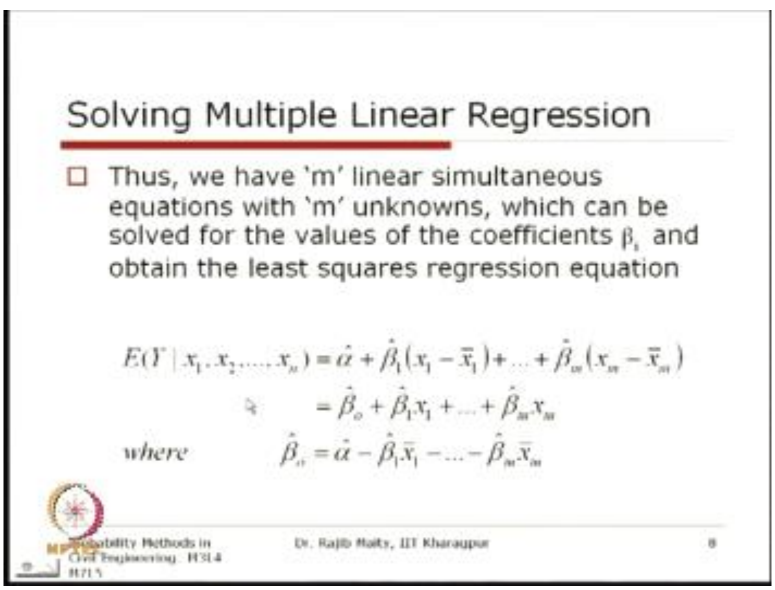
And at ßm it is (xmi - $\bar{X}$m)^2 which is equals to xmi - $\bar{X}$m x xyx, I sorry yi - y bar multiplication summation for 1 to n. So one thing that you can see here when we are taking the partial derivative with respect to ß1 that is the first equation this quantity is becoming square basically this quantity is that x1 i - x1 bar is basically getting multiplied for all these left hand side of this equation for all other variables so these things x x1 i - x 1 bar this one is there for all these entry and here also right side also you can see that this is the target this 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 β 1.

So we can say that this quantity is this like that for this last one which is the m-the equation the partial derivative taken with respect to the β 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 other expression in the left hand side and also on the right hand side with which is this function is related to the target variable which is also multiplied by this X_i - x m bar so similarly you can see that for other variables also so if it is for this β 2 then the first quantity will be that X_{1i} - X_1 bar multiplied by X_{2i} - X_2 bar and the second quantity will be $X2_i$ - X_2 bar whole square.

And like that right hand side will be X_{2i} - X_2 bar multiplied by y_i - y bar so in this way these are the simultaneous m equations are there and there are m unknowns.

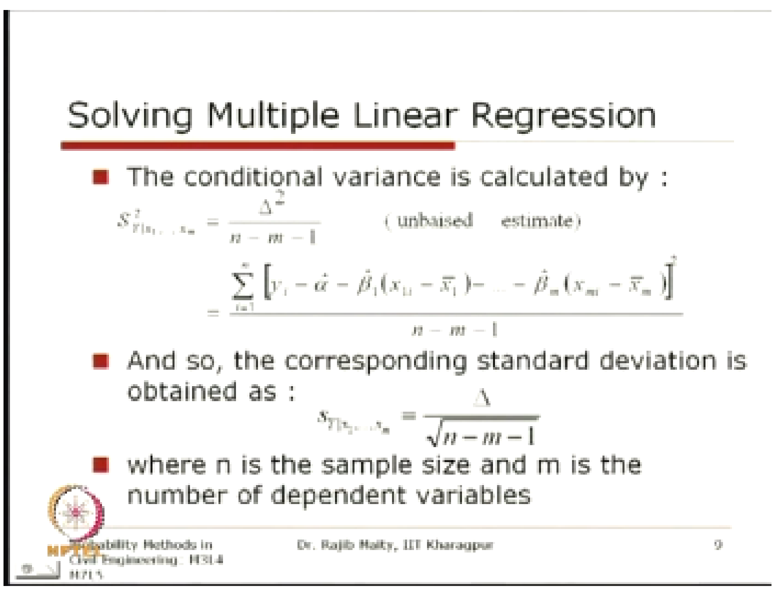
(Refer Slide Time: 19:44)



So thus we have the m linear simultaneous equation with m unknowns which can be solved for the values of the coefficient β_i β_1 β_2 up to βm and obtain the least squares regression equation. once we get this estimate finally the expression that linear regression expression that we are getting is that α cap + β1 cap into X_1 - X_1 bar + up to this βm hat means that the estimate of β m into xm - $\bar{x}_m$ and now if we write in terms of this first equation that is the β_0 hat + β1 hat x 1 + β1 β2 hat x 2 + βm hat x m this $\beta_0$is now that β the β_0 hat that is the estimate of this β_0 = α hat estimate of this α - β1 hat $\bar{x}_1$ - β2 hat $\bar{x}_2$ - beta m hat $\bar{x}_m$.

So this is the final expression for this multiple linear regression that we will get the conditional variants given that all those input variables will be that Δ^2 this is the sum of square error divided by n - m - 1 which you can see that this delta square is the summation of this y_i - this α hat – β_0 cap so this is the sum of square error divided by n - m - 1 so now you recall that when there is one only one input was there in case of the simple linear regression if you recall that equation for this conditional variants was Δ^2 divided by n - 2 so there m was only 1 so one input was there so if you put m = 1 here you will get.

(Refer Slide Time: 20:57)

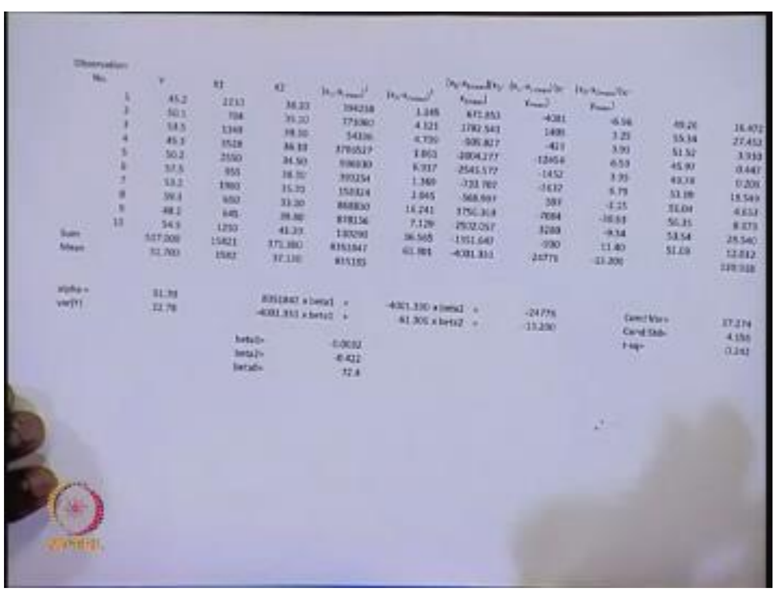


And the conditional variance given that all those input variables will that Δ2 this is sum of square error divided by n – n – 1 which you can see that this Δ^2 is the summation of this y_i – this α hat – β1 cap so this is the sum of square error divided by n – n – 1 so now we call that n there is 1 only 1 input was there in case of the symbol linear regression it will re call that equation for a conditional variance was Δ^2 divided by n- 2 so there m was only one so 1 input was it will put m = here will Δ^2 divided by n - 2 so if m increases here their m is more than 1 you have to put that value and this n is this number of data points that is available.

So this is one unbiased estimate why you we have already discuss earlier that unbiased estimate so that number of degrees of freedom is lost that should be reduced so here you see that n – m so there are m estimates are there for this X_1 bar X_2 bar up to x m bar this is the estimate that we are there are those estimates are there so m degrees of freedom is lost here again one more parameter here the α hat is there that is that we are calculating so it is that basically m + 1 degrees of freedom is lost and in case of simple linear regression.

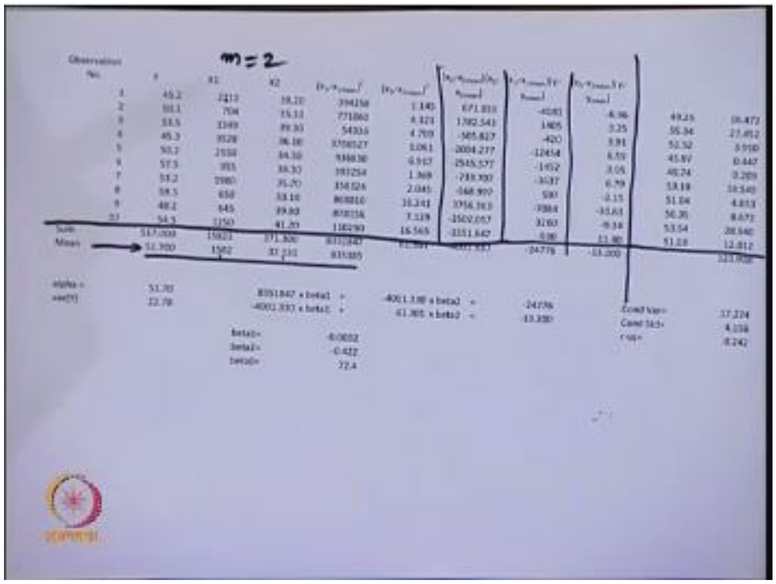
There are two constants were there that is two estimates so 2 degrees of freedom was lost there so it was Δ^2 divided by n - 2 in that simple linear regression so this is the unbiased estimate of the conditional variants of y similarly if that corresponding standard deviation is the positive square root you know this is $\Delta\sqrt{n - m - 1}$ where n is the sample size and m is the number of dependent variables so now we will take up one problem here so you can see that there are many applications can happen in the civil engineering problem.

(Refer Slide Time: 23:27)



If you just see here for this particular problem here so this could be any say for example that we are talking about what should be the temperature with respect to its altitude and latitude generally the altitude increases so we know that the temperature will decrease.

(Refer Slide Time: 23:52)



And if the latitude also increases the temperature generally decreases so suppose that type of data if we just take here so this is the number of observation that10 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 can see 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 your x_2.

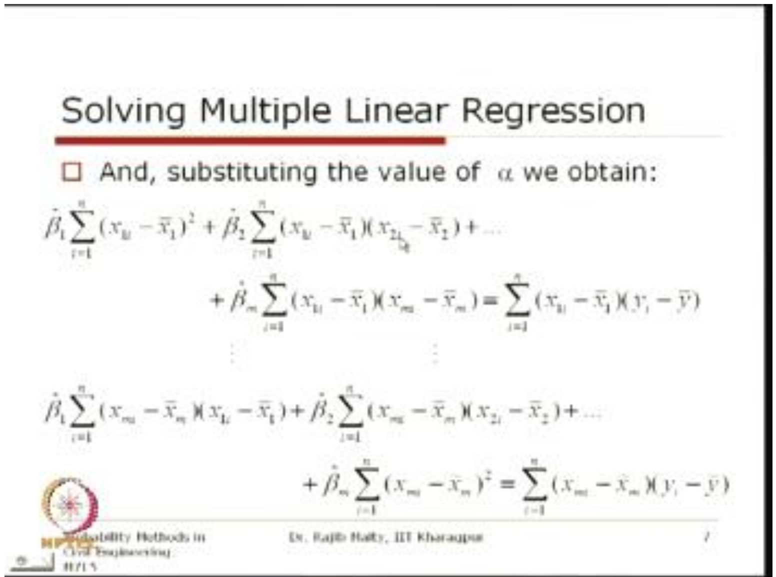
So this is up to this you can see that this is that data that we are getting now to get that estimate if you want to know that so these three columns that you can see is the input so first thing that you have to calculate is that this could be used in a general trade sheet; just to explain that how these things can happen first what we can calculate that this x_1 x_{1i} - x_1 bar that mean that we calculated here so that square up for this one so basically this x_1 - this mean is calculated here so this - this that square will be give you this value and similarly for this all such values we can calculate this one similarly this is for the x_2- x_2 bar that square that you can calculate and then this is your x1 - x_1 mean multiplied by x_2 - x_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 mean row that you can see here and this is the summation and there are ten observations are there so this 2 to 10 - this value multiplied by this 38.2 - this 37.13 this is the mean if you multiply this we will get this one similarly this column is for that x_1 - x_1 mean multiplied by y, y_i- y_i mean so if you see this one here also basically this minus this mean multiplied by this minus this mean we will get these values.

So here again this column if you see that then this y_2 - y_2 bar multiplied by y_i - y bar so this x_2 - this x_2 mean into that y - y bar that multiplication if you take we will get this value similarly for all the set en observations we have calculated and 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 α is equals to your that y mean so this is directly the 51.7 that we have seen for the mean for this Y, okay.

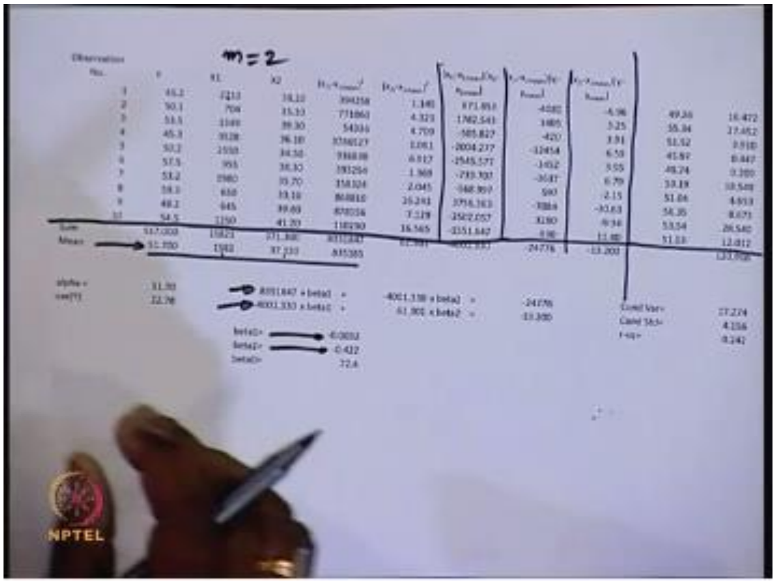
Now the variants of Y also you can calculate whatever the we have seen this that Y that we can calculate the variants of that from this sample estimate that we discuss earlier you will get this 22.78.

(Refer Slide Time: 27:44)



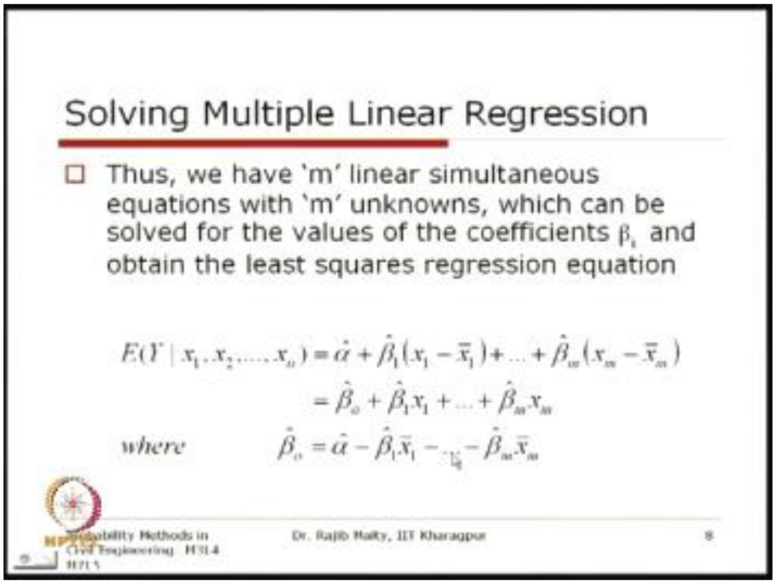
Now, this expression if you refer to this equation here that is that; just here you have to put that m=2. So, this value minus this square multiplied by this β^2 into this one this value will be equal to this one. So, all these quantities have calculated now here, so using this information, these quantities that just now what we have calculated we will just set. So, here there will be two simultaneous equations.

(Refer Slide Time: 28:11)



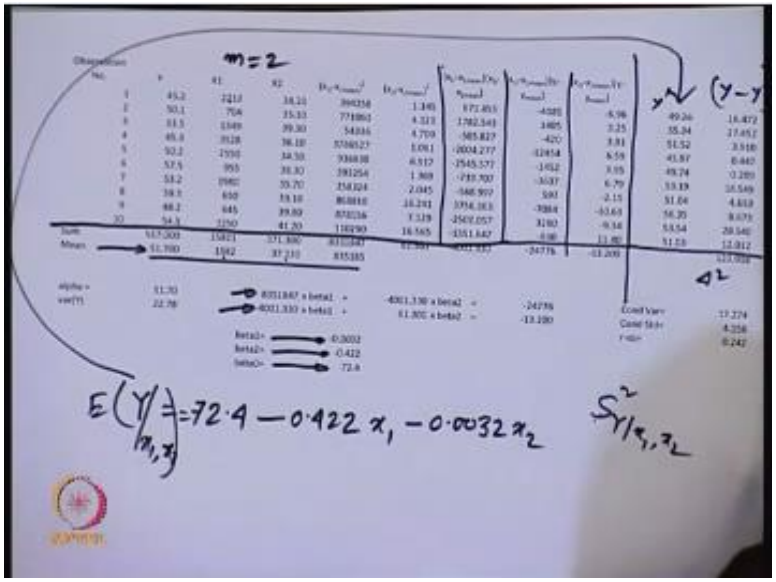
So, these two equations are written here. So, this is your that first quantity multiplied by beta1 plus this quantity multiplied by beta2 is equals to the right hand side. Similarly, so this is the first equation this is the second equation that we get and if we solve these two equations and there are two unknowns beta1 and beta2. So, you will get that this is your beta1 and this is your beta2, beta1 is -0.0032 and beta2 is -0.422. So, with this estimate and that alpha also we know so, that beta0.

(Refer Slide Time: 28:47)



We can calculate that is beta0 as you have seen here that β0 is equals to your α^-β_1^ $\bar{x}_1$ -β_2^ $\bar{x}_2$.

(Refer Slide Time: 29:00)

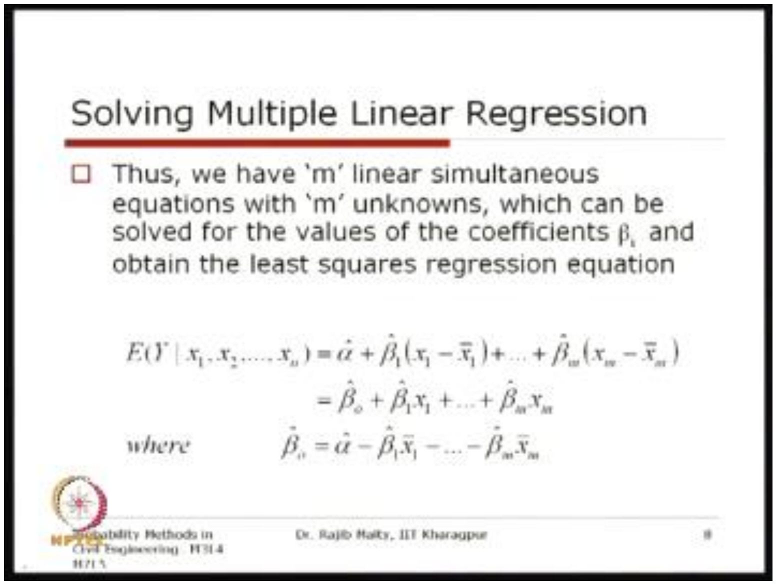


So, like if you use that one so, β_1 minus this mean of this that α-β_1 estimate multiplied by the mean of x_1-β_2 estimate into that $\overline{x}_2$ then we will get what is your β_0. So, once we get these three information that is your this is your beta0 this is beta1 and this is beta2 then we are getting this full expression like this that Y is equals to your 72.4 which is your β_0 -0.422 this is the estimate for this β_1 multiplied by x_1- 0.0032 x_2 is that expression.

Now, if we use this expression and use this what is this input then what we are getting this is the estimate of this Y basically you can say that is the expectation of Y given x_1, x_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′ and these are the error so this y-y′ so these are the error and that square. So, if I just square for all these 10 observations and sum them up this is basically is your that Δ^2 that we are getting here.

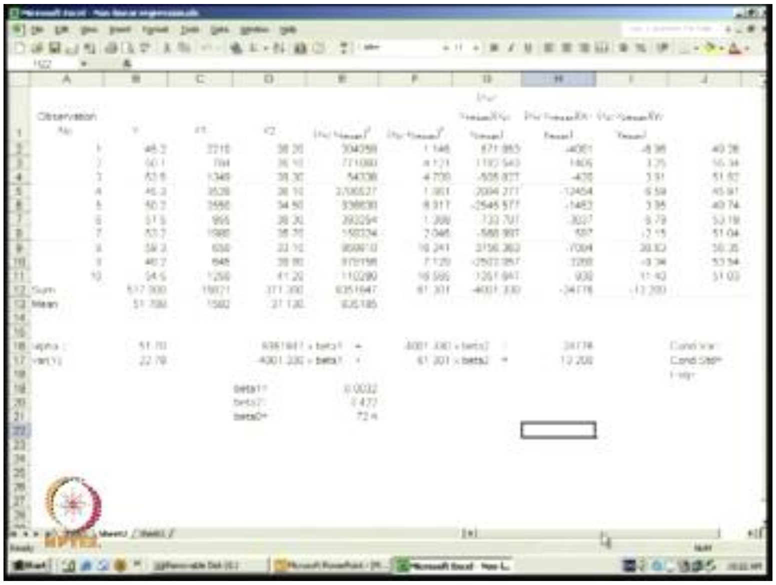
So, once we get that Δ^2 then we can calculate what is the conditional variants. This conditional variants means as we have seen that S_Y^2 are given that x_1, x_2. So, this if we calculate we will get this 17.274 and the conditional standard deviation is this positive square root of this one which is 4.156 and the r^2 value that we 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 can see here and this that r^2 is the so that 22. Sorry 24.2 percent has is the reduction.

(Refer Slide Time: 31:25)



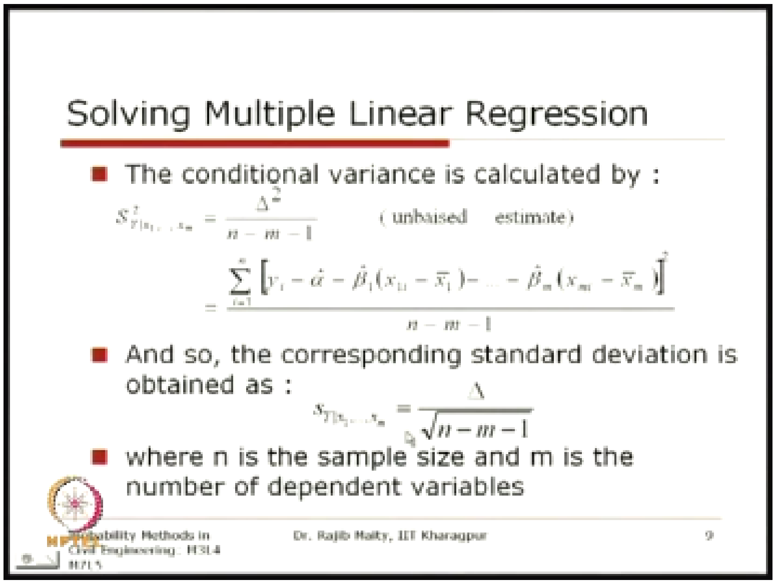
So if you want to see this one in this larger font.

(Refer Slide Time: 31:27)



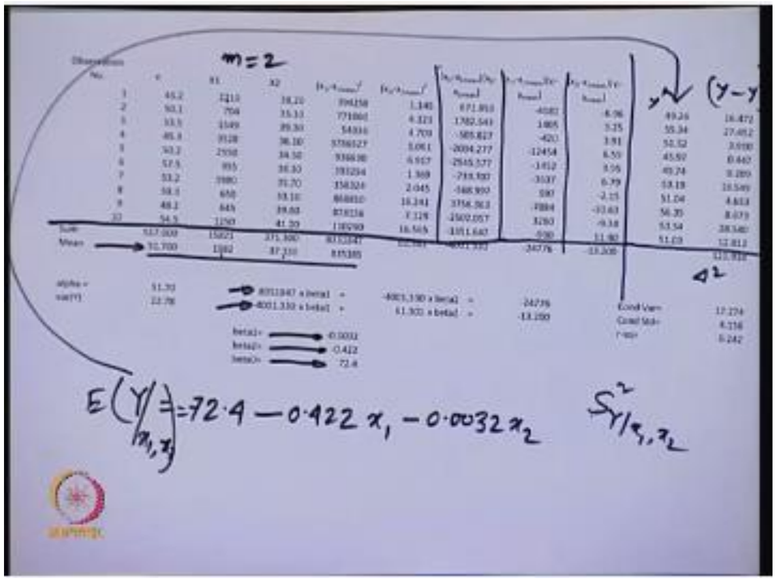
You can refer to this excel file which you can see here more clearly whatever the calculation that we have done.

(Refer Slide Time: 31:53)



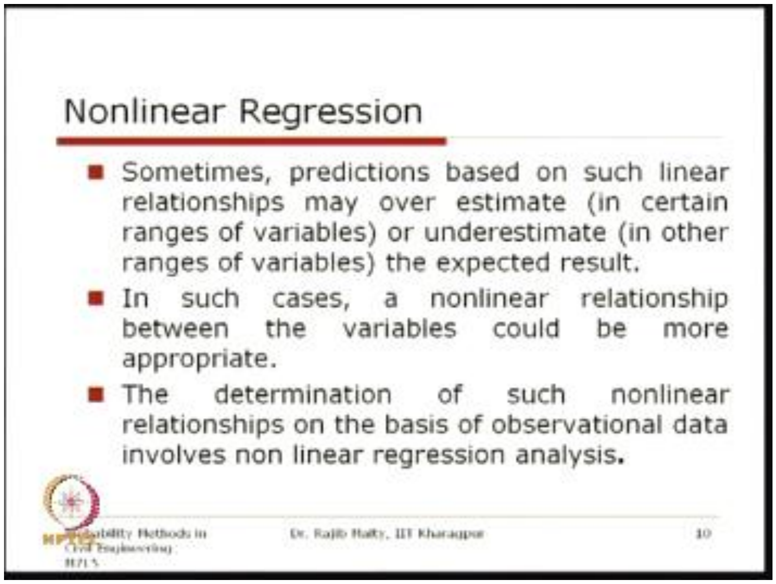
Okay, so, with this we can say here once again that.

(Refer Slide Time: 31:58)



Now when should I say that this model that what we have, whatever we have we got that is strong 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 this hypothesis testing that we have covered earlier. So, through that hypothesis testing we can see that how significant that estimate is and the hypothesis testing was covered in the earlier lecture.

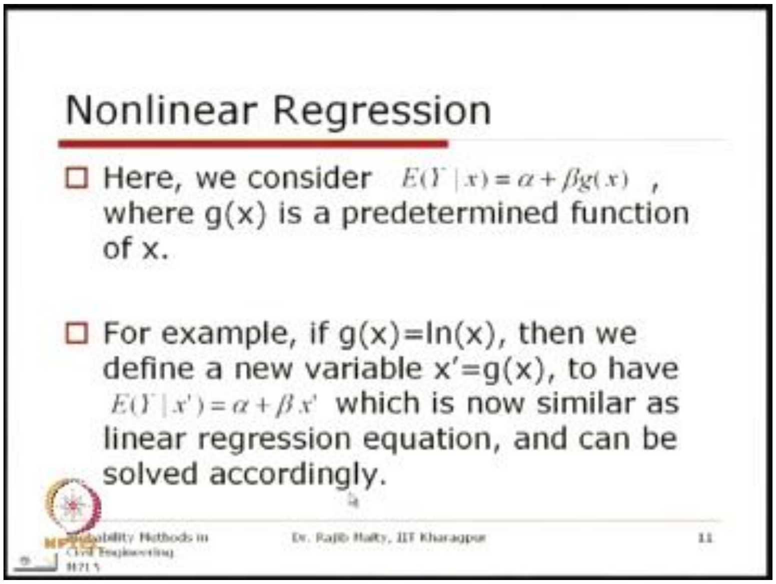
(Refer Slide Time: 32:39)



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 so in case of this simple linear regression we generally get one straight line and for this if 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 the variability fully, what we get sometimes is that the predictions based on such linear relationship may over estimate in certain ranges of this variable or under estimate in other ranges of this variable of this expected result.

So in such cases, a non-linear relationship between the variables could be more appropriate the determination of such non-linear relationship on the basis of the observational data involves a non-linear regression analysis.

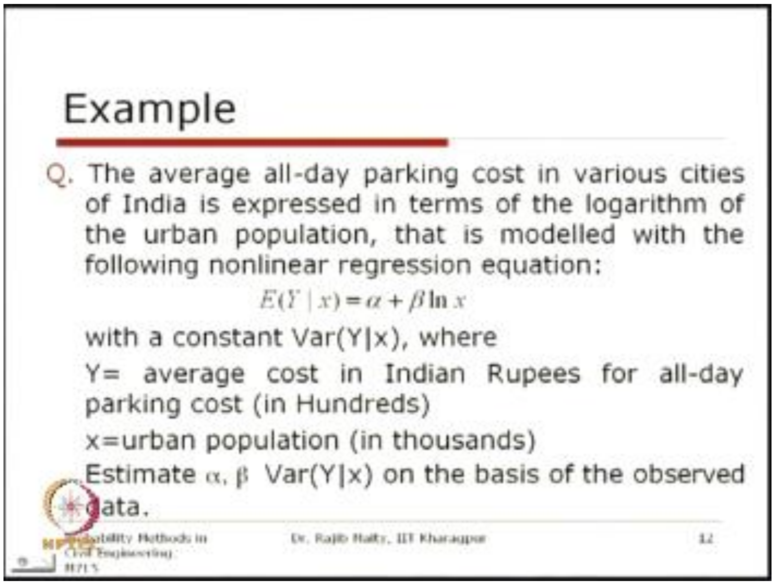
(Refer Slide Time: 33:38)



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 $\alpha+\beta g(x)$. So, this g(x) is a predetermined function of this x. What we can do is that whatever the input x is there we will transfer, we will get a another new variable through this g(x) and that variable I can use with respect to this Y and follow again that either the simple linear regression or multiple linear regression because, once we have converted it then you can see this form of this equation is a linear regression form.

Say for example, that if that g(x)=ln(x) then we can define a new variable x′ of g(x) to have that expectation of y given x′ now it is the converted variable is equals to $\alpha+\beta x'$ which is now similar as the linear regression equation and can be solved accordingly.

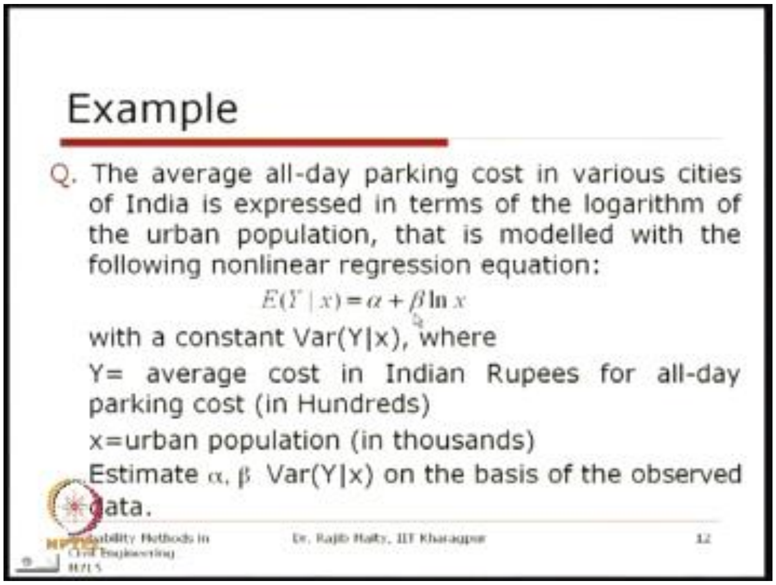
(Refer Slide Time: 34:43)



We will take up one example, the average all day parking cost in various cities of India is expressed in terms of the logarithm of the urban population that is modeled with the following non-linear regression equation. The expectation of Y given x is equals to your $\alpha+\beta \ln x$ 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.

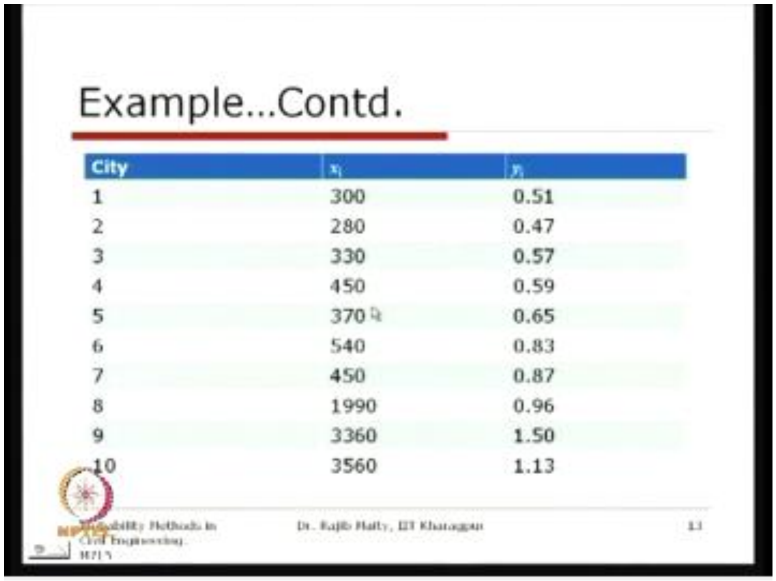
So this relationship is sometimes what happens if we just plot the data that is say for example, here the Y and x if we plot it through a scattered plot that time that is nature can be visible whether a linear or a non-linear expression should be more appropriate or not. So these are some initial guess.

(Refer Slide Time: 35:46)



So, based on that if we see that this log transform of that x might be the better estimate for this case, so that is why the proposed equation is that $\alpha+\beta\ln$ of that of x. So, we have to estimate that α and β. As we have told that we have to first transform that values of this urban population through this log natural function and then we will get a newest of this new variable, new observation in place of this x and then we can follow whatever we have seen in the linear regression equation.

(Refer Slide Time: 36:27)



So, here there are 10 cities; this expression that is the x i that in 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? So with these two data.

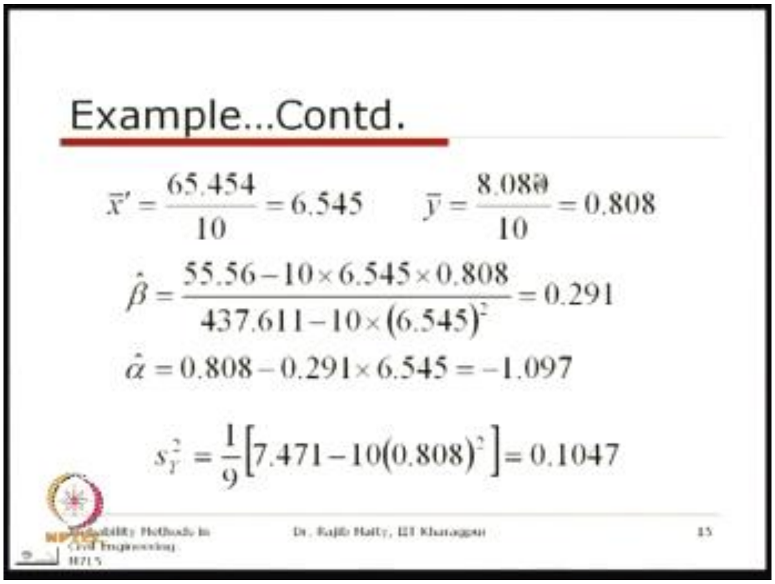
(Refer Slide Time: 36:49)

Example...Contd.

City	x_i	y_i	$x_i'=\ln x_i$	$x_i'y_i$	$(x_i')^2$	y_i^2	$y_i'=\alpha+\beta x$	$(y_i-y_i')^2$
1	300	0.51	5.704	2.909	32.533	0.260	0.563	0.003
2	280	0.47	5.635	2.648	31.751	0.221	0.543	0.005
3	330	0.57	5.799	3.305	33.629	0.325	0.591	0.000
4	450	0.59	6.109	3.604	37.323	0.348	0.681	0.008
5	370	0.65	5.914	3.844	34.970	0.423	0.624	0.001
6	540	0.83	6.292	5.222	39.584	0.689	0.734	0.009
7	450	0.87	6.109	5.315	37.323	0.757	0.681	0.036
8	1990	0.96	7.596	7.292	57.698	0.922	1.114	0.024
9	3360	1.50	8.120	12.180	65.929	2.250	1.266	0.055
10	3560	1.13	8.178	9.241	66.872	1.277	1.283	0.023
Σ		8.080	65.454	55.560	437.611	7.471		0.164

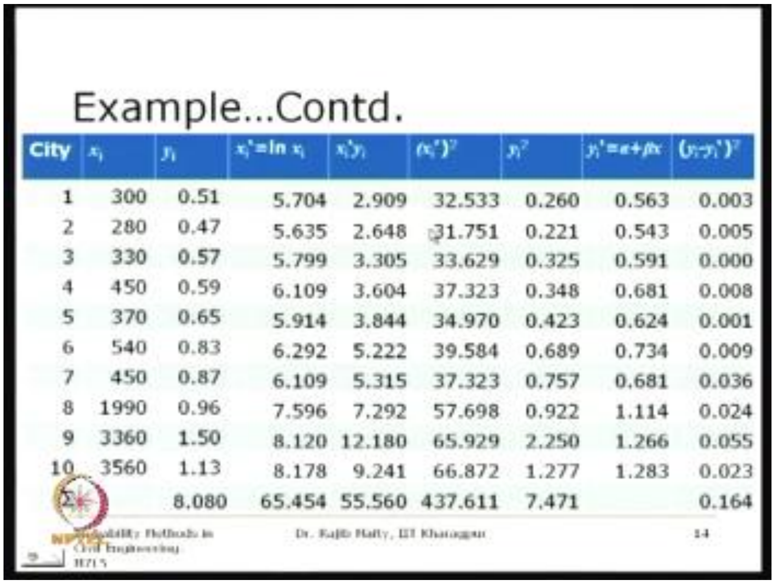
What we can first do is that, this x i that is the input that is converted through this log natural and we get this expression. So, using this one as the input, we have to model this y i. So, all those quantities that we require for this least square estimate that is that should be estimated only in 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 square this we can we can calculate. So, up to this column whatever the data that is available, we can calculate and we can take their individual ∑ also. Now, with the help of this information this ∑- the ∑ x i prime y i ∑ x i prime square ∑ of y i square.

(Refer Slide Time: 37:42)



We can get that. What is their mean of this x prime which is6.545? y bar is 0.808. So, this estimate of this β hat as if you refer to this expression of this least square estimates then we can get that estimate of this β hat will be0.291 and the α hat will be - point oh sorry - 1.097 and the variants of this y which is unconditional which is your 0.1047mow using this α hat and β hat. What we can calculate?

(Refer Slide Time: 38:29)

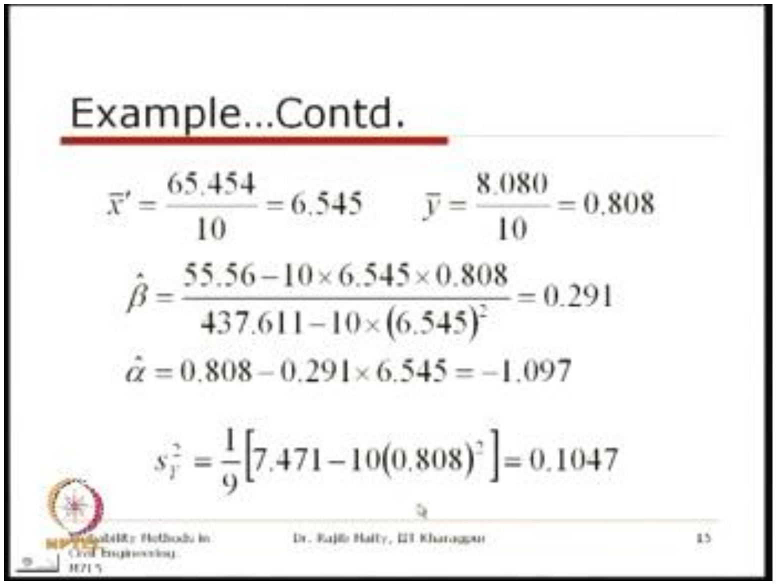


Example...Contd.

City	x_i	y_i	$x_i'=\ln x_i$	$x_i'y_i$	$(x_i')^2$	y_i^2	$y_i'=\alpha+\beta x$	$(y_i-y_i')^2$
1	300	0.51	5.704	2.909	32.533	0.260	0.563	0.003
2	280	0.47	5.635	2.648	31.751	0.221	0.543	0.005
3	330	0.57	5.799	3.305	33.629	0.325	0.591	0.000
4	450	0.59	6.109	3.604	37.323	0.348	0.681	0.008
5	370	0.65	5.914	3.844	34.970	0.423	0.624	0.001
6	540	0.83	6.292	5.222	39.584	0.689	0.734	0.009
7	450	0.87	6.109	5.315	37.323	0.757	0.681	0.036
8	1990	0.96	7.596	7.292	57.698	0.922	1.114	0.024
9	3360	1.50	8.120	12.180	65.929	2.250	1.266	0.055
10	3560	1.13	8.178	9.241	66.872	1.277	1.283	0.023
Σ		8.080	65.454	55.560	437.611	7.471		0.164

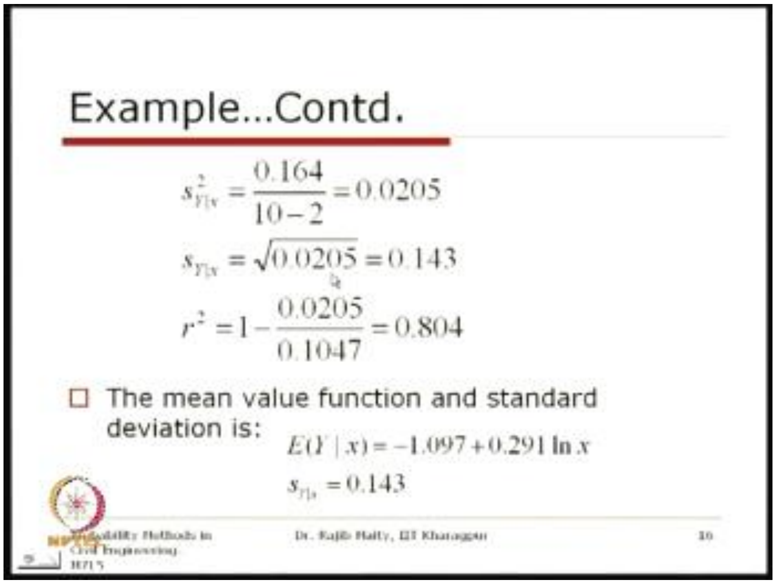
That this modeled values, we can get for this y i. So, this α and β whatever we have estimated now and now we will use this inputs this x i prime as this input as this x and we will get what is the model estimate of this y i. So, that we will get and after we get this one then we can calculate what their square errors so, this 0.51 - 0.563 square will give you this value similarly for all 10 observations which you have been calculated and sum them up to get that sum of square error which is0.164.

(Refer Slide Time: 39:05)



Now, using.

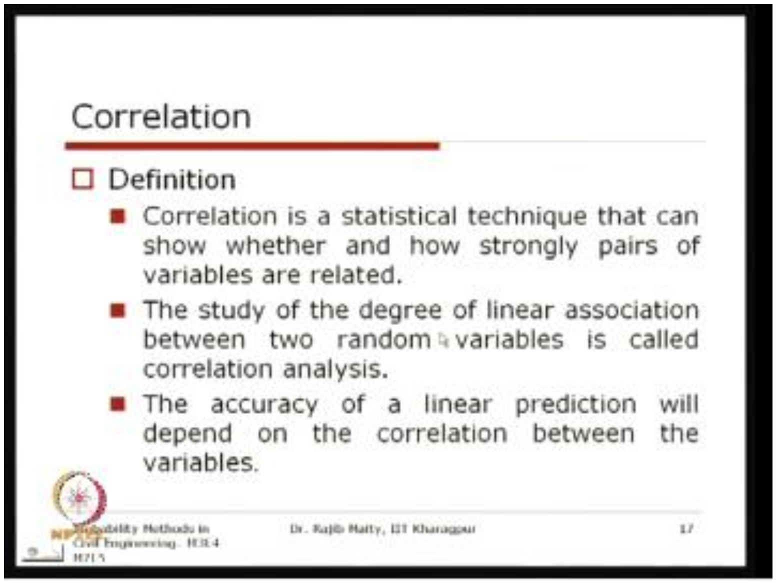
(Refer Slide Time: 39:07)



That 0.164 we know that this is conditional variants of that target variable Y is that sum of square error divided by n - 2 and here you know that two means that in number of input variable is one. So, 1 +1 it is 2 that is for just now we have seen for this multiple linear regression that it is that sum of square error divided by n - m - 1. So, m + 1 degree of freedom is lost; here this 10 - 2. So, if we calculate this one. So, these conditional variants will be 0.0205. So, this conditional standard deviation positive square root of that which is 0.143 and that percentage of that reduction of this variants.

So, this 1 - what you got for this conditional0.0205 and what was the unconditional which is 0.1047 which is equals to 0.804.Finally, the equation - that final equation that we get the mean value mean value function and the standard deviation is that expectation of Y given x equals to -1.097 which is the estimate for this α + 0.291 estimate for the β into log natural of that x. And, conditional variants given the x is constant, sorry, conditional standard deviation given that x is constant which is equals to 0.143.Aswe have seen and this r square that you got is that 0.804. So, you can say that here that80.4 percent of this variability has been explained through this model.

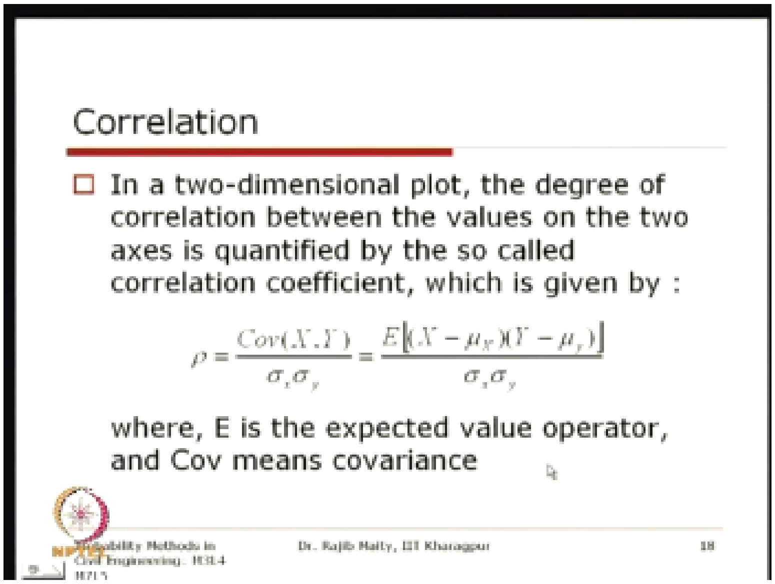
(Refer Slide Time: 40:55)



Well, now, we will just take that correlation part that we have discussed earlier and you know that the correlation 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 yes that just now we have seen that in terms of this percentage reduction of these variants in through this r square which is we can say that. So, that much variability is can be explained through that developed model. So, like that this accuracy of the linear prediction will depend on this correlation between those variables.

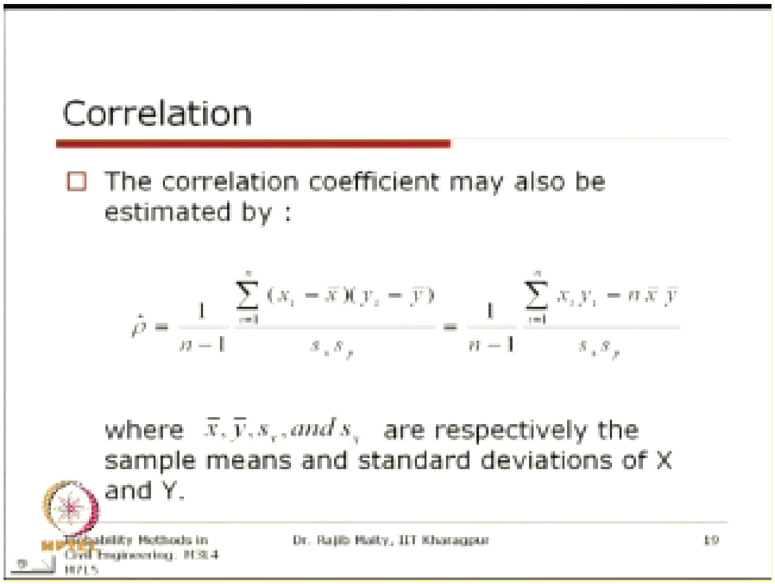
(Refer Slide Time: 42:03)



In a 2 dimensional plot now, if we just take that only one input 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. You know that is x - the is a expectation of x - x into y - μ y; where this e is the expected value of this operator and C o v is that operator, means the covariance between these two; basically the sum.

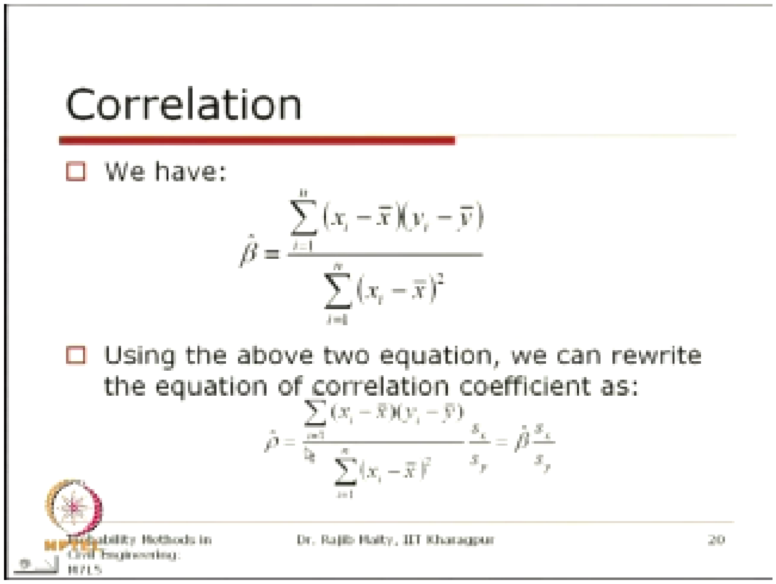
In that, after we develop this regression model we have to see that what is the correlation coefficient between that what we have observed why and what we have modeled through. So, basically here even though we are expressing this one, just to relate our earlier discussion and y basically, we have to see it for this yi and that y estimates that is the y hat from that regression expression.

(Refer Slide Time: 43:12)



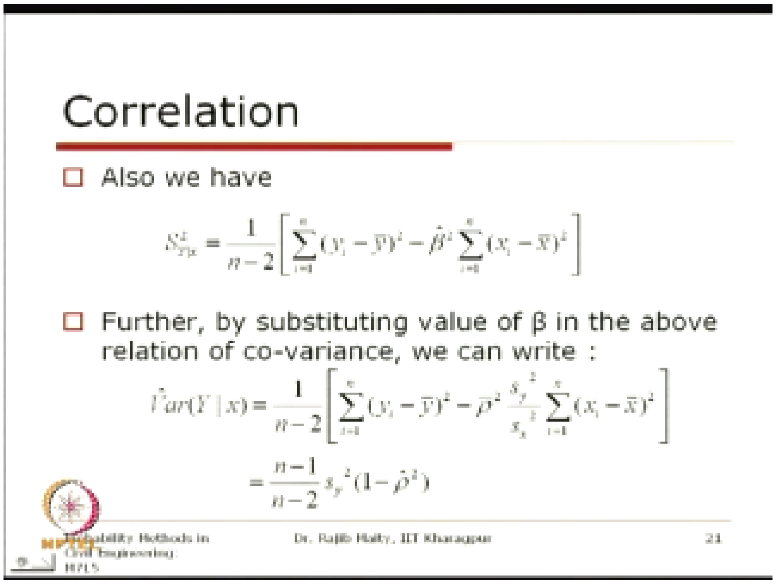
Now this correlation coefficient may also be estimated by this ρˆnow this estimation that we are getting 1 / n -1 i = 1to n x_i - x / y i minus y / divided by S $_x$ S $_y$. So, this one you can see that this expression can be written as this i = 1 to x I into x $_i$ y $_i$ - n x x / x y bar. These are the means of this x and y whatever we have observed in this data divided by that S_x x S_y where these things x̄ ȳ S_x 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 β is having this form of this. If you just put this one in whatever in the expression of this ρ then, we can get this expression; that this $\hat{\rho} = \hat{\beta}$ multiplied by this ratio of this S_x and S_y. So, that ratio between the standard deviation of x and standard deviation of x and standard deviation of y multiplied by this beta hat will give you that estimate of this correlation coefficient.

(Refer Slide Time: 44:39)

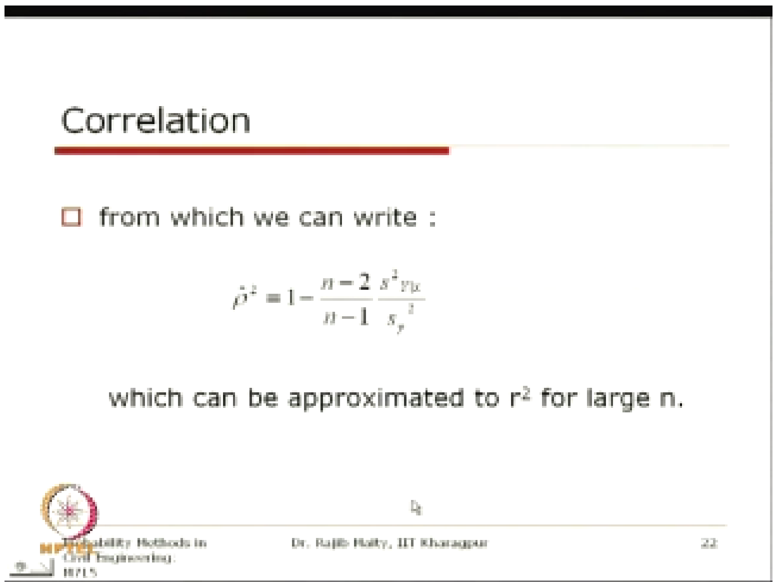


Now also what we have seen that the conditional variants of y given that x = 1 /n - 2 that multiplied by i = 1to n y $_i$ - $\bar{y}^2$ - β^2 equals to 1to n into x_i - $\bar{y}^2$. This expression that we have seen in that simple linear egression to express what is the conditional variant. So if we just put that estimate of this beta hat in terms of their correlation coefficient and their variance of those y and thus and x then this conditional variants that we can write that 1 /n - 2 this expression and in place of this β^2 we can write that this $\hat{\rho}$, sorry this will be $\hat{\rho}$, the estimate of this correlation coefficient that rho square multiplied by this Sy^2 and Sx^2.

These are the variants of y and this is the variants of x. So, if we just express this one then we can write that this is n - 1 / n - 2 s_y^2 x 1 - t $\hat{\rho}^2$. So this one this part that we can take and we can express that what is coming is that Sy^2 square will come. So, this Sx and this one we can relate 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 Sx square.

So, that can be absorbed here and final expression that we are getting is that conditional variants of y given this x will be equals to n - 1 / n -2 into sy^2 x 1 - $\hat{\rho}^2$. Now, if we can say that this estimate if we say that this n = 1, sorry, if we say that this n_i is very large. So that means, that when we say that there are large numbers of observation is available. So, for n when n is very large then we can say that this quantity can be equated to unity and that we can say that this will be equals to your Sy^2 x 1 - $\hat{\rho}^2$.

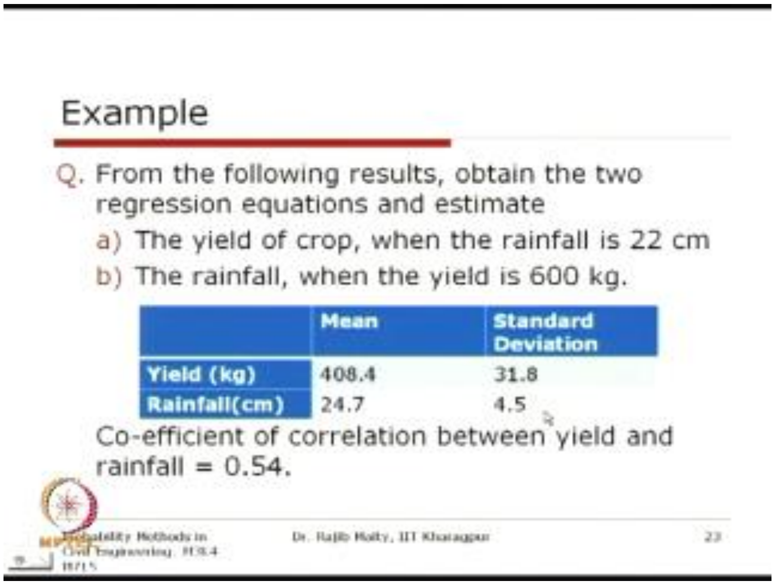
(Refer Slide Time: 47:40)



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 just consider this factor to be multiplied that which can be approximated to this r^2 for this large n. What we have seen earlier that this r^2 that this we have explained in terms of this percentage reduction of these variants which is equals to 1 minus the conditional variants divided by conditional divided by unconditional variants that was the r^2 percentage of the reduction or that there is a percentage.

How much is explained through that regression model. So, here we can see that for if the n is large than this quantity can be approximated to this r^2.

(Refer Slide Time: 48:29)

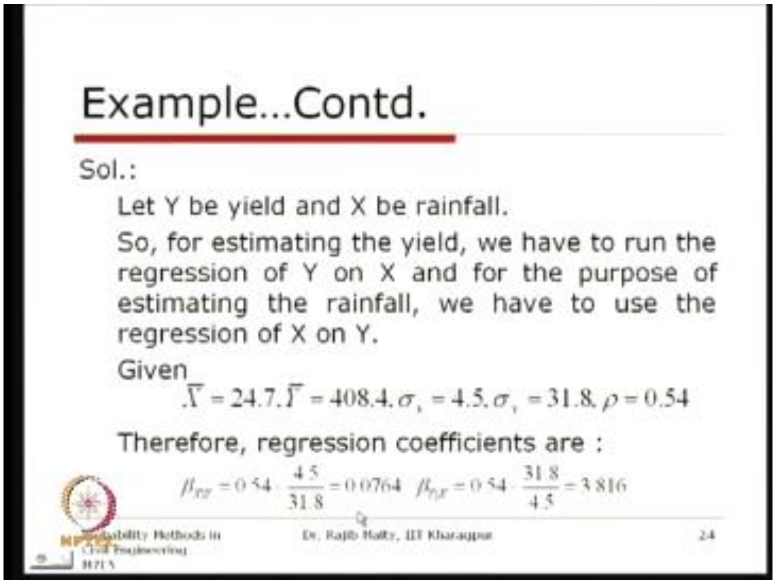


So what we have seen so far is that if we know that the correlation coefficient between the variables that we are modeling that is what is your target variable and what is your input variable if you know, that what is their correlation coefficient between then and if we know that, what is their respective mean then, basically what we can do? We can develop; we can get the estimate of those parameters of the regression. So, we can basically develop their regression equation.

One such example we will just see now. This example is on that; from the following results, obtain the two regression equations and estimate the yield of crop when the rainfall is 22 centimeter and the estimate the rainfall when the yield is 600 kg. Basically, this is the relation between this yield of 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 is31.8 k g and standard deviation of this rainfall is 4.5 centimeter and the correlation coefficient between the yield and rainfall is point 54. So, this information are available to us. So, if you know this one then we have to develop the two regression equations, that is the what is the expression for this yield given the rainfall and what is the expression for the rainfall given by yield.

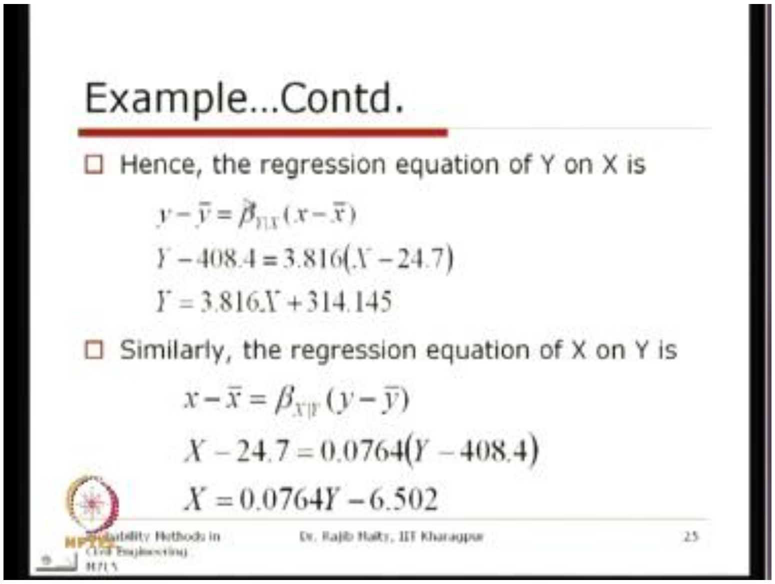
(Refer Slide Time: 50:23)



So, this both this equation we can use with the help of this their information of this correlation coefficient. 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.

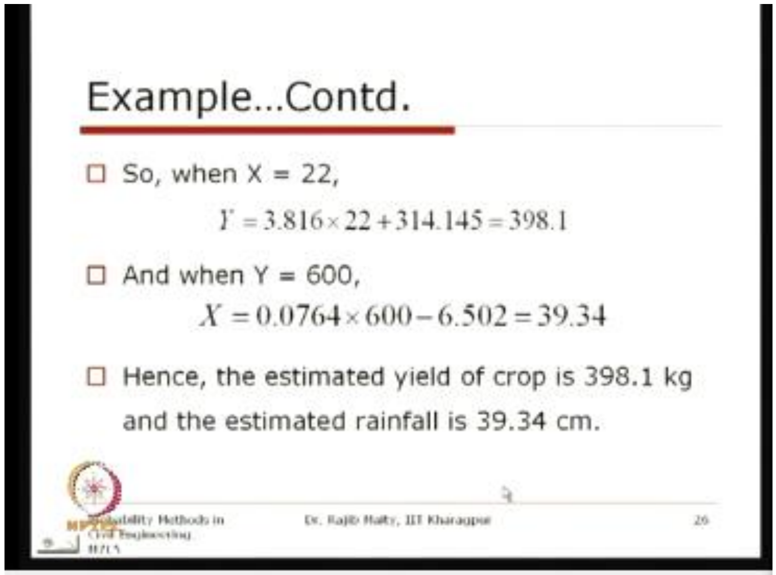
So, this β X given Y that is when we are when we regress X on Y. So, this is that correlation coefficient divided by their ratio of their standard deviation. So, what we get that 0.0764 and on the other hand when we regress that Y on X then that β coefficient will be 3.816.So, the regression equation for this Y on X will be y - y bar equals to β y. So, the coefficient when we are regressing Y on X that is the noted like this that β y slash x multiplied by x minus x bar.

(Refer Slide Time: 51:29)



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 - 6.502. So, 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? That is when the rainfall equals to 22 centimeter what is the yield?

(Refer Slide Time: 52:26)



So, putting that in this expression we get that Y is equals to your 398.1 and when that yield is equals to 600 then the estimate of this rainfall is your 39.34. So, the estimated yield of this crop is 398.1 kg and the estimated rainfall is 39.34 centimeter. So, this is when this yield is when rainfall is 22 centimeter and this rainfall is estimated like this when the yield is 600 kg. So, in this lecture or including this last lecture we have discussed the regression and different regression technique including their simple linear regression, multiple linear regressions then non-linear regression and that in terms of this correlation, how we can estimate that? We have discussed.

So, in this entire module what we whatever we have seen in this probability and style statistics. We have started with the sample statistics then we have covered this hypothesis testing. Then, how to test what the data follows what distribution through this probability paper and different test different statistical test to test that, what is the distribution of this of the parameter and finally, we have covered the regression analysis and correlation thank you.

Probability Methods in Civil Engineering

End of Lecture 40

Probability Methods in Civil Engineering –Complete

NPTEL Video Recording Team

NPTEL Web Editing Team

Technical Superintendents

Computer technicians

A IIT Kharagpur Production

www.nptel.iitm.ac.in