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

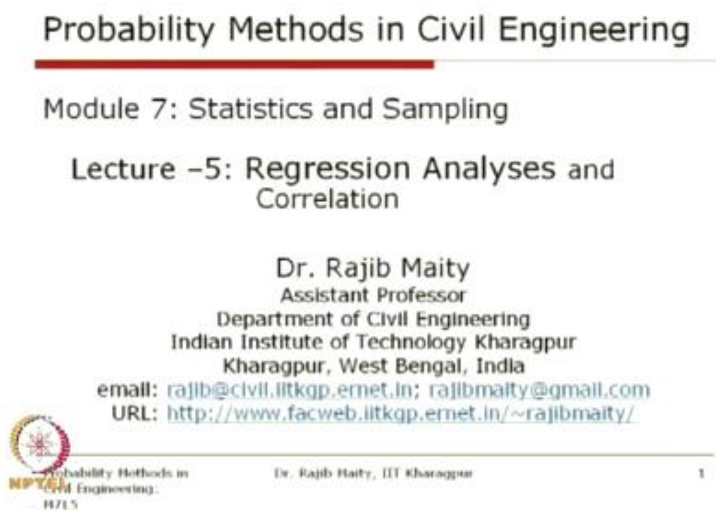
Lecture – 24

Topic

Regression Analyses and Correlation

Hello and welcome to this lecture in this lecture may be in this or the next lecture, we will cover the topic on this regression analyses.

(Refer Slide Time: 00:29)

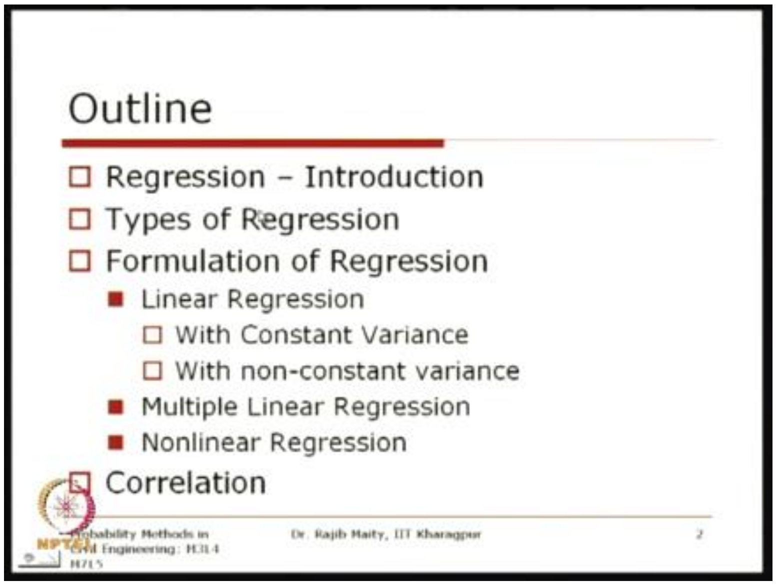


There are different types of regression analyses there say, first we will start with that simple linear regression and after that there are different types multiple regressions are there and then non-linear non linear regressions are there. So we will see and basic fundamental things basic concept we will understand basically, when in a, in many application fields obviously including civil engineering there are many random variables are there which are supposed to have some relationship to their and in this through this analyses we tried to capture that we try to model that relationship.

Now if the relationship is linear then we generally go for this linear regression and sometimes we have seen that may be the linear relationship is not sufficient. So there we have to go to the non-linear regression analyses. Sometimes the target variable is dependent only on one variable or sometimes that response variable or the target variable can depend on more than one and dependent variable. So, in that case we generally go for these multiple regressions. So, all these things we will learn in this lecture or this may continue to the next lecture also.

And so this is our today's lecture title is regression analyses and correlation and this correlation means here that we have already discussed earlier that this correlation when we discuss this random variable and all. So, here also we will see that how this regression analyses in this regression analyses correlation is an important part. So, we will just see in the light of this regression analyses also towards the end of this lecture.

(Refer Slide Time: 02:24)

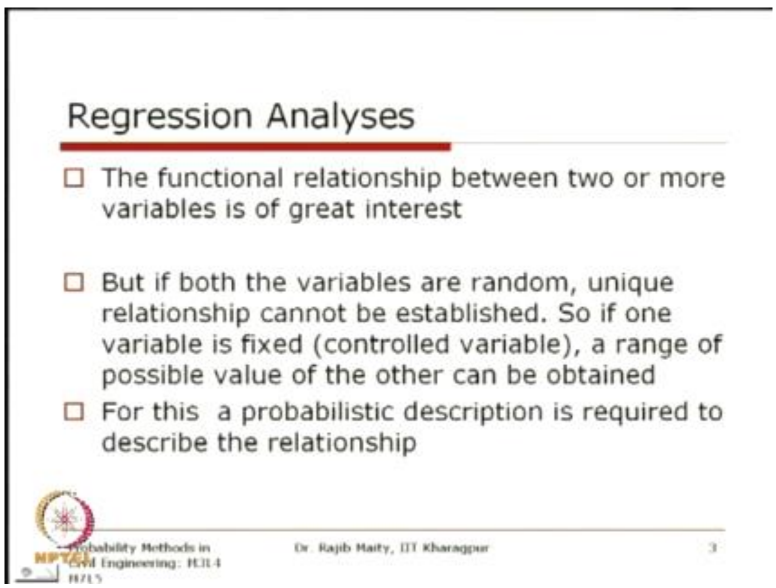


So, our outline of this today's lecture is that first we will go through some introduction and then we will discuss about some that different types of regression then you formulation of this regression in this there are linear regression as I mentioned and in this linear regression also it may have the constant variance or it may have the non constant variance. So, this non constant may be the variable variance maybe the other word but just to avoid to similar word. So, it has used as this non constant variance.

So this non constant variance and this constant variance means in general for the linear regression when we refer to we refer to this constant variance we means over the entire range of the dependent variable the variance of the response variable remains same that is what is the I can say that by default case but sometimes or it can be observed that these variance may also vary over the different range of that dependent variable.

So in that case we have to go for this non constant variance also then, if there are more than one dependent variable then we have to go for this multiple linear regression and if the relationship we see that may not be linear sometimes some other non-linear relationship may have better can better extend the target variable then we can go for this non-linear regression and then as I told that there is so, we will see that correlation basically this will be a major that how strong the relationship has been captured through that model that we have developed through this regression so that we will see.

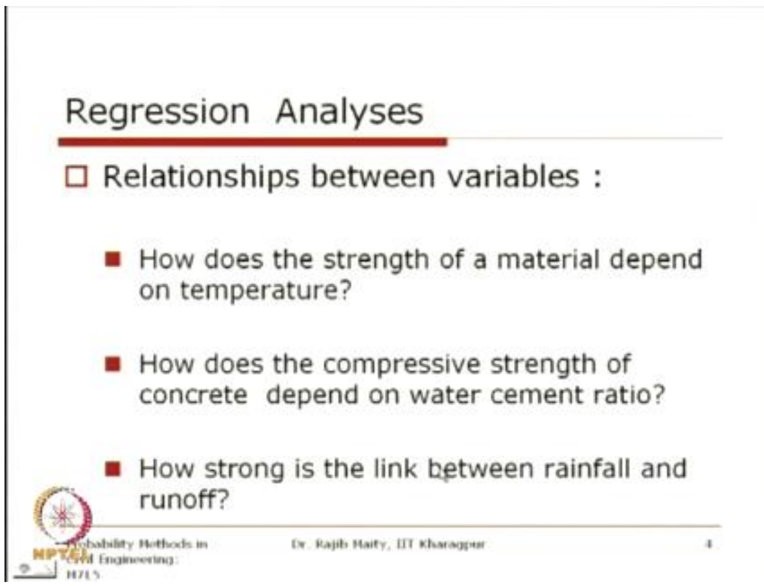
(Refer Slide Time: 04:10)



Well in this regression analyses the fundamental that to sorry the functional relationship between two or more variables is of great interest as I mentioned that there may be many variables which are which we can see that there could be a could be a relationship by the linear or non-linear and this kind of relationship basically if we just take the observed data and plotted through some scatter diagram and then itself by visual inspection itself we can see that there are whether there are some types of relationship is there or not.

So if we can see than we can we can think of this type of regression analyses to capture that particular relationship. So, here if both the variables are random, unique relationship cannot be established. So you know that unique relationship here what is meant that it may not be that one to one relationship there could be some even there could be some randomness in both the variable. So, if one variable is fixed and that is known that is termed as a control variable or that what I mentioned is that is the dependent variable the range of possible values of the other can be obtained through this analyses for this a probabilistic description is required to describe this relationship and basically this is what is you will get through this regression analyses.

(Refer Slide Time: 05:45)

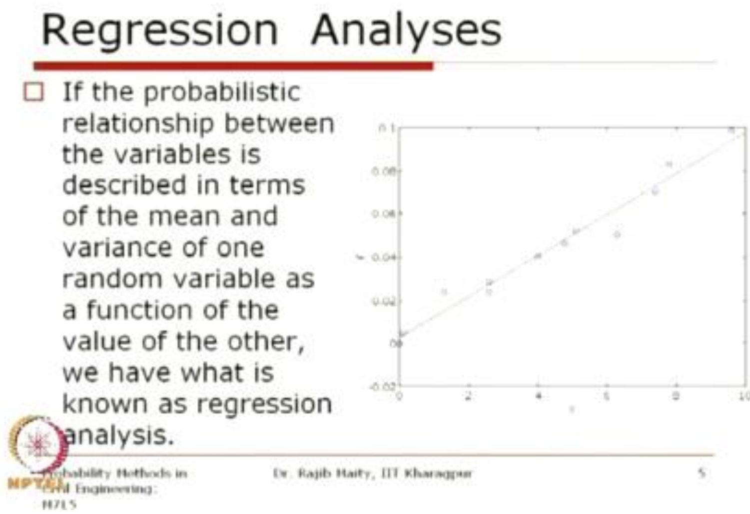


So, in this the type of question particularly if I concentrate to this different a field of application in civil engineering then this type of analyses will give me the answers to a kind of this type of question say that how does the strength of material depend on the temperature. So if the temperature I vary. So, how the strength of material whether it will increase or decrease or how the relationship is second is say that how does the compressive strength of the concrete depend on the water cement ratio.

So, if I increase the water cement ratio then what will happen to the compressive strength or if I decrease it what will happen. So these are some two variables are considered similarly, what we can say that whether that target variable here is the compressive strength may have instead of this only that water cement ratio. There could have been other factors as well that can be influencing to this. So then what will happen that one target variable and more than one dependent variable – so that multiple regression can come into the picture third so that, how strong the link between the rainfall runoff for a given catchment or for given area.

So how so rainfall and runoff if the rainfall is more runoff can be more so how strong is that relationship so this type of answer we can get through this regression analyses.

(Refer Slide Time: 07:17)

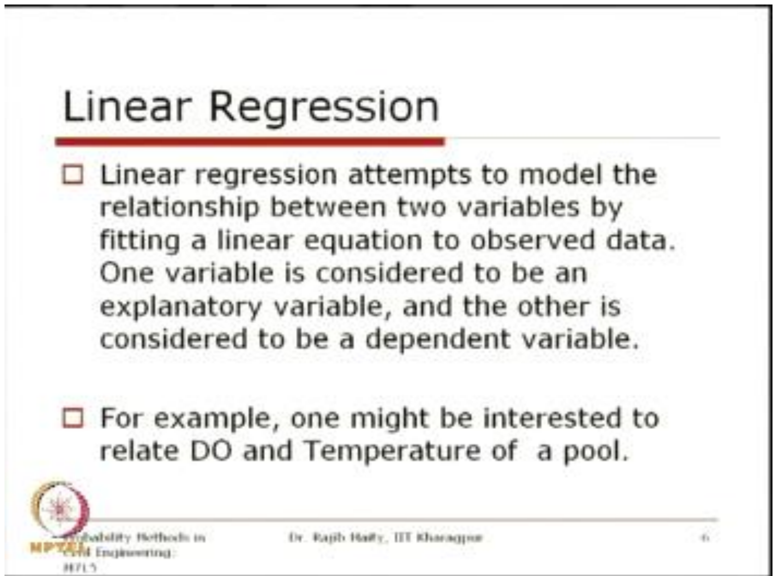


The probabilistic relationship between the variable is described in terms of the mean and variance of one random variable as a function of the value of the other we have what is known as the regression analyses. So, say for example as I was telling just by if I just plot that through a scatter plot the what is the observe data that we are having the paired observe data paired in the sense here that we are talking about the two variables first. So this is one variable is x and other one is the y.

So now, if I just plot it these blue circles you can see that this is the paired data and. So, we can see that if x increases y can y is also increasing and vice versa if x is decreasing y is decreasing.

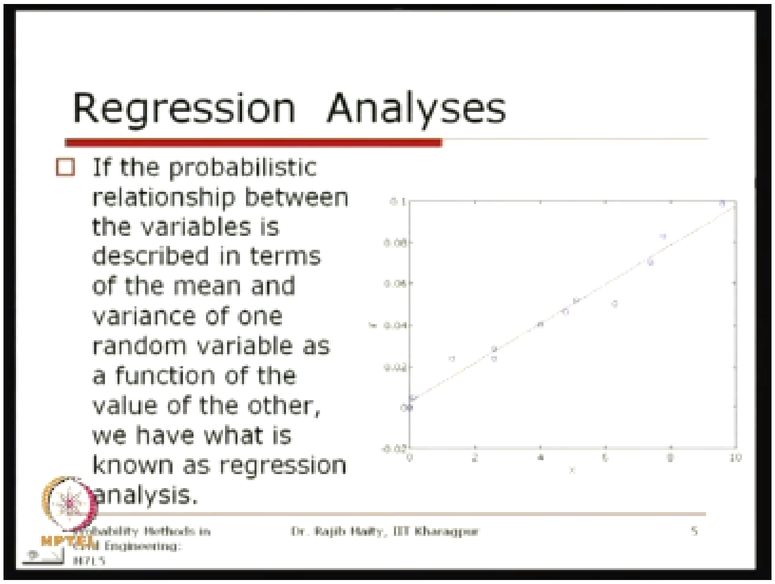
So whether now can we just estimate one relationship between this x and y. So that estimate that estimate of this functional relationship is that regression analyses that we will get through this analyses.

(Refer Slide Time: 08:24)



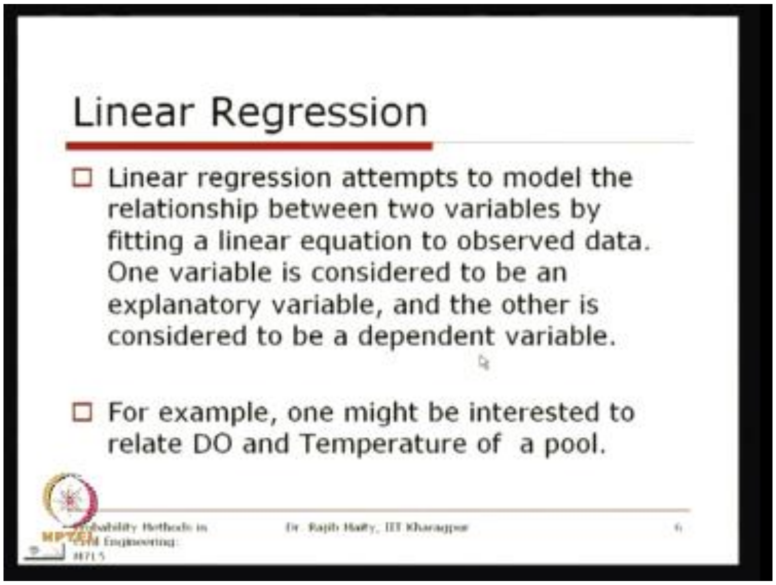
So first we will take that linear regression for examples.

(Refer Slide Time: 08:30)



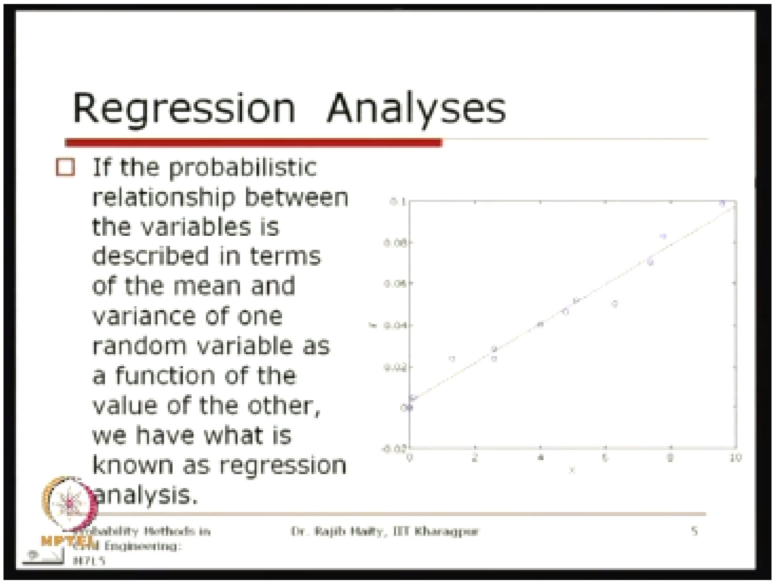
The diagram that is shown here we can see and we can expect that there could be a linear relationship can have it here. So but in many other cases where if just is looking this scatter plot we can see that initially it may be increasing and later on it may not increase in that rate. So there could be we can expect that there might be a non-linear relationship can happen. So the first what we are taking up is that linear regression.

(Refer Slide Time: 08:57



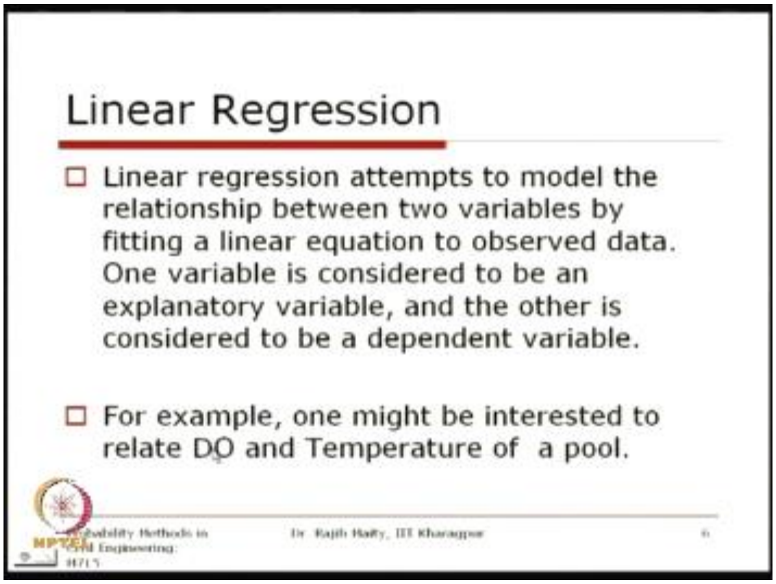
Where the expectation that the relationship is linear between the dependent variables and the target variable. So the linear regression attempts to model the relationship between two variables by fitting a linear equation to the observed data, one variable is considered to be an explanatory variable and other is considered to be a dependent variable. So that is what so our target.

(Refer Slide Time: 09:33)



So in this example that we have seen what we can use is that this variable x we can use as to be that your dependent variable and this is a y is my target variable. So I can use the information of x and I can model this y, it can be it could not be opposite also if we can we if we estimate x with respect to the variable y. So then we generally say that x is regressed on y and in other way the y is regressed on x.

(Refer Slide Time: 10.03)

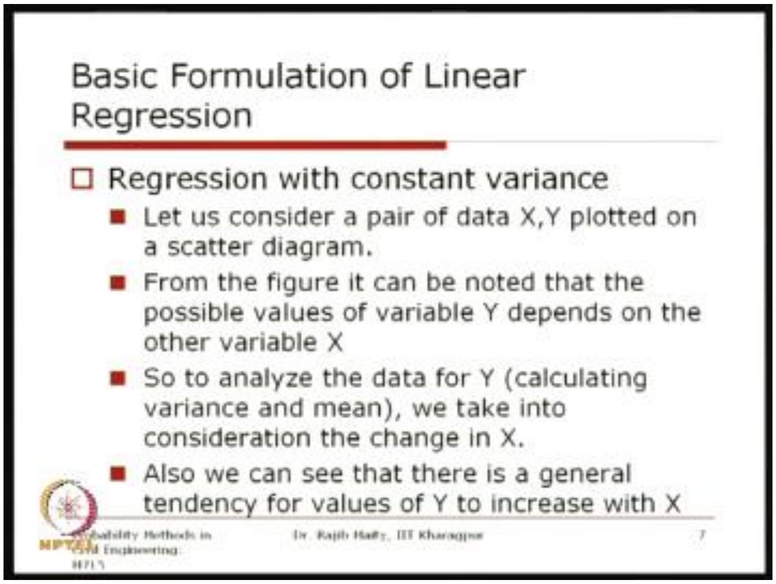


For example that one might be interested to relate the dissolve oxygen and the temperature of a pool. So whether the dissolve oxygen and temperature these two data is generally first collected and then we can see that whether the their relationship how the relations how they vary with respect to each other whether in the sense that I can see it in the both sides whether that DO given the temperature or the temperature given what is the DO.

But sometimes in case of this the practical consideration may be we are interested to know that our what is our target what is the what should be the dependent variable and what should that target variable for example, the example that is given here the dissolve oxygen and the temperature generally what we see is that temperature we use as a dependent variable and this dissolve oxygen is the target variable.

So this depends on the in what area in what practical field that we are that we are applying this analyses.

(Refer Slide Time: 11:10)

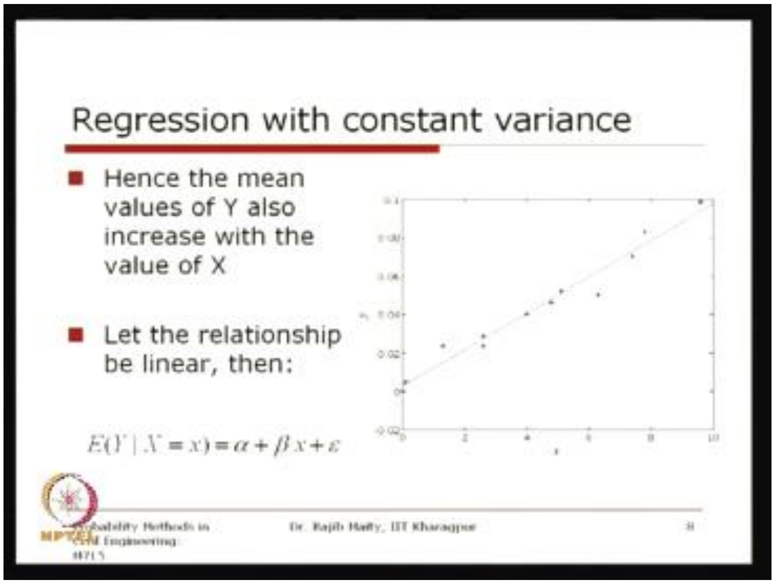


The basic formulation of linear regression with the constant variance is first. So here as I was state telling as a starting that when we are taking that the constant the variance of the dependent variable over the entire range of the dependant variable it remains constant. So in that case we generally say that this regression with a constant variance and by default when you say the regression analyses we generally mention that it is with the constant variance.

So the non constant case is a special case that we will take that we will see after sometime. So in case of this regression with constant variance let us consider a pair of data XY plotted on the scatter diagram as just few slide before you have seen, from the figure it can be noted that the possible values of the variable Y depends on the other variable X. So to analyze the data for Y calculating variance and mean.

We take into consideration the change in X and also we can see that there is a general tendency of the values of Y to increase with the X. So these are some of this example is given with respect to that plot the scatter plot that we have seen few slides before.

(Refer Slide Time: 12:35)



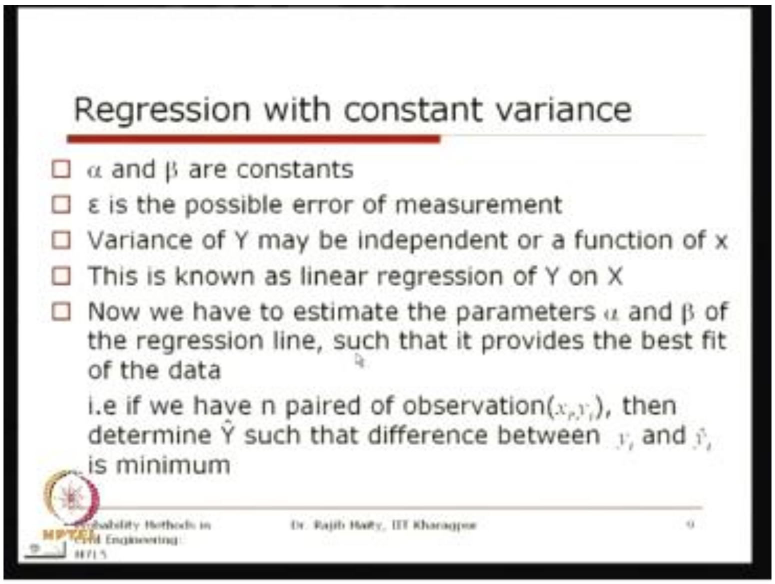
So here again the similar plot has been shown here. So here that one variable is X and other one is the Y. So here we are taking the case that we will regress Y on X. So X is our dependent variable and Y is our target variable. So here you can see that when X increases Y also increases and vice versa. So we have to fit a linear relationship between this X and Y. So hence the mean value of mean value mean values of Y also increases with the value of the X.

So as X increases that mean value or the in the statistical sense the expected value of the Y also increases. So the relationship let the relationship be linear because we are discussing this in a linear regression now. So the expected value of the Y given X a particular value X. So you know. So this is the conditional expectation. So if I just take what is the expectation of the Y you know the expectation of the y means without any other information.

So whatever the Y we see that it can from this diagram we can see that it varies from 0 to0.1 say. So whatever the values the range that we see we will just take its mean and that is the expected value of the Y. Now this when you are fitting this relationship that means that means it is a condition on the given value of this X now if I give some value of this X at 6. So in this part what is the expected value of this Y.

So this is now becomes the condition and this conditional mean is expressed through this linear relationship which is α + ß + ε. So this is you know this is the equation of that of the straight line plus some error term should be there to express that what is that value of that the mean value of this y.

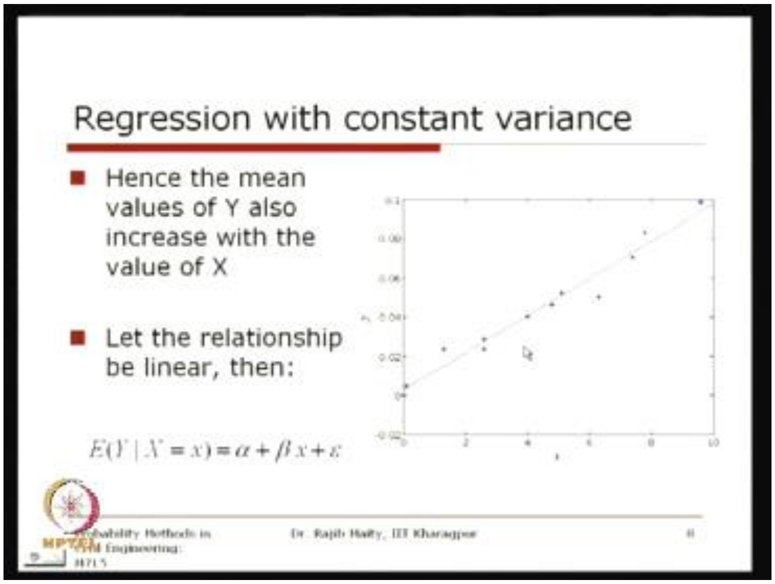
(Refer Slide Time: 14:42)



Now this α and ß are the constant and ε is the positive error of this measurement of the sorry possible error of measurement. So if when we take that data observe that data, so there could be some in that measurement, so there could be some errors. So that error is expressed through this ε variance of Y may be independent or a function of X this is known as the linear regression of Yon X that is what I was telling.

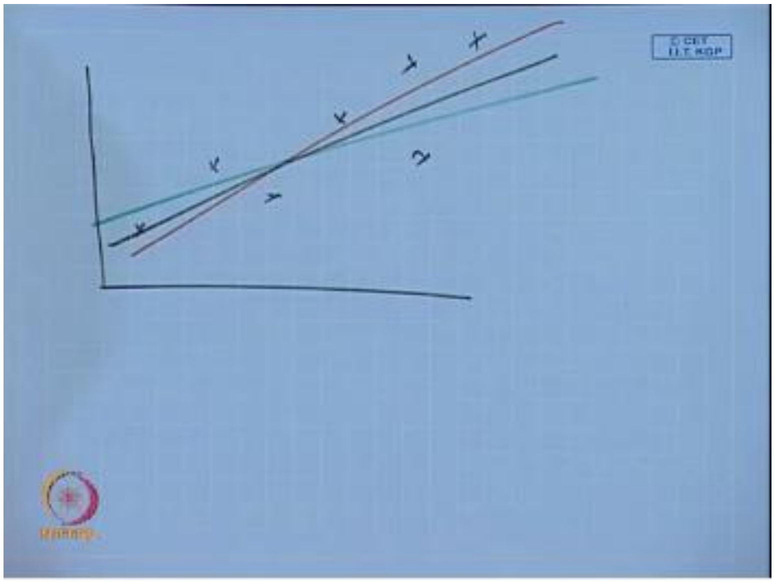
So it is Y on X, it can be expressed in other way also that is X on Y. So the relationship will change that expectation of X given Y is equals to some constant plus the ß multiplied by that your Y + ε. So that is the observational error. So now we have to estimate the parameters of this α and ß of the regression line such that it provides the best fit of the data, now this best fit of the data means if I just see this one.

(Refer Slide Time: 15:48)



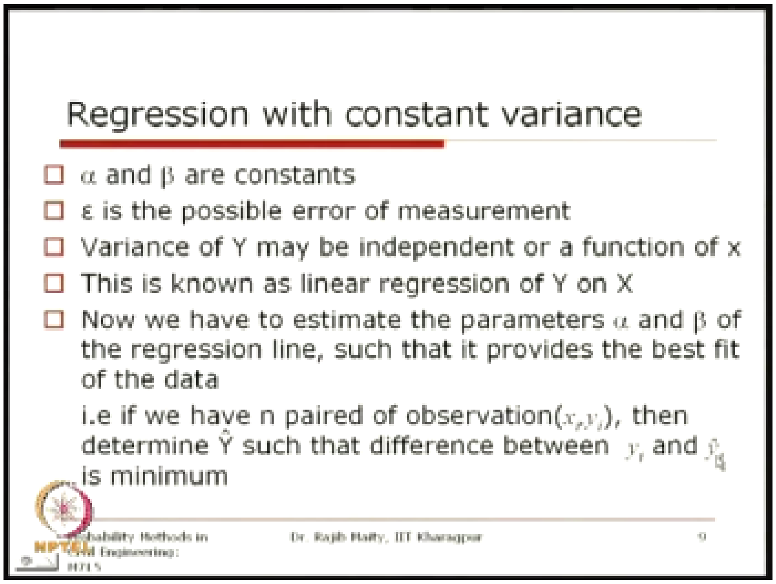
This scattered diagram. So there could be the various possible lines that I cannot think through these points, now which line should be the best fit line so what is meant here is this.

(Refer Slide Time: 16:09)



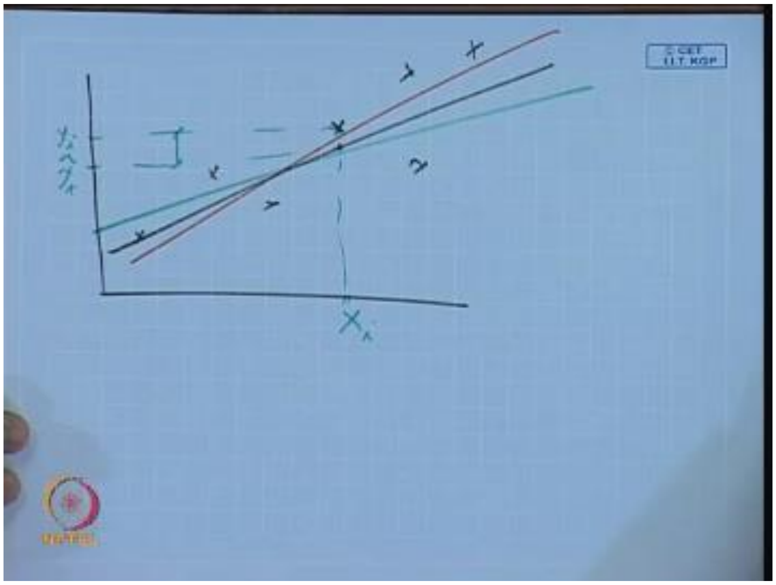
So if this is these are the data points then there could be the there could be some lines which can be described through different straight lines now out of these lines the possible lines which one should be the best fit now this to get that best fit so to get that best fit we have to follow some methodology which is known as the method of least square to estimate that to get that the line that is best fitting through this points and based on that we will get what is the that estimate of those regression constants so this is what is mentioned here.

(Refer Slide Time: 16:51)



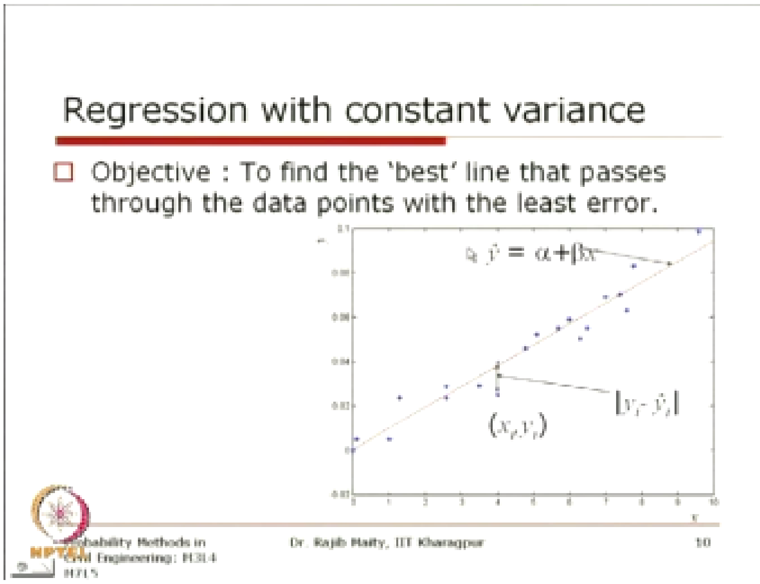
That is such that it is provide the best fit of the data that is if we have n paired of this observation X_i Y_i so these are the paired observation and these are one pair is one point on that diagram then determine y cap in such that the difference between that y_i and y_i cap is minimum now what is this y i I and y_i cap is if I referred to this diagram is this so this is your that point.

(Refer Slide Time: 17:26)



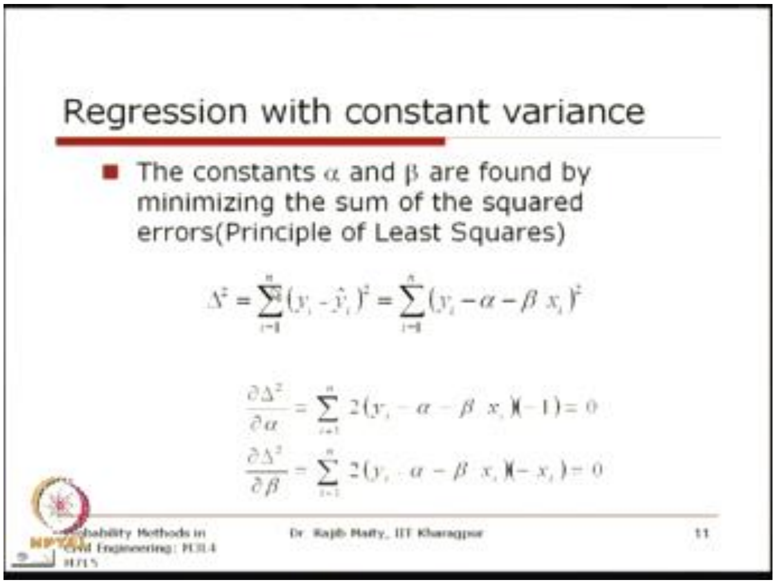
Where you can see that this is your some this is that y_i and if the whatever suppose the this black line is your best fit line then this with respect to this x with respect to this x_i so the estimate is this one so this is your y_i cap so the difference between these two is the error which should be minimized so now as close as this point to this observation and this is for all the points then that line should be the best fit line so the difference between that is why the difference between this y_i and y_i cap should be minimum to declare that the line is a best fitting through the data.

(Refer Slide Time: 18:23)



So it is now explained here so our objective here to find the best line that passes through the data points with the least error so now this blue stars are the observed data and this is the estimate of this regression line which is α + βx now this difference from what is the point that you can see and what is this corresponding point on this regression line there is a red line shown here is your error so this mode of this y_i - y_i cap is the absolute error for the point xi y I so y_i is known is the is the observed one and that y_i cap if I just put here this x_i then α +β these two as a constant putting this x_i what we will get that will be your y_i cap.

(Refer Slide Time: 19:22)



So the constants α and β are found by the minimizing the sum of squared errors sum of squared errors and this is known as this principle of least square so what is done is that this is the error that is y_i - y_i cap this is the error and that error is squared and summed up for all the observations so in this diagram if we see this is the error y_i - y_i cap and this is obtained for all these data points and this errors for individual point is first square up and that so that square error is summed of for all then observations that is available.

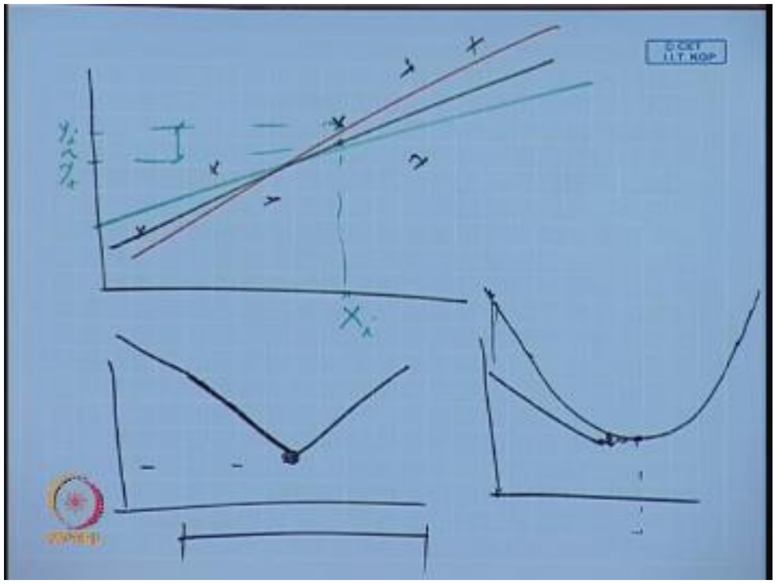
So this is giving is the sum of square errors and now this sum of square error now if I just replaced this y cap from that regression line which is your $\alpha + \beta x_i$ and we are taking this - so $y_i - \alpha - \beta x_i$ whole square is give you that sum of square errors now to get that estimate of this α and β so this error should before this α and β the value of α and β should be such that because these are the two constants which is basically which is determined everything about that straight line so error - this quantity should be minimum now to ensure that we have to take this partial derivative of this sum of square error with respect to each this parameters.

α and β and this has to be equated to 0 so we are having two unknowns and we are having two simultaneous linear equation that we can solved to get what is the estimate of this β before I proceed I need to take some time to explain this one why we have taken this square and this and we are using this as that total error so because you can see the first thing the first direct thing that you can that you can have from this diagram is that so for some points the error will be negative and some point the error will be positive -the error will be positive.

Depending on whether the point is below the regression line or above the regression line now when we are taking this square obviously those sign is going because we are interested to this what is the deviation from this regression line whether it is on the positive side or on the negative side that we are not interested when we are looking for the best fit line so whatever the error that we get if we take the square obviously that sign will go but this can also be think of that if we just take that absolute value of that error as it is shown it here then also that sign can go and if we just add them up then what we will get is that also it will give an the absolute error summation of the absolute error.

But generally when we go for this least square technique we take it to be the square and then we do this partial derivative this is because you know when we take.

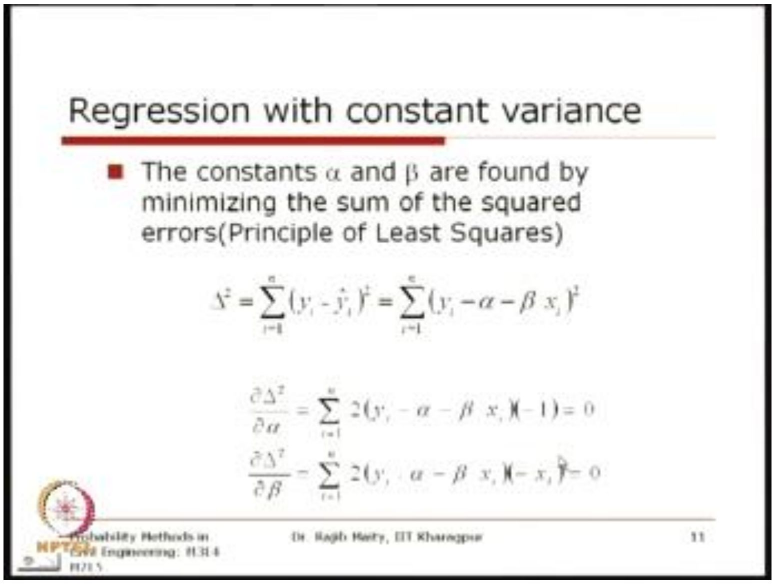
(Refer Slide Time: 22:58)



The error now if we just see that error and that error if we take it as a linear function basically what we are if we are minimizing it then this one basically our point is suppose this is our target point now this is the over that the possible range of the parameters now when we take this absolute error then the change with respect to that parameter it will be the linear one and when it take it to this to the square or the so this will become basically a quadratic function now what will happen if our estimates are far away from what is the optimum value then you know from the optimization technique.

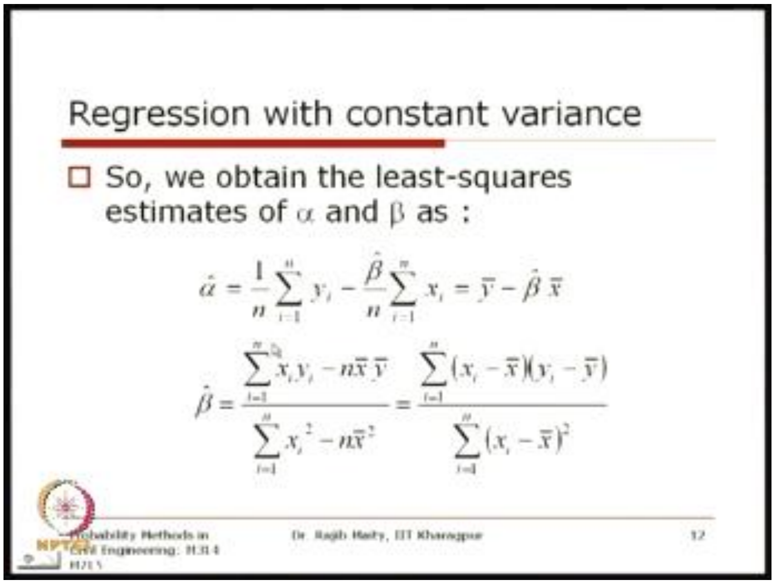
So if it is far away then the next step basically it will go very close to that optimum value and once it is comes to the optimum value then the steps will be smaller steps. But in this case generally that the step size are always same because this variation is linear here, but means this is basically when you go for this optimization optimizing the parameters that time it has been seen that this taking square is better than this taking this linear function.

(Refer Slide Time: 24:31)



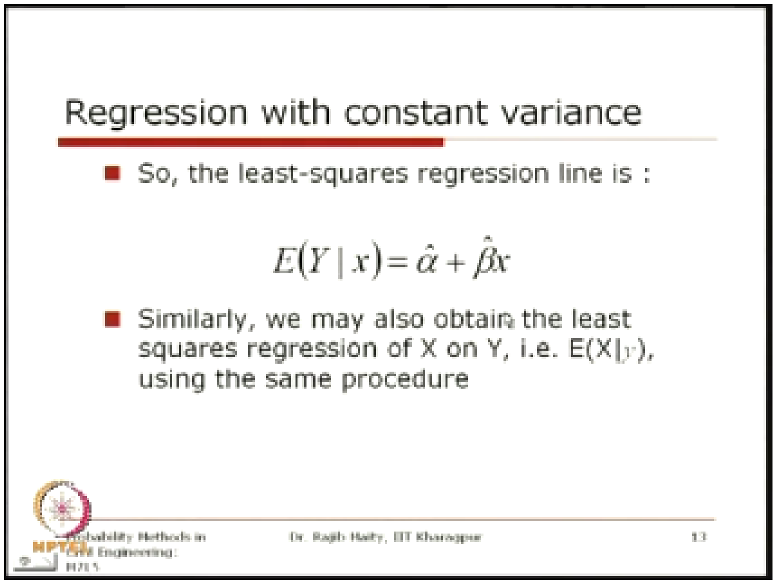
So, that is why so far as that sum of it is we are all generally interested for this sum of square error. So, what is done in this principle of least square technique. Well, so we got this after, through this partial derivative we get these two simultaneous linear equation where.

(Refer Slide Time: 24:50)



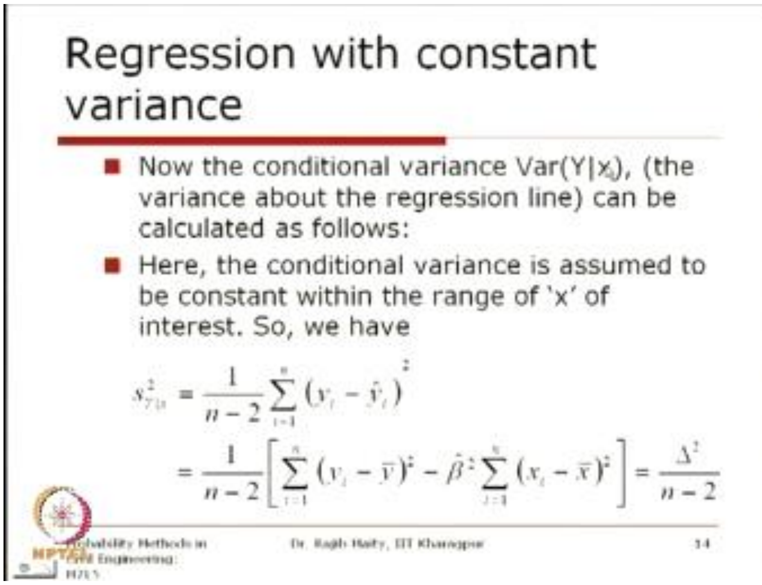
By solving them we can get the estimates like this that $\alpha^{\wedge}$ so this is now, this ^ symbol is given when we are referring to that it is the estimate. So, this $\alpha^{\wedge}$ if we can solve it and we can so that it will be the $\bar{y}$-$\beta^{\wedge}\bar{x}$ this $\bar{y}$ and $\bar{x}$ are the mean of this observed data and this $\beta^{\wedge}$ is the estimate of this β is can be shown that it is the ∑ of this $(x_i - \bar{x})(y_i - \bar{y})$ multiplication of them sum it over the all n observation divided by $(x_i - \bar{x})^2$ sum it over this all n observation.

(Refer Slide Time: 25:34)



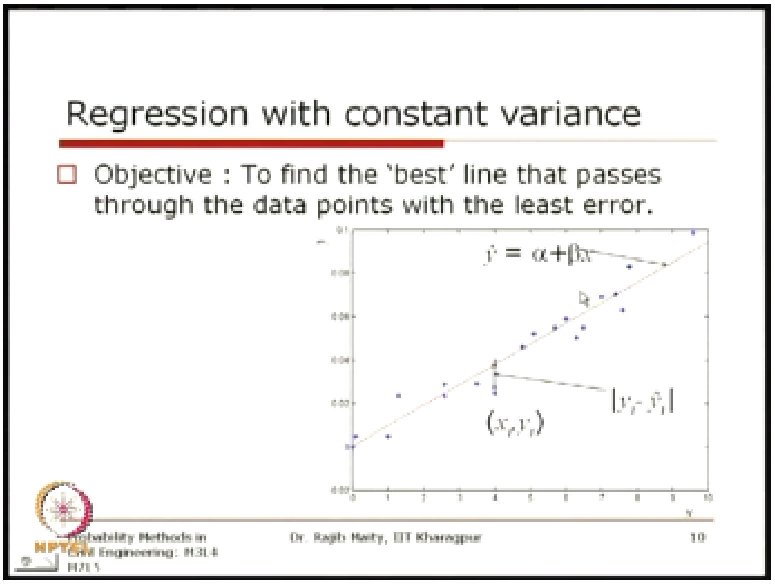
So, these two are the estimate of this α and β. So, the least square regression line is the expected value of this Y given that $x=\alpha^{\wedge}+\beta^{\wedge}x$. Similarly, we may also obtain the least square regression of this X on Y as I was mentioning that is that expected value of this X given y using the same procedure. But, here it will come as that and the dependent variable will be y. Obviously, that α and this β the estimate of these regression parameters obviously, will change through that if we follow that procedure whatever we have done.

(Refer Slide Time: 26:10)



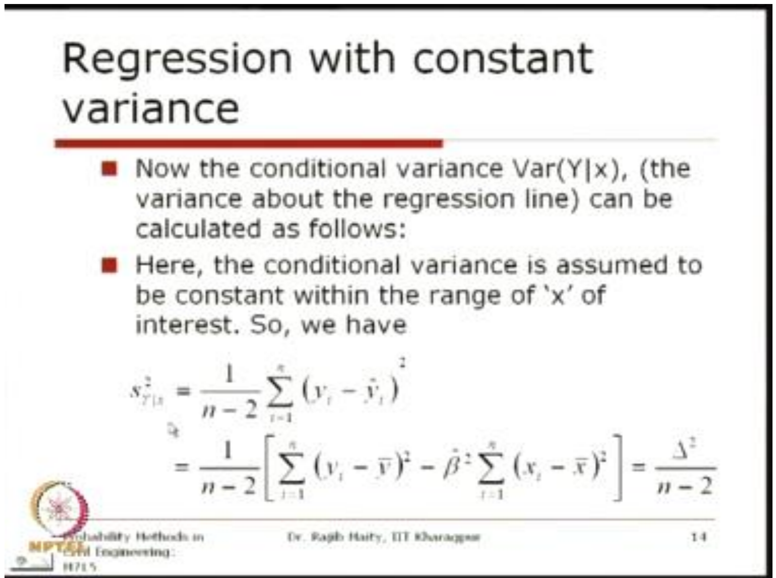
So, now the conditional variance that is now the conditional variance of this now the variance of Y given x. So, whatever we have got if that just now is that expected value of Y given x. So, now, we are interested to know, what is the conditional variance of Y given that x, so the variance about the regression line basically, what we meant here is that if I just referred to this diagram.

(Refer Slide Time: 26:47)



That this is the, if this is the y so we can see that it is varying from this 0 to 1 so, whatever the y observed data that we have got that we got and which we know how to obtained that it is a sample estimate of this variance if we do so that will give you the variance of the y. Now, after we get this regression line now what is the variability of the y with respect to the regression line. So, basically we are looking through this access and we see that how it is varying across this regression line.

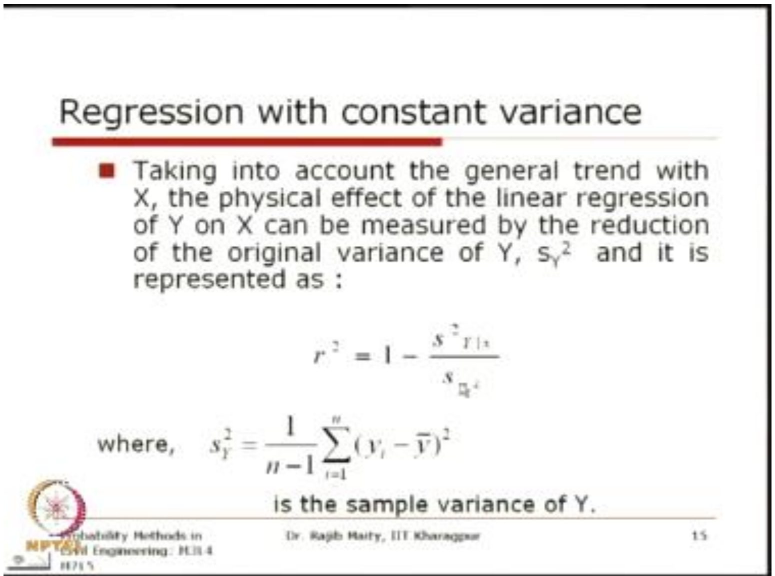
(Refer Slide Time: 27:26)



So, that is what is refer to as means pictorially as this variance of Y given x. Now if we want to estimate that one if we want to calculate that one this can be calculated as follows. Here, the conditional variance is assumed to be constant within the range of this x. So, we have this $s^2_{Y|x}$=1/n-2, i=1 to n(yi-$\hat{y}_i$). So, this is the basically the estimate from this regression and there are this n-2 is to make it that what it is called that unbiased and this you know that in the standard deviation we have seen that one degree of freedom is lost and that is why we make it that n-1 that we discussed in the earlier lectures and here one more degrees of freedom is lost when we are estimating that regression line.

Basically, there are if we see that there are two parameters that both α and β has to be estimate through this regression line and that is why the two degrees of freedom is lost. So, make this estimate unbiased it is 1/n-2 that we have to make. So, we can just do this we can sometimes for this we can make that $(y_i-\bar{y})^2$-$\beta^{\wedge 2}$ i to1 to n, $(x_i- \bar{x})^2$. So, this is just from this equation and you can see that this is basically that error and so, sum of square errors. So, which is that Δ^2/n-2 that is that conditional variance of Y given x.

(Refer Slide Time: 29:12)

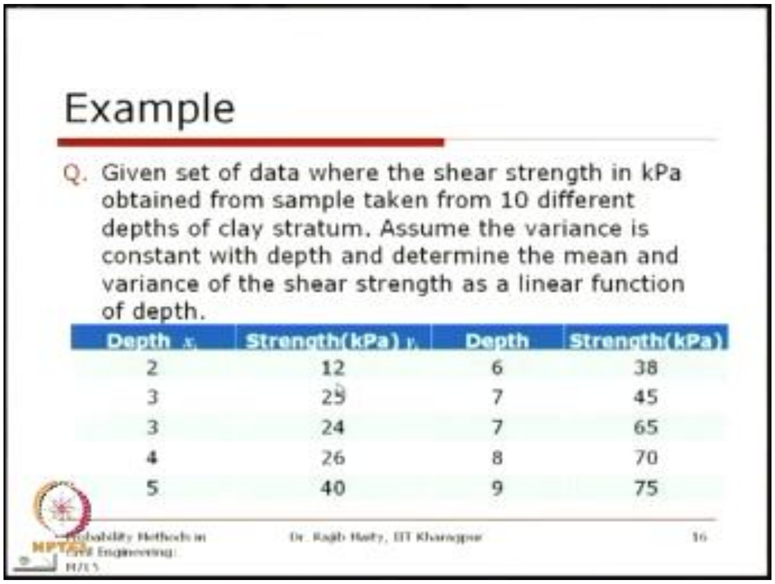


Now, taking into account the general trend with X, the physical effect of this linear regression so what actually is happening through this linear regression is that, Y on X that is the regression Y on X can be measured by a reduction of the original variance of this Y. So, the original variance of this Y you know that which is that a S_Y^2 which I can get the data of this Y and we can estimate what is that, what is this variance from the sample and that but through this regression when we do it that there is a regression there is a reduction in that variance.

So, which can be expressed through this one minus that variance of this Y given x divided by variance of this variance of Y, sorry, this square will that S^2. So, what you can see is that this is the original variance that was there in the data Y and this is the variance after the regression that we that we got. So, this is basically how much is the reduction then that reduction is that what is the total minus what we got after this regression divided by what was their the total.

So, this is that reduction in that variance and later on, we will just show you that this can be approximated to basically the correlation coefficient. Obviously, the square root of this one is, can be approximately equal to the correlation coefficient and that is basically the measure of how strong the relationship that we have measured. Where this S_Y^2 means this is the one by that you know it is the sample estimate of this variance of the data Y. So, $1/n-1\sum$ of from 1 to n (y_i- sorry, $\bar{y}^2$.

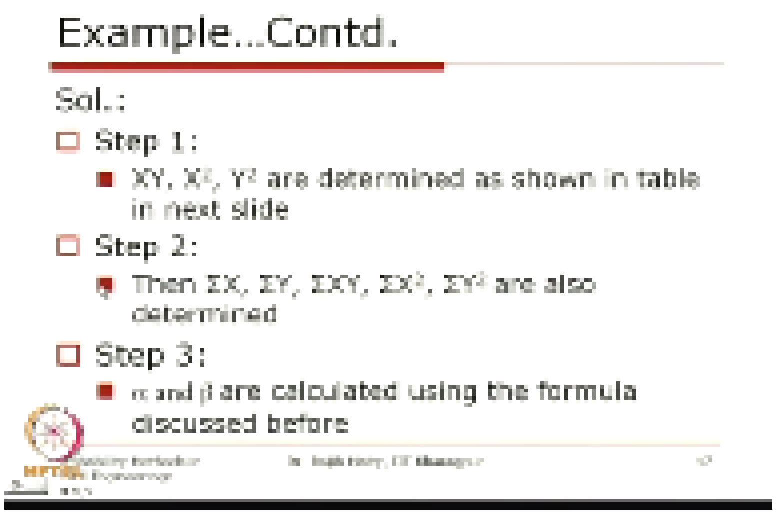
(Refer Slide Time: 31:07)



We will take one problem whatever we have discussed through the, for this linear regression so given the data where the shear strength in kPa obtained from the sample taken from 10 different depths of this clay stratum. Assume that the variance is constant with the depth and determine the mean and variance of the shear strength as a linear function of the depth. So, here you can see that there are depths are given the depth at 2,3 again this 3.

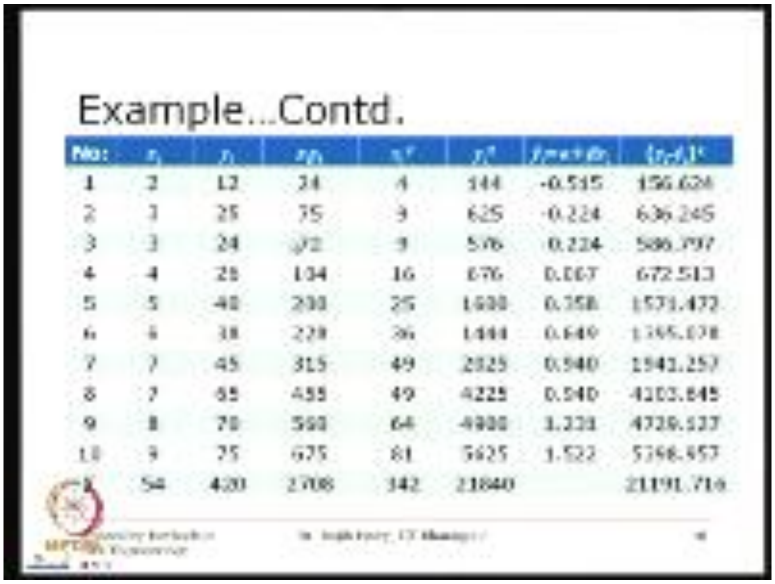
So there are 10 such depths are taken there are some depths are same, you can see here and we are getting this data. So, for the strength that in kPa. So we are having these10 different dataset this is the depth and this is basically this depth is going to 3,3,4,5,6, 7,7,8,9 and these are the corresponding strengths. So, this 10 data set that we are having and we will follow whatever we have discussed just to find out the relationship between strength and the depth. So, we will regress the strength on depth. So, we can say that our variable y is here the strength and the x is here depth.

(Refer Slide Time: 32:26)



So, to get this estimate through this least square technique, we will first get the estimate of this parameter; that is XY,X square are determined and this will be, we will show in this table and then the ∑ of this X∑ of Y∑,XY∑X square and Y square and then, the α β are obtained for using the formula that we obtained through that from that least square estimate.

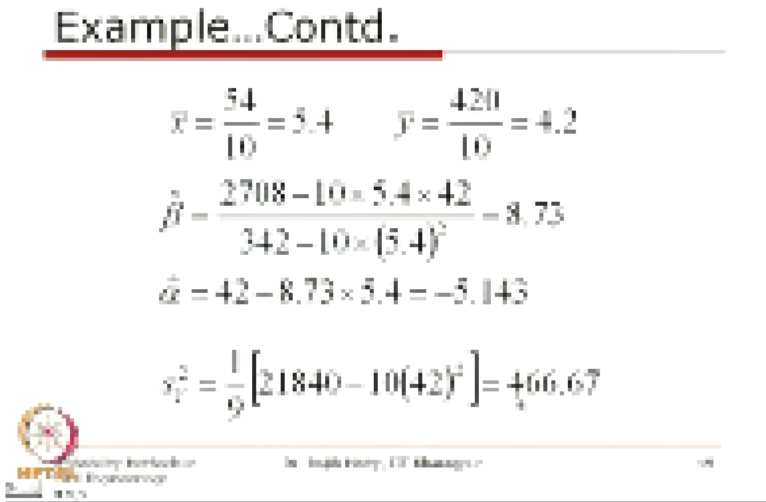
(Refer Slide Time: 32:52)



No.	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	[illegible]	[illegible]
1	2	12	24	4	144	-0.515	156.624
2	3	25	75	9	625	-0.224	636.245
3	3	24	72	9	576	0.224	586.797
4	4	26	104	16	676	0.067	672.513
5	5	40	200	25	1600	0.358	1571.472
6	6	38	228	36	1444	0.649	1395.078
7	7	45	315	49	2025	0.940	1941.257
8	7	65	455	49	4225	0.540	4203.845
9	8	70	560	64	4900	1.231	4720.527
10	9	75	675	81	5625	1.522	5398.457
Σ	54	420	2708	342	21840		21191.716

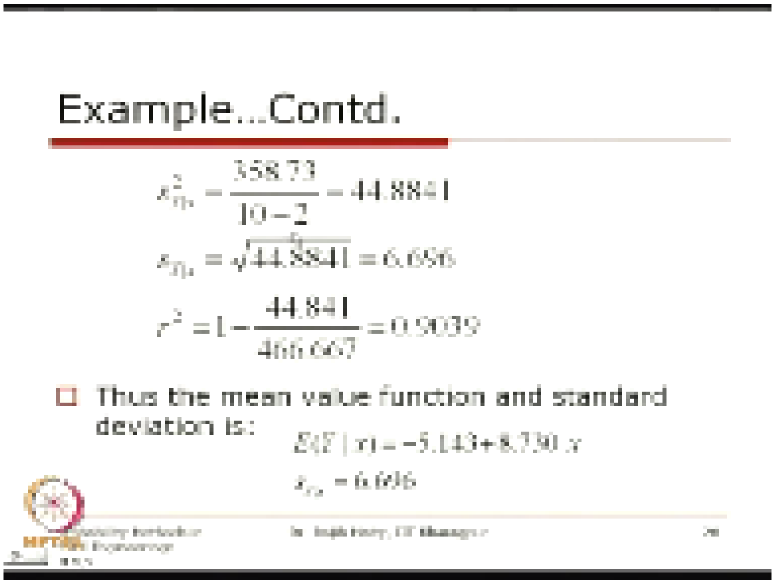
So, this is the data for this different depth and for this 10data sets are there xy, xi square y i square and then, these things we will just see. So, first we are having up to this and we are having their ∑ also. So, up to this of this table we know.

(Refer Slide Time: 33:13)



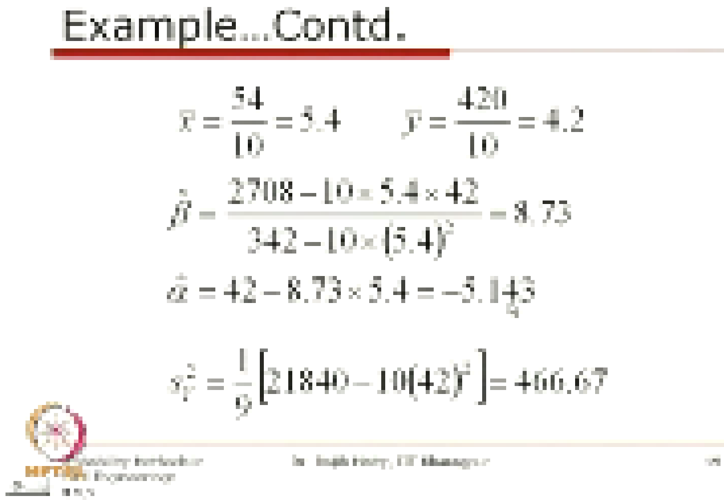
And using this information that is what is the power x bar is 5.4y bar is 4.2,sorry,it is 42thiswill be 42 420b/10; so, it is 42.So, this β 1you know that this expression we will use. So, and we will get that estimate of this 8.73 and α cap is the estimate of the - 5.143 and this sy square there is a variance the or the total variance I can say now the total variance of this y is466.67 now.

(Refer Slide Time: 33:50)



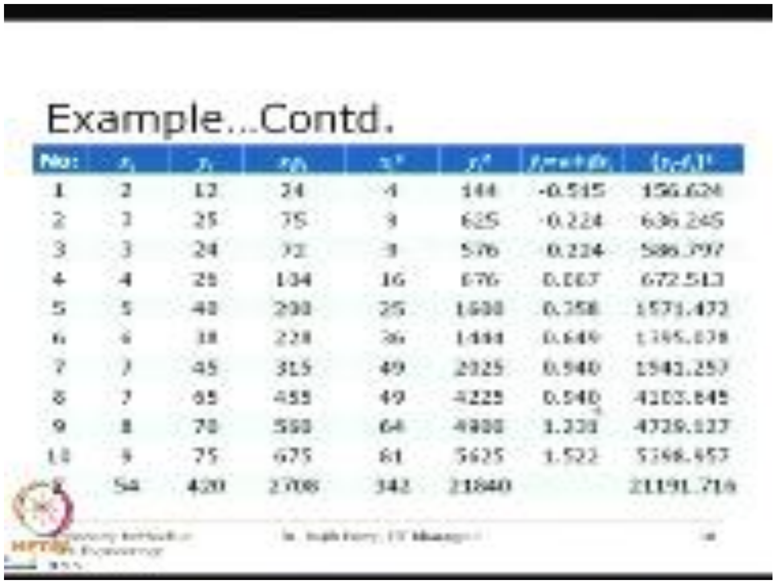
So, this now S y given that x is your this 44.88 and the standard deviation is 6.696andthisr square is equals to 1-these how much is the reduction is that 0.9039; and this α and β.

(Refer Slide Time: 33:50)



That you can see it here’s, α is - 0.50.143 and β is this1. So, this regression equation comes like this; that expected value of this y even x is that - 0.50.143 + 8.730 x. Now, using these things basically.

(Refer Slide Time: 34:40)

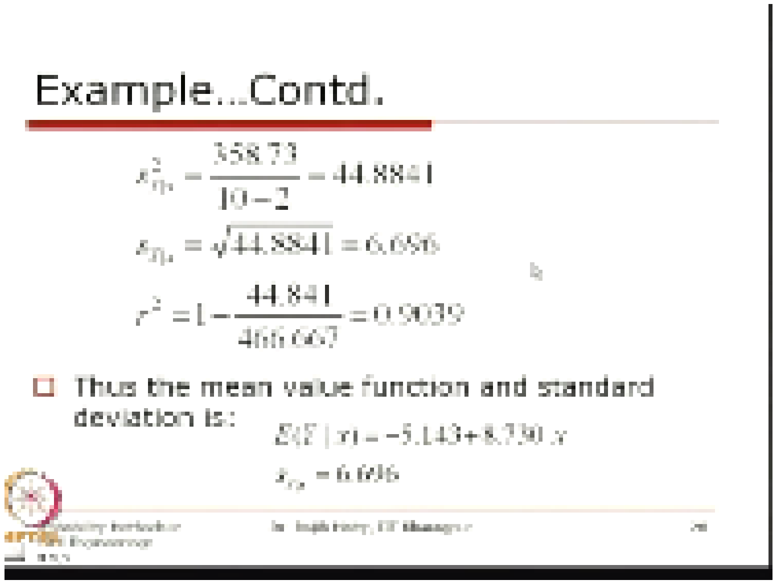


Example...Contd.

No.	x_i	y_i	$x_i y_i$	x_i^2	y_i^2	$\hat{y} = \hat{\alpha} + \hat{\beta} x_i$	$(y_i - \hat{y}_i)^2$
1	2	12	24	4	144	-0.515	156.624
2	3	25	75	9	625	-0.224	636.245
3	3	24	72	9	576	-0.224	586.797
4	4	26	104	16	676	0.067	672.513
5	5	40	200	25	1600	0.358	1571.472
6	6	38	228	36	1444	0.649	1395.038
7	7	45	315	49	2025	0.940	1941.257
8	7	65	455	49	4225	0.940	4103.845
9	8	70	560	64	4900	1.231	4739.127
10	9	75	675	81	5625	1.522	5366.457
Σ	54	420	2708	342	21840		21191.716

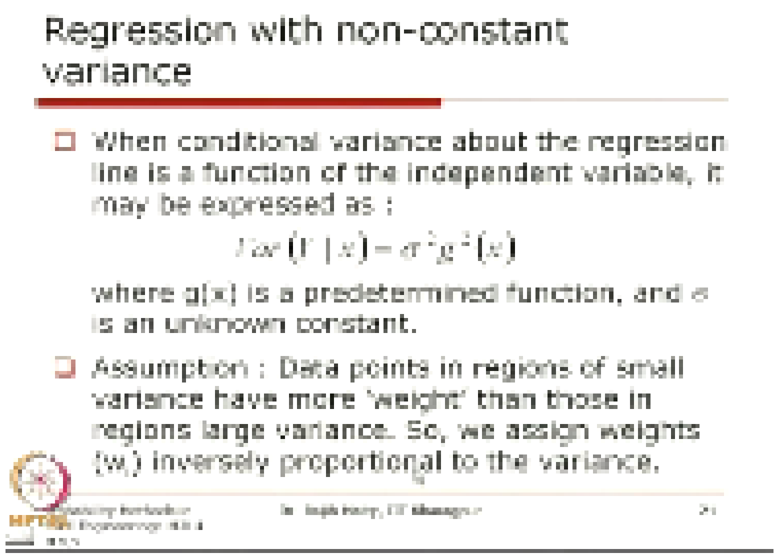
This relationship in the table we got this expression first. So, we are putting this x input and that α estimate and β estimate and we get this one. From here we are getting what is their error of this that is y - y caps. So, this -this and that square will give you that basically what is this one that we get that error square now if we sum it up this is basically the sum of square error and we are using that information

(Refer Slide Time: 34:58)



n to estimate this what is that reduction in this that variability variance in y. So, this one we have seen. So, this is finally, that expected mean is expressed through this expression and expected variance is 6.697 and obviously, this is constant over the entire range of x.

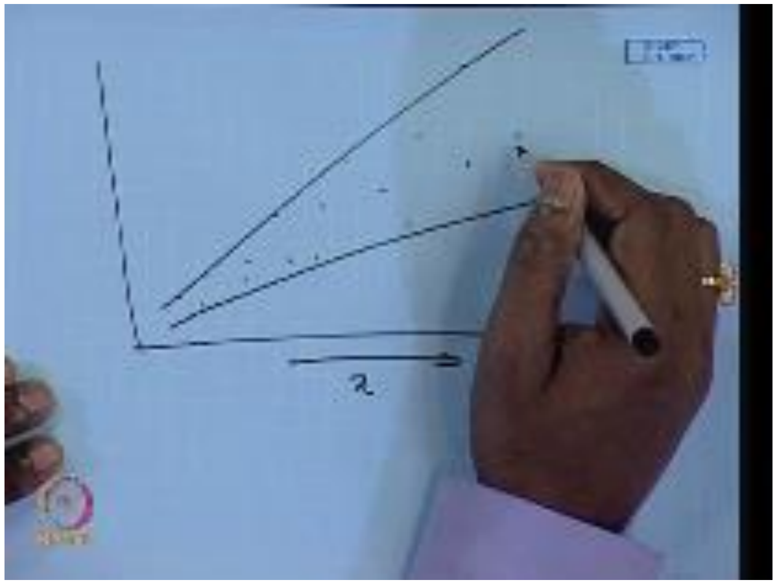
(Refer Slide Time: 35:23)



So, now we will deal with the regression with the non-constant variance now when the conditional variance about the regression line is a function of independent variable, it may be expressed as variance of Y given x is equal to sigma square multiplied by g square x. Now, this g x basically is the predetermined functions. Some function is that, how it is varying and this should be multiplied with the sigma square and when it is variance you know that any function or constant that is multiplied with this variance so that, we make it square. That is, we discussed in the earlier lectures. So, this is that sigma square g square x.

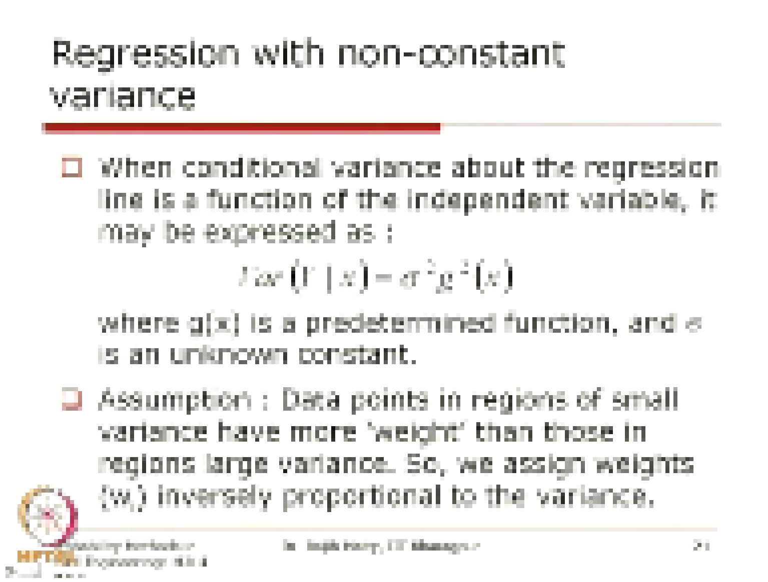
So, now, this sigma is an unknown constant and here the assumption is that data points in the region of this small variance have more weight and then those in the region of this large variance. So, we assign the weight wi inversely proportional to the variance. So, some weight we have to put and our assumption is that when the data is having the small variance.

(Refer Slide Time: 36:34)



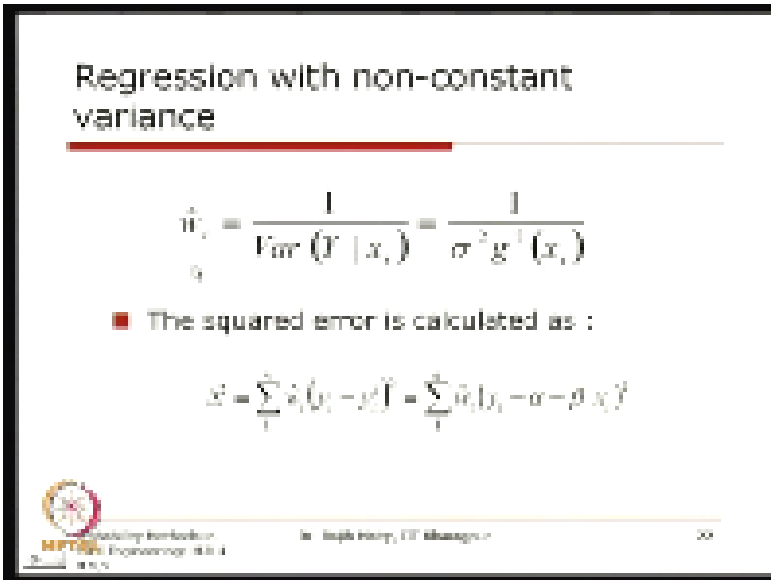
Now, if you see this diagram basically, if I say that this is varying means, suppose this what we can see in this literally. So, we can easily see that as it is going and so, if this the x as this x is varying basically the range is changing. So, here the weight will be more.

(Refer Slide Time: 36:59)



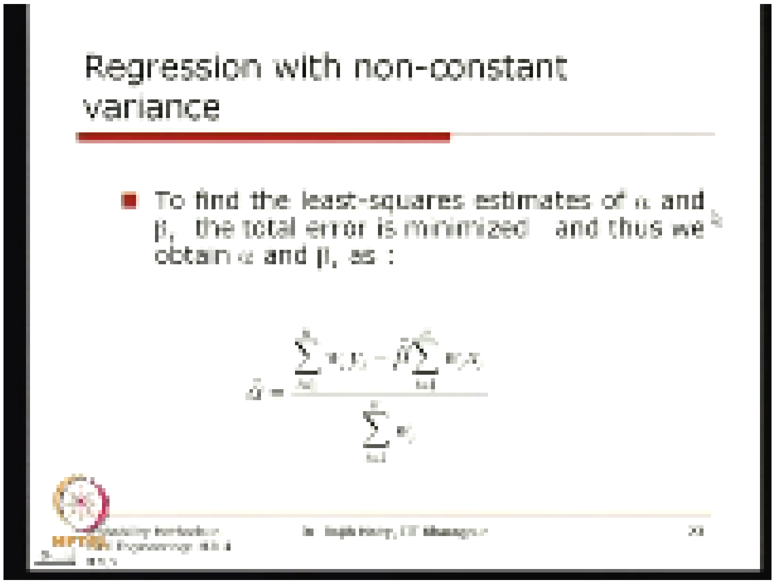
In this zone where the variance is less and here the weight will be less where the variance is more basically that is what. So in this way it is inversely proportional to the variance.

(Refer Slide Time: 37:03)



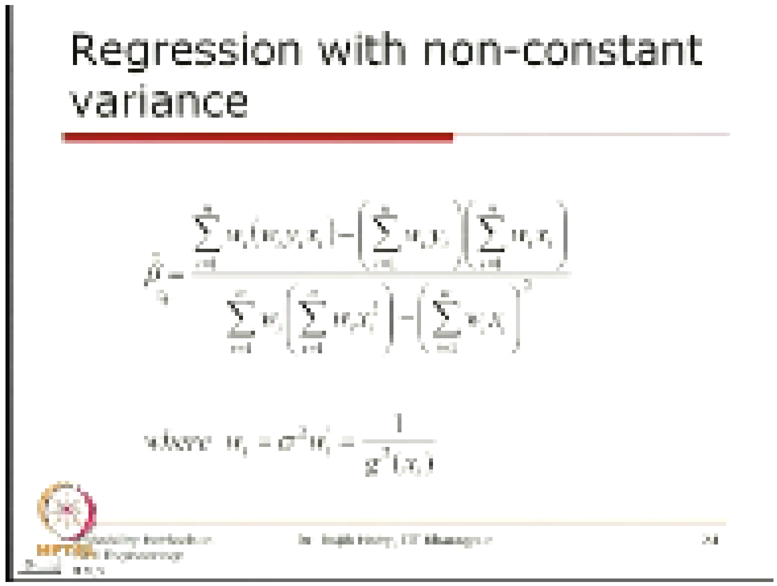
So, this how the weights are given. So, 1byvariance of Y given xi which is the 1by sigma square g square xi the square error is calculated as this sigma square is equals to that this weight is weights we will put and then that your that difference square and sum it up. So, this is the I equals to1to n. So, now this yi cap again that estimate will get from this α - β xi.

(Refer Slide Time: 37:37)



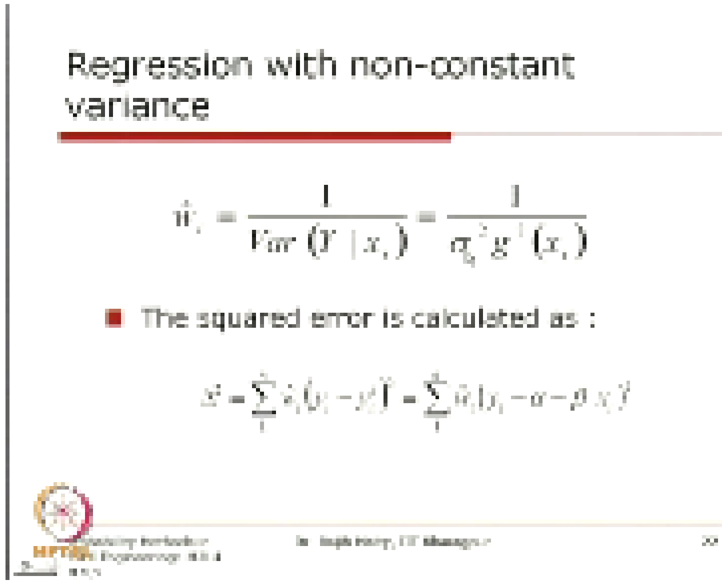
Now to find that least square estimate of α and β the total error is minimize and thus we obtain this α and β and following the same principal that we have discussed for this constant variance we will just get that error and error is partial derivative is taken equated to 0and after solving those equation we will get the estimate of α is equals to through this expression say wi into yi - β cap of this wi xi divided by wi ∑ of all this w i.

(Refer Slide Time: 38:08)



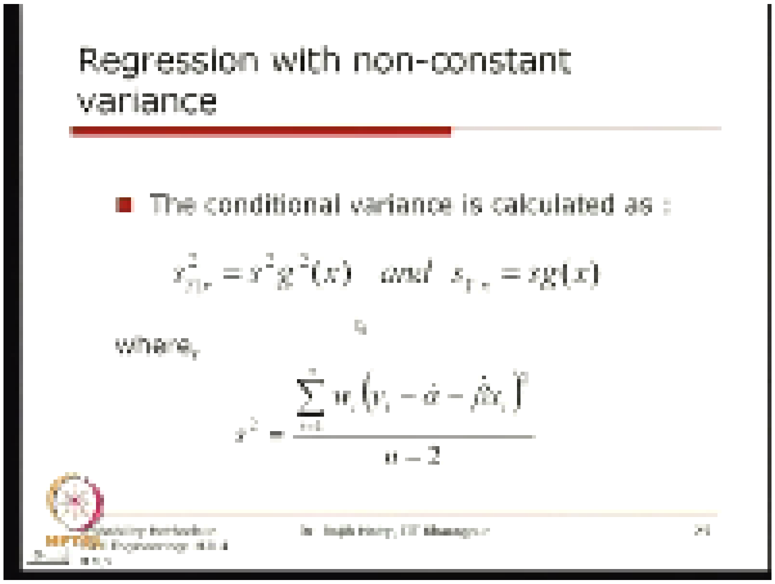
And, β cap will be obtain through this expression even though this expression looks through like a little bit cumbersome, but, thing is that this is we get following the same principal that we have done for this constant variance only thing here the one that weight function is coming and which the weight you can see that this weight is equal to sigma square wi prime and this sigma square. If we just multiply whatever the equation that we have used here.

(Refer Slide Time: 38:38)



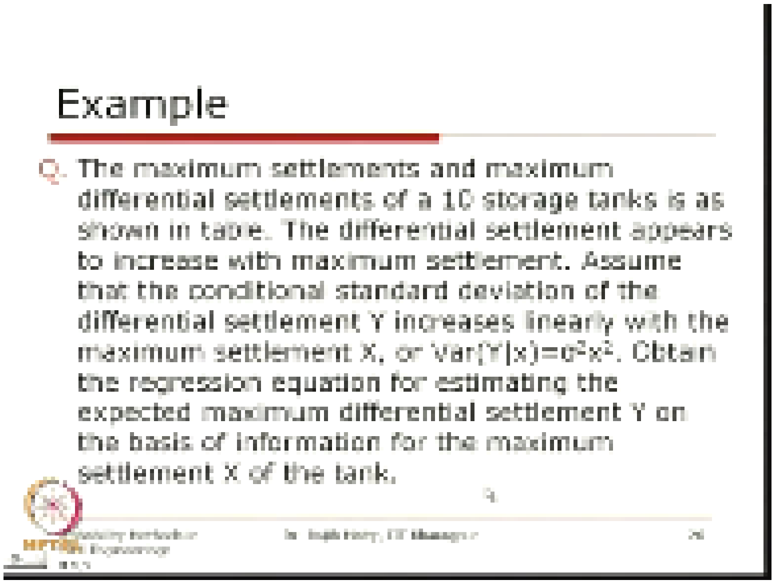
So, this sigma square will be cancelled. So, it will be 1by gx square. So, now this g x square generally same function we will used and that function of this x should be there to when we are determining this wi to get the estimate of this α and β.

(Refer Slide Time: 38:56)



And the conditional variance is calculated as S i square is the s square g x square and S xi –this is the standard deviation, the square root of this positive square root of this. So, s multiplied by the g x; you can see here that this conditional standard deviation is a function of that x where this Share is that ∑ o fi equals to 1- 1to n wi into yi - α cap - β cap xi whole square divided by n - 2.

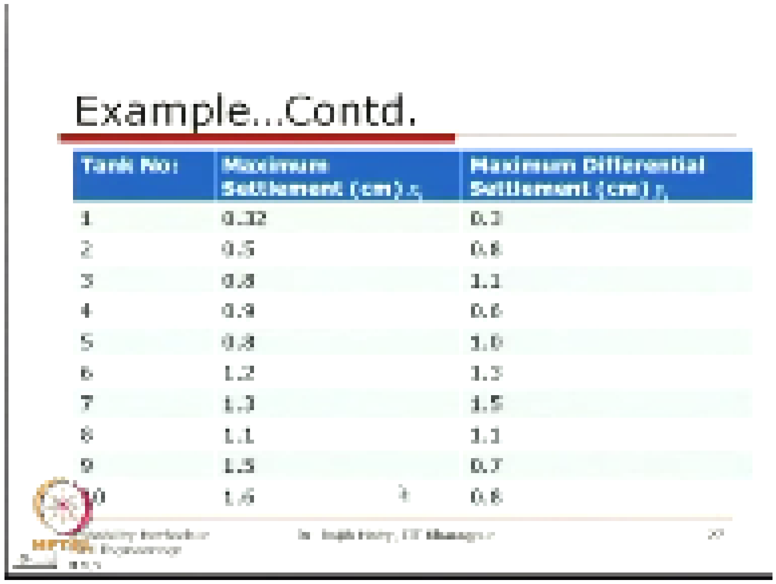
(Refer Slide Time: 39:30)



Now we will take one example on this one. It will be more clear in that way where the variance is dependent on the value of X. The maximum settlement and the maximum differential settlement of 10storage tanks, this is wrong; of 10storagetanks is as shown in the table. The differential settlement appears to increase with the maximum settlement assumed that the conditional standard deviation of the differential settlement Y increases linearly with the maximum settlement X or this is what is told.

That is, linearly it increases; that means, that g x that the function that we have told this is g x is equals to X .So, the variance of Y given X is equals to sigma square x's square; that function square obtained, the regression equation for estimating the expected maximum differential settlement on the basis of the information for the maximum settlement of X of that tank.

(Refer Slide Time: 40:38)

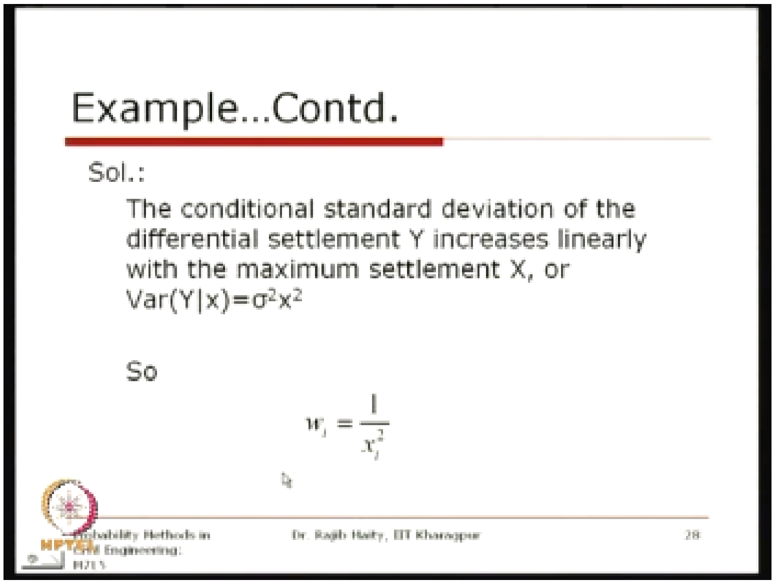


Example...Contd.

Tank No:	Maximum Settlement (cm) x_i	Maximum Differential Settlement (cm) y_i
1	0.32	0.3
2	0.5	0.8
3	0.8	1.1
4	0.9	0.6
5	0.8	1.0
6	1.2	1.3
7	1.3	1.5
8	1.1	1.1
9	1.5	0.7
10	1.6	0.8

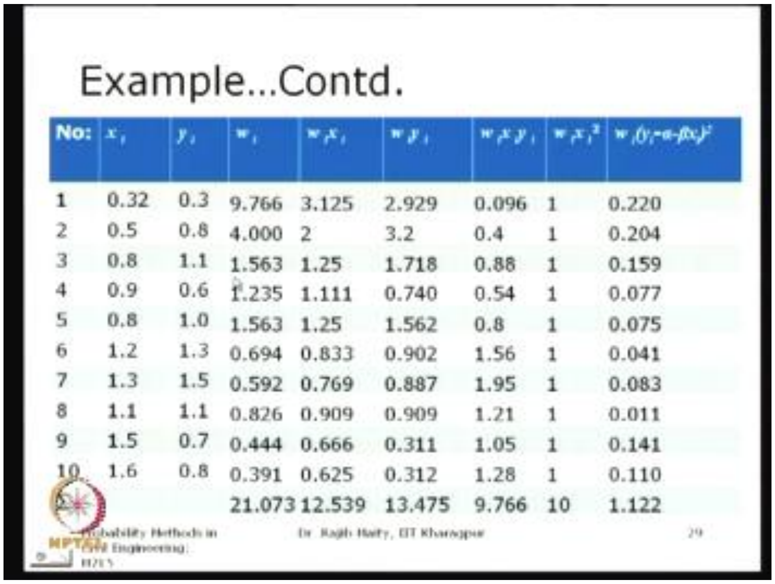
To do this one; this is the data that 10 different dataset is given here. So, this is the maximum settlement there is a maximum differential settlement. So, here we have to regressed that maximum differential settlement on this maximum settlement. So, our variable here in this following the notation that we have used is the this is of our y and this is our x.

(Refer Slide Time: 41:06)



So, and the conditional standard deviation of the differential settlement Y increases linearly with the maximum settlement X or variance Y on condition x is equals to $\sigma^2 x^2$. So, this is the relationship that is given. So, this x this function actually is predetermined as we have seen in that theory. So, here that w_i is the inverse of that functions.

(Refer Slide Time: 41:42)

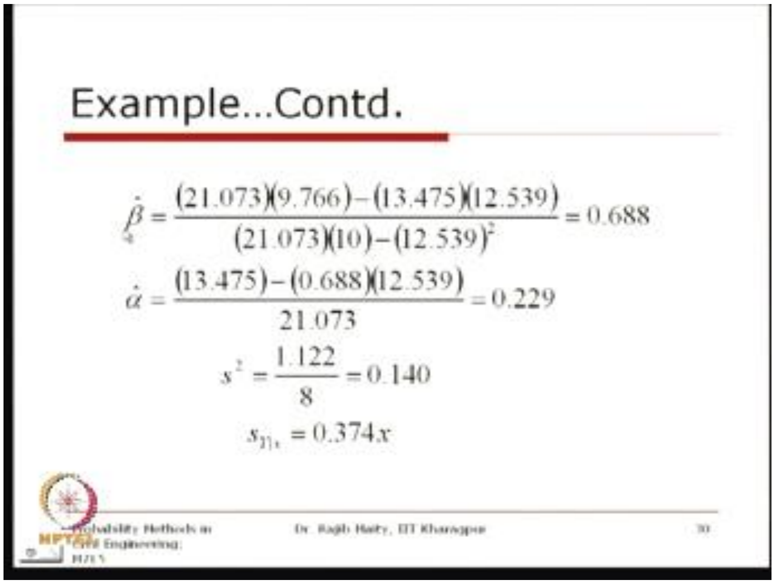


Example...Contd.

No:	x_i	y_i	w_i	$w_i x_i$	$w_i y_i$	$w_i x_i y_i$	$w_i x_i^2$	$w_i (y_i - \alpha - \beta x_i)^2$
1	0.32	0.3	9.766	3.125	2.929	0.096	1	0.220
2	0.5	0.8	4.000	2	3.2	0.4	1	0.204
3	0.8	1.1	1.563	1.25	1.718	0.88	1	0.159
4	0.9	0.6	1.235	1.111	0.740	0.54	1	0.077
5	0.8	1.0	1.563	1.25	1.562	0.8	1	0.075
6	1.2	1.3	0.694	0.833	0.902	1.56	1	0.041
7	1.3	1.5	0.592	0.769	0.887	1.95	1	0.083
8	1.1	1.1	0.826	0.909	0.909	1.21	1	0.011
9	1.5	0.7	0.444	0.666	0.311	1.05	1	0.141
10	1.6	0.8	0.391	0.625	0.312	1.28	1	0.110
			21.073	12.539	13.475	9.766	10	1.122

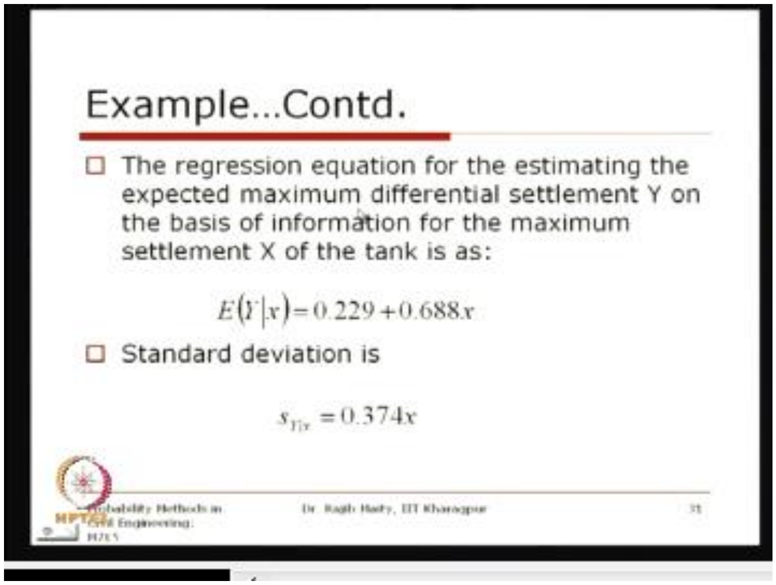
So, $1/x_i^2$ So, this is the weight is that is the wi and with that wi so for all these we will get this weight and basically, this is input. This is also we know the observed data this is the weight which is a inverse to this xi and. So, these things we can calculate wi xi ,wi yi, wi xi yi and wi xi square .

(Refer Slide Time: 42:10)



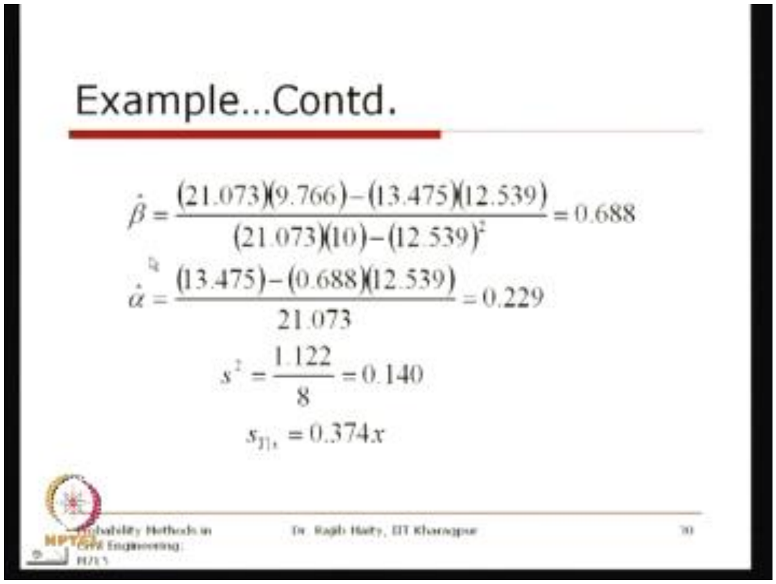
So, if we use this one and then up to this of this table we can calculate and based on this we will estimate that α and β and here the β is estimate of this β is 0.688 and estimate of this α is 0.229 and the S square is your 0.140 and this standard deviation of y given x equals to 0.374 x. So, you can see that as x increases this standard deviation also increases which is the function of this x of the variable x.

(Refer Slide Time: 42:48)

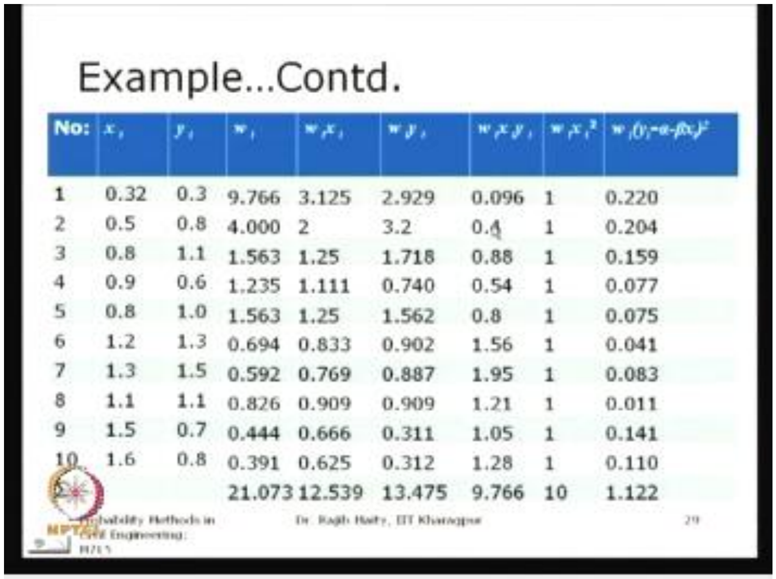


Now, the expected value of this y is the regression equation for the estimating the expected maximum differential settlement Y on the basis of information for the maximum settlement X of the tank is as expected value of this y given x is equals to 0.229 plus 0.688 x. So, this is that expected value of y given x and this is expected value of this is the standard deviation of y given x and now, using this relationship basically when how we are getting this 0.345 here that we have seen that S square is this one.

(Refer Slide Time: 43:21)



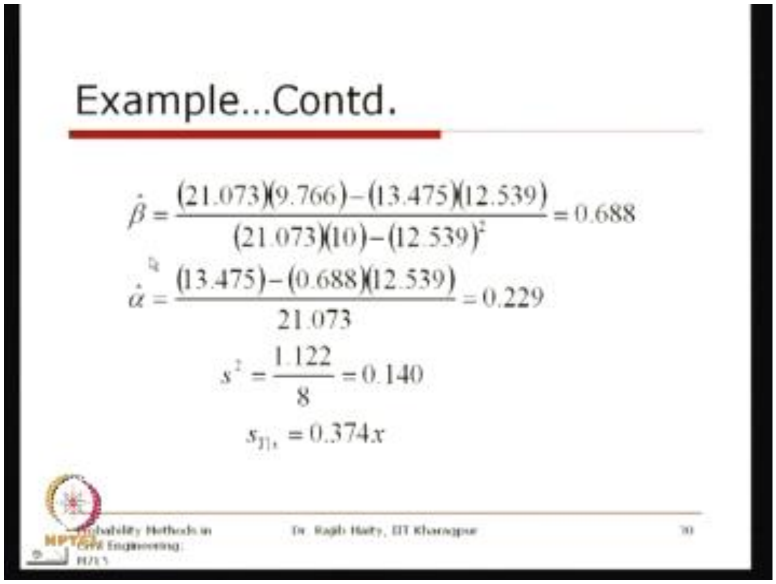
(Refer Slide Time: 43:22)



Example...Contd.

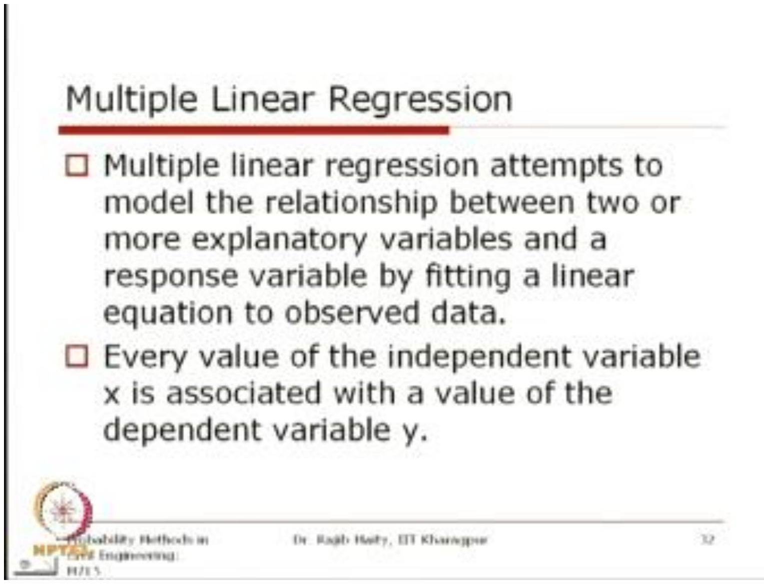
No:	x_i	y_i	w_i	$w_i x_i$	$w_i y_i$	$w_i x_i y_i$	$w_i x_i^2$	$w_i(y_i-\alpha-\beta x_i)^2$
1	0.32	0.3	9.766	3.125	2.929	0.096	1	0.220
2	0.5	0.8	4.000	2	3.2	0.4	1	0.204
3	0.8	1.1	1.563	1.25	1.718	0.88	1	0.159
4	0.9	0.6	1.235	1.111	0.740	0.54	1	0.077
5	0.8	1.0	1.563	1.25	1.562	0.8	1	0.075
6	1.2	1.3	0.694	0.833	0.902	1.56	1	0.041
7	1.3	1.5	0.592	0.769	0.887	1.95	1	0.083
8	1.1	1.1	0.826	0.909	0.909	1.21	1	0.011
9	1.5	0.7	0.444	0.666	0.311	1.05	1	0.141
10	1.6	0.8	0.391	0.625	0.312	1.28	1	0.110
			21.073	12.539	13.475	9.766	10	1.122

(Refer Slide Time: 43:35)



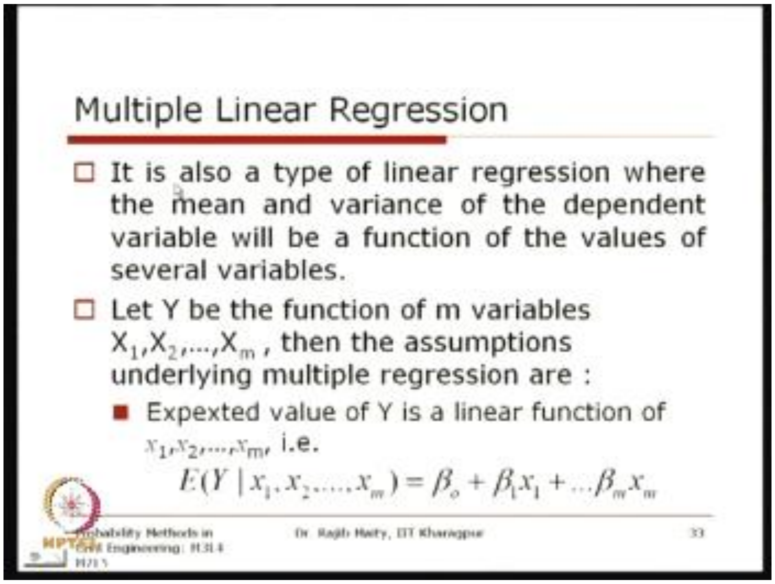
Basically, we are using this α and β estimate to calculate this one first and this total we are getting this is the sum of square error weightage sum of square and from there we are getting this s square and from there it we getting that given that variance, sorry, standard deviation of y given x 0.374x. So, next we will take that multiple linear regression and here you know that. So, far whatever we have discussed it is that regression and one dependent variable, one response, variable one target variable was there.

(Refer Slide Time: 44:05)



Now, in case when we are having that more than one random variable then we have to go for this more than one dependent variable then we have to go to this multiple linear regression. So, this multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed data. Every value of independent variable is associated with a value of the dependent variable y.

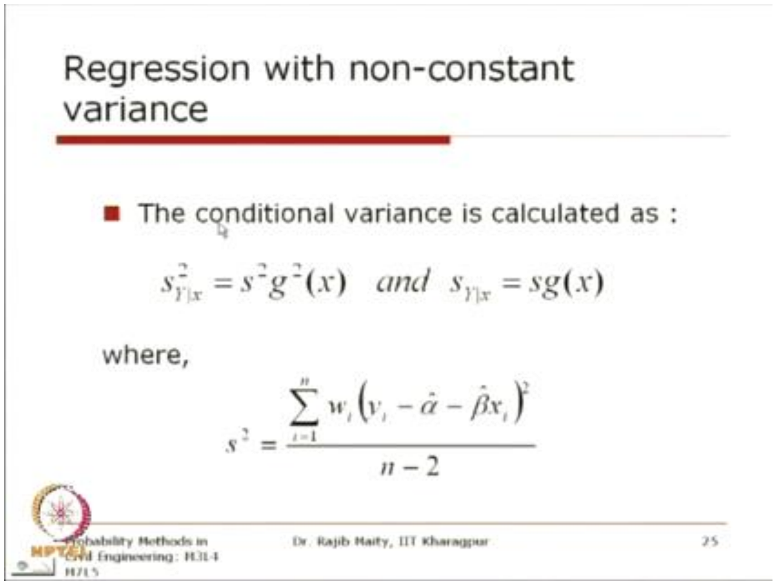
(Refer Slide Time: 44:44)



It is also a type of linear regression where the mean and variance of the dependent variable will be a function of values of this several variables. So, here instead of that using that one independent variable that is the x; that x and our dependent variable was a y earlier case. So here what you can see is that Y be the function of m variables instead of only one.

So, far what we have discussed here is, Y is the function of m variables which is x 1x 2 up to x m then the assumptions underlying the multiple regression are the expected value of Y is a linear function of x 1,x 2 up to x m that is the expected value of y given the information of this independent variable x 1,x 2up to x m equals to that β naught β 1x 1plus β 2x 2plus up to in this way β m x m. One thing I will just, one correction I will just do before I proceed further in this linear regression with one between x and y.

(Refer Slide Time: 46:00)



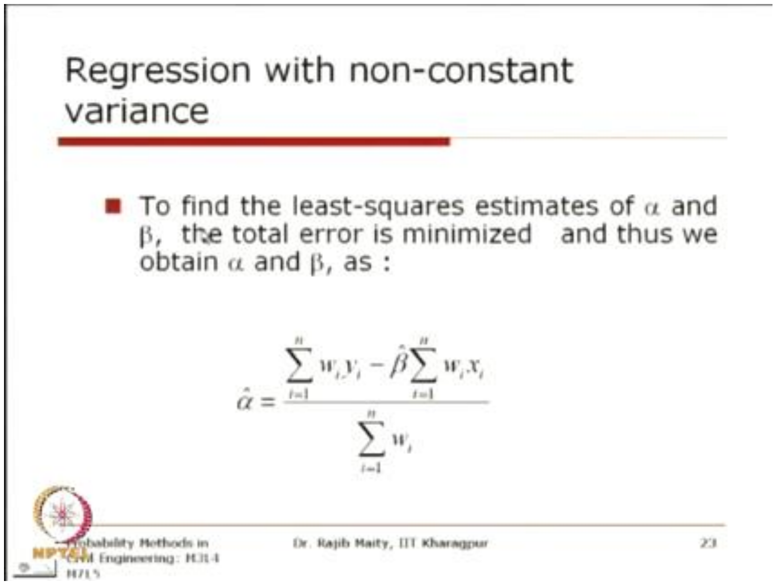
(Refer Slide Time: 46:01)

Regression with non-constant variance

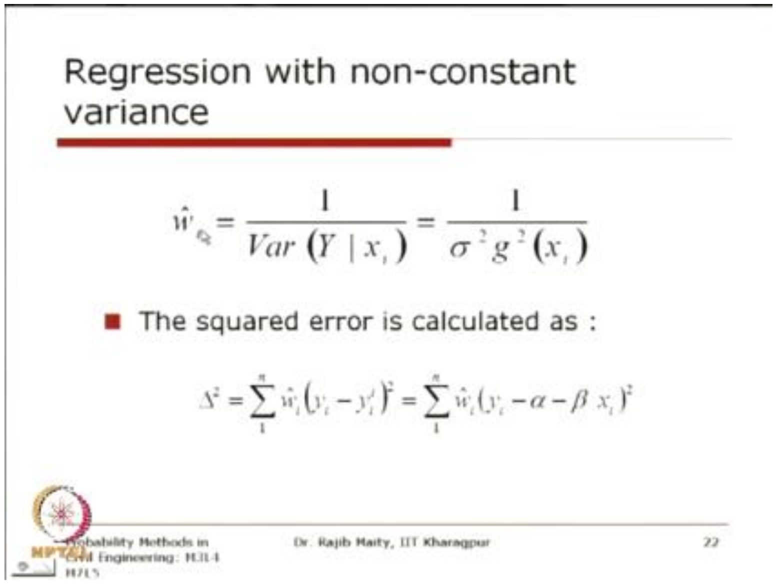
$$\hat{\beta} = \frac{\sum_{i=1}^{n} w_i \left(w_i y_i x_i\right) - \left(\sum_{i=1}^{n} w_i y_i\right)\left(\sum_{i=1}^{n} w_i x_i\right)}{\sum_{i=1}^{n} w_i \left(\sum_{i=1}^{n} w_i x_i^2\right) - \left(\sum_{i=1}^{n} w_i x_i\right)^2}$$

$$where \;\; w_i = \sigma^2 w_i' = \frac{1}{g^2(x_i)}$$

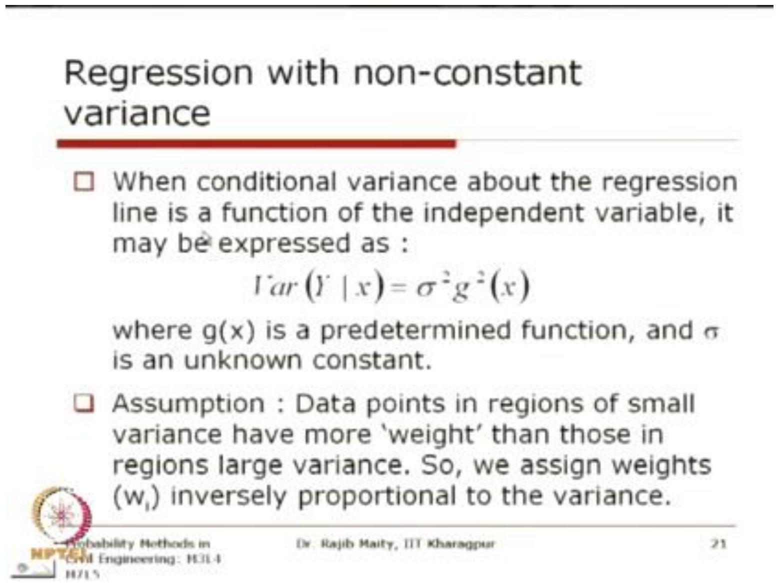
(Refer Slide Time: 46:02)



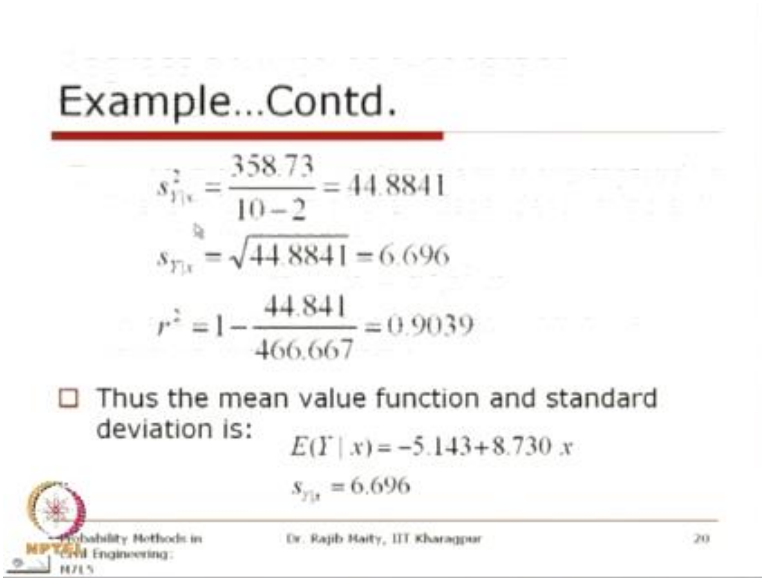
(Refer Slide Time: 46:03)



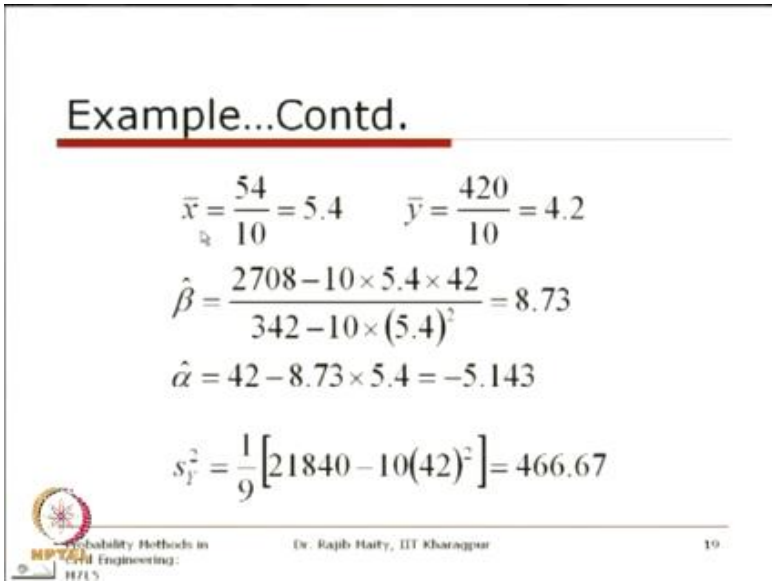
(Refer Slide Time: 46:04)



(Refer Slide Time: 46:04)



(Refer Slide Time: 46:06)

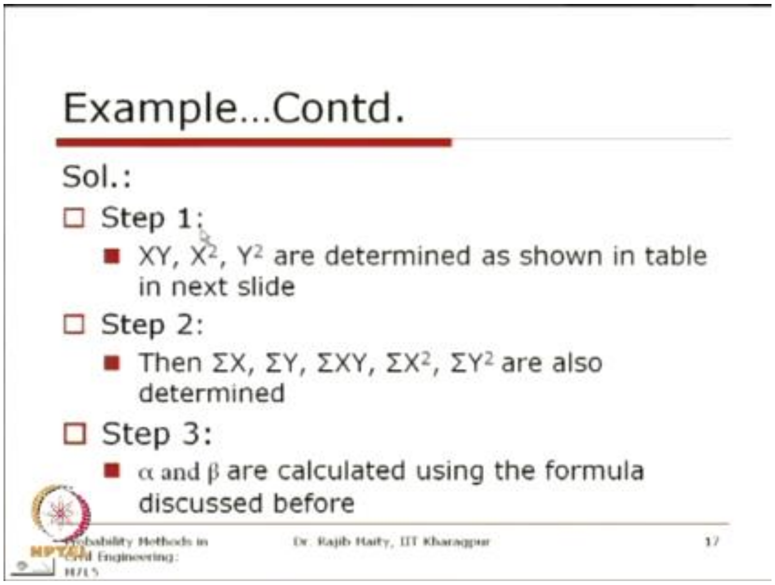


(Refer Slide Time: 46:07)

Example...Contd.

No:	x_i	y_i	x_iy_i	x_i^2	y_i^2	$\hat{y}_i=\alpha+\beta x_i$	$(y_i-\hat{y}_i)^2$
1	2	12	24	4	144	-0.515	156.624
2	3	25	75	9	625	-0.224	636.245
3	3	24	72	9	576	-0.224	586.797
4	4	26	104	16	676	0.067	672.513
5	5	40	200	25	1600	0.358	1571.472
6	6	38	228	36	1444	0.649	1395.078
7	7	45	315	49	2025	0.940	1941.257
8	7	65	455	49	4225	0.940	4103.645
9	8	70	560	64	4900	1.231	4729.127
10	9	75	675	81	5625	1.522	5398.957
Σ	54	420	2708	342	21840		21191.716

(Refer Slide Time: 46:08)



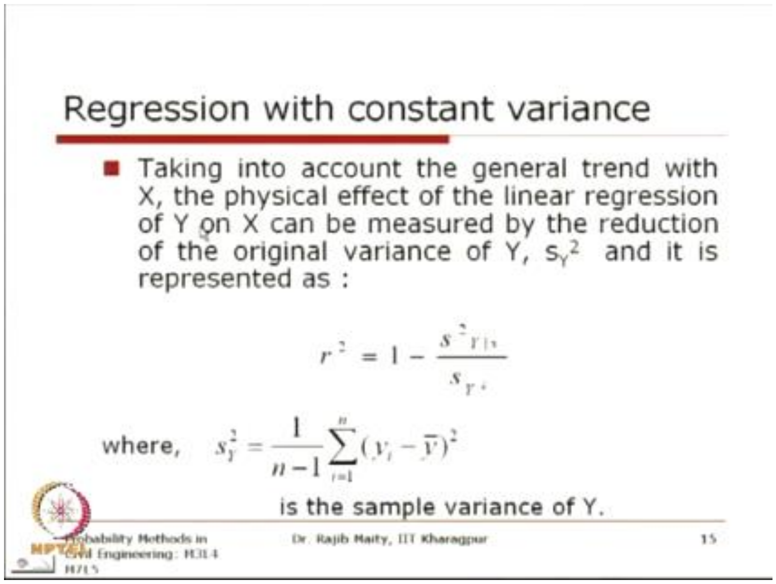
(Refer Slide Time: 46:08)

Example

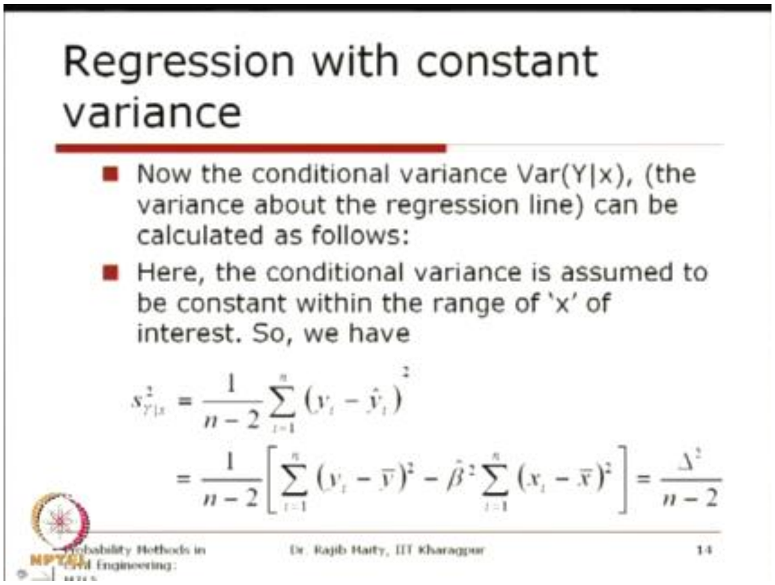
Q. Given set of data where the shear strength in kPa obtained from sample taken from 10 different depths of clay stratum. Assume the variance is constant with depth and determine the mean and variance of the shear strength as a linear function of depth.

Depth x_i	Strength(kPa) y_i	Depth	Strength(kPa)
2	12	6	38
3	25	7	45
3	24	7	65
4	26	8	70
5	40	9	75

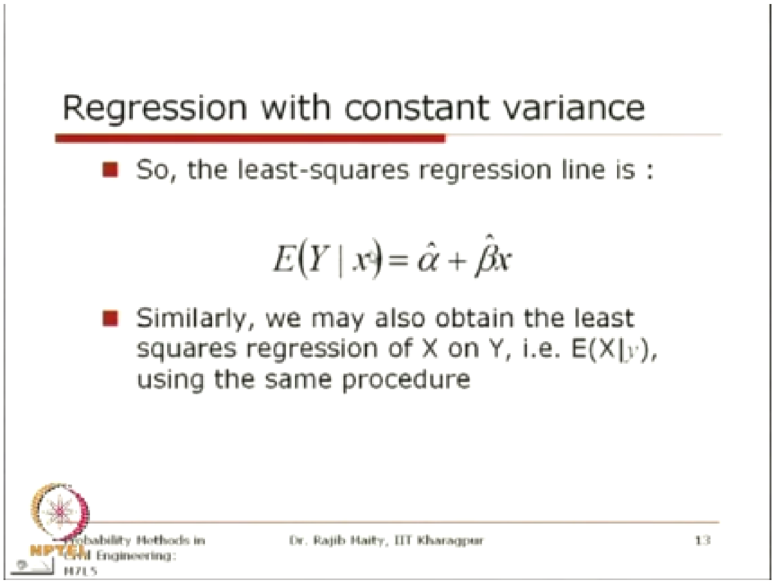
(Refer Slide Time: 46:09)



(Refer Slide Time: 46:09)



(Refer Slide Time: 46:10)



(Refer Slide Time: 46:10)

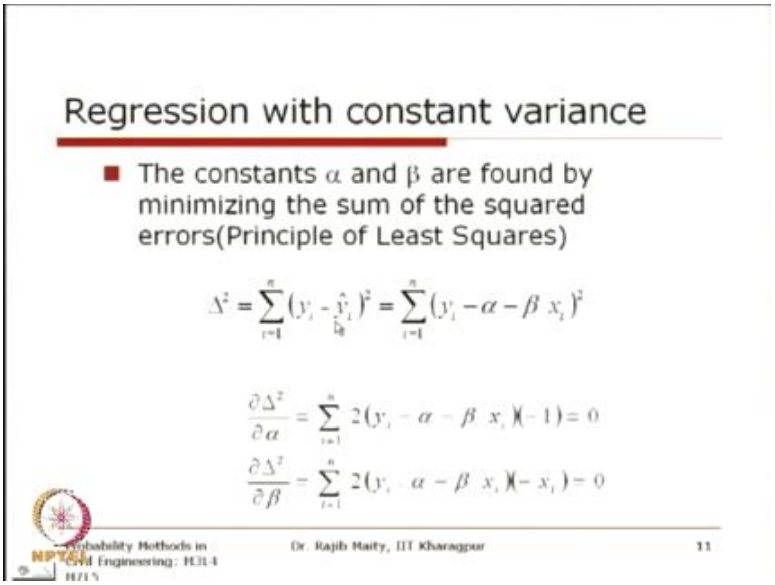
Regression with constant variance

- So, we obtain the least-squares estimates of α and β as :

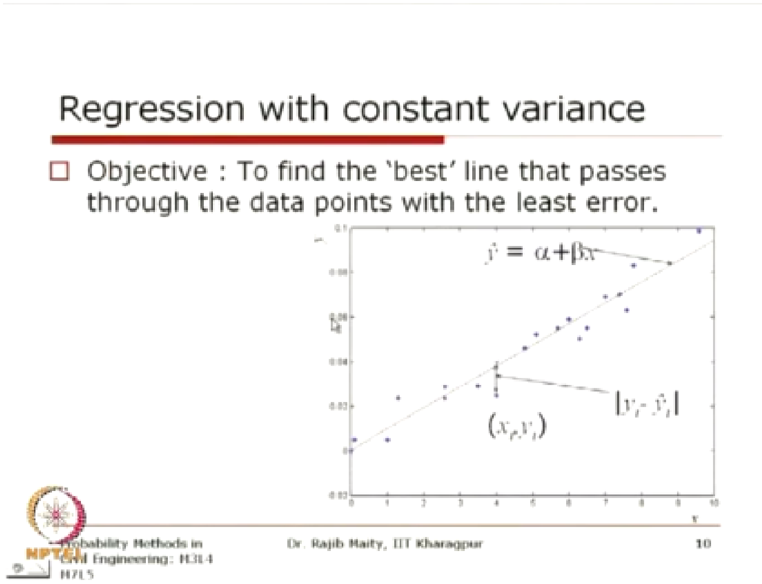
$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{\hat{\beta}}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}\,\bar{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^{2} - n\bar{x}^{2}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(Refer Slide Time: 46:11)



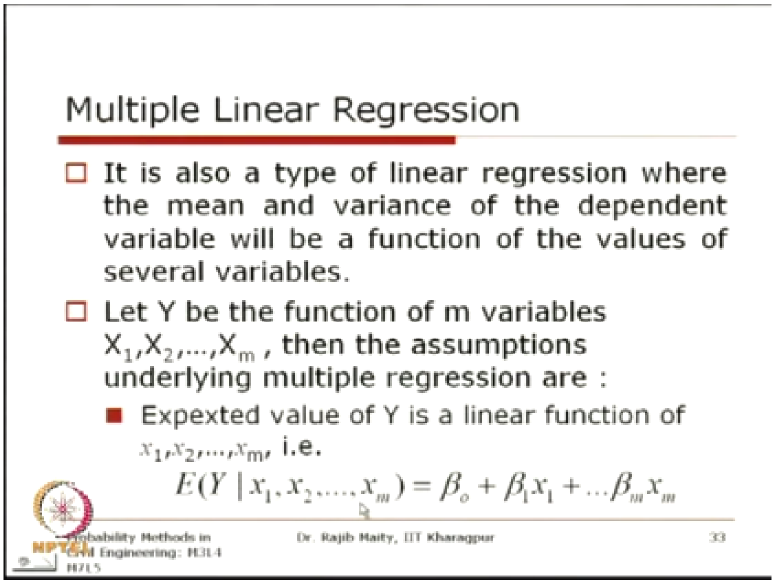
(Refer Slide Time: 46:11)



So, where only one input was there I might have sometime mentioned that this x is in this expression when we are regressing y on x, I might have sometime mentioned that this x is your dependent variable and y is the target variable. So, the correction will be that x is your independent variable and y is your dependent variable.

Sometimes, for the y we can mention that this is the target variable, response variable, dependent variable and all and basically, when we are referring to this o x this is the independent variable. Earlier in this case, when we were discussing the simple regression that time, only one dependent variable was there. Now, what we are discussing here is that, we are having more than one dependent variable.

(Refer Slide Time: 47:00)

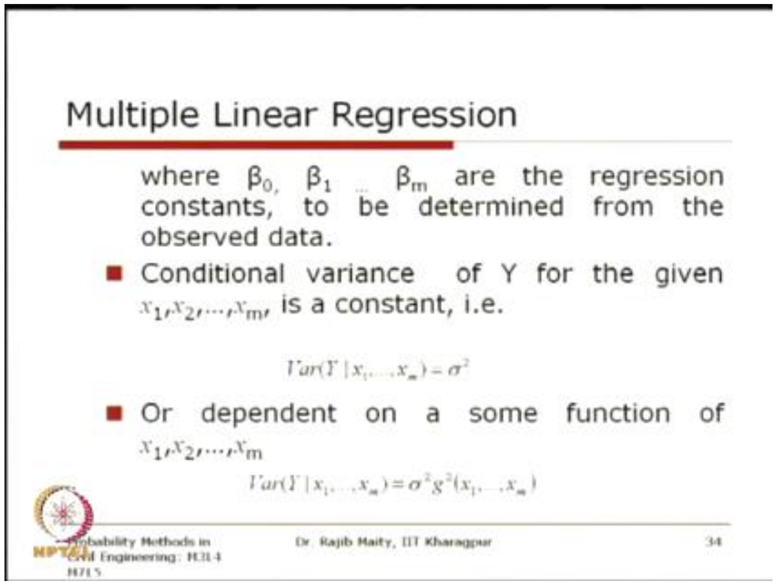


To model that dependent variable y and this is through a linear function which is that βo + β 1x 1+ up to this; up to the β m x m. now, basically how the concept is taken through is that. Now, for that when you see that there is only one independent variable and one dependent variable between x and y, we are basically fitting a straight line through the observed data point.

Now, when we are having more than one input say for example, if I just say there are 2 inputs x 1and x 2and our target – our dependent variable is y then basically, you can visualize, you can conceptualize in this way that this is a 3 dimensional space over which the two axis is one for the x 1, other for the x 2and we are basically fitting one surface; one straight, one linear surface through the data point in the 3 dimensional space.

So, this is in case of when there are two independent variables and one dependent variable y. now, similarly you can extend it to the higher dimension and this; so that, for the m independent variable, the relationship is generally that b naught plus b 1x 1plus b up to the b m xm. Now, we will follow the similar procedure to estimate these parameters as well. That is, we have to first find out what is the error and that error should be squared up; sum them. So, there is the sum of square error then it is minimized with respect to the parameters to get those expressions.

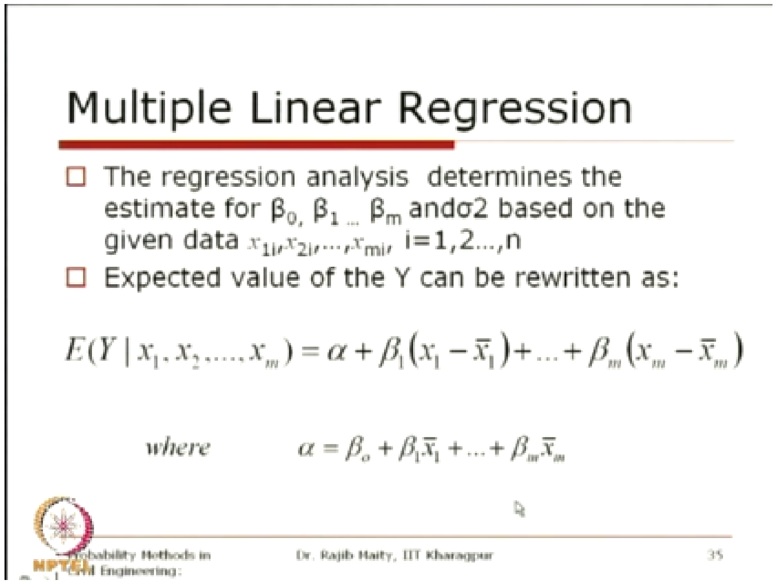
(Refer Slide Time: 48:44)



So, where this βo, β1, βm are the regression constant to be determined from the observed data and the conditional variance of Y for the given x_1, $x_2$up to x m is a constant that is variance of this Y given this input is equal to σ^2. Or, this is in case of when it is constant or it may be dependent on some function of this x_1, x_2, x_m. So, when it is dependent, when it is varying when it is non constant as we have used in the simple regression case.

So, this variance of y given $x_1 x_2 x_m$ is equals to σ^2.multiplied by the square of some function of this x_1, x_2 up to x_m.

(Refer Slide Time: 49:32)



Now that expression that is the regression analyses determines the estimates for this βo, β1, βm and the σ^2, sorry, this will be square and the σ^2based on the given data x $_1$ I, x $_2$ I, x $_3$ I, up to x $_{mI}$ and i is varying from the 12n. So, we are having then n set of I can say that n set of data that is y 1, x $_{1,}$ x $_2$,x $_3$,x $_m$. Similarly, I will have another set of this data. So, there are, m is the number for the number of the dependent variable and n is the number of what? How many sets of the observed data that is available to us.

So, based on this we can, whatever the expression that we have seen in this expression can be slightly modified as this one. That is, the α + βx $_1$- x $_1$/ + up to this that βm x $_m$ - x m bar. So, how we get this one is that, this x bar is the mean of whatever we have seen in this x, in the variable of $x_1$and the x $_m$ bar is the mean of that observed data of the, for the independent variable x_m. So, basically what we are replacing is that this constant beta naught is basically a adjusted is basically replaced.

So, this alpha you can see that this alpha is equals to that βo+ β_1 x_1/β_2 x 1bar plus β $_2$x 2/ + up to this β m x m bar. So, here what we can get is that, from this expression we can estimate that is α, β_1, β2, β m and from that estimate of this α and obviously, β_1, β2 the same. If we put it here, we will get what is the estimate for this beta naught and this one we will see. We will continue from this point onwards in our next lecture and what is in this linear regression part, what we have seen in today's lecture is the linear regression and linear regression with respect to the constant variance and some or the non constant variance.

We have seen one example for each case and this. So, in the next class what we will see is that multiple linear regression and then we will also see the non-linear regression and we will see the co relation as a measure of how strong the relationship is captured; that we will see in the next class; thank you.

Probability Methods in Civil Engineering

End of Lecture 39

Next: "Regression Analyses and Correlation (Contd.)" in Lecture 40

NPTEL Video Recording Team
NPTEL Web Editing Team
Technical Superintendents
Computer Technicians

A IIT Kharagpur Production
www.nptel.iitm.ac.in