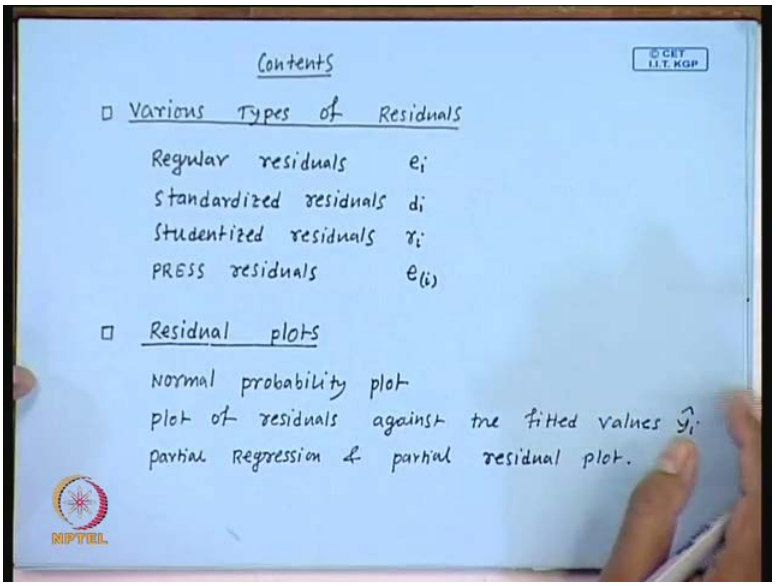
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian institute of Technology, Kharagpur

Lecture - 18
Model Adequacy Checking (Contd.)

(Refer Slide Time: 00:28)

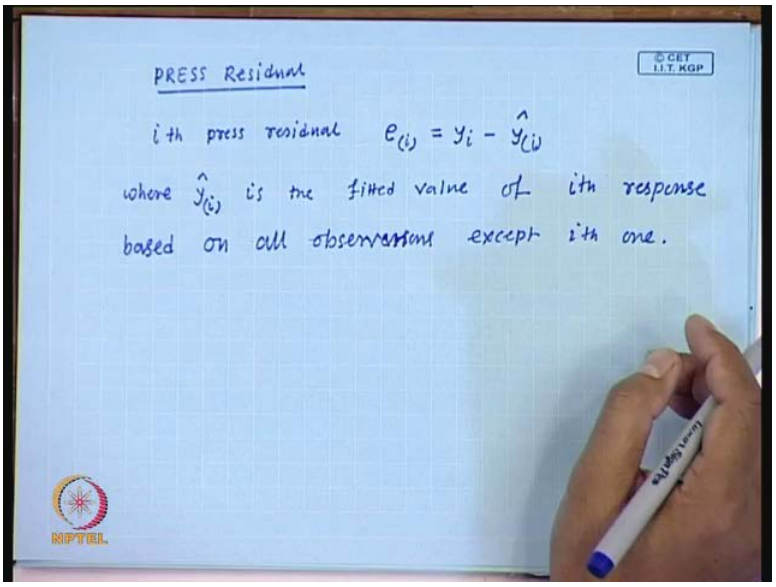


This is my second lecture in model 5, that is model adequacy checking and here is the content of this model various type of residuals. And in the previous class we talked about regular residuals, standardized residuals, studentized residuals. And today we will be taking about you know the press residuals, and also next we will be talking about several types of residual plots. Like, you know normal probability plot, plot of residuals against the fitted values y i hat and in the next class maybe we will be talking about partial regression and partial residual plot.

So, before I start talking about you know the press residual, just I want to repeat once more the objective of this model here, you know if you can recall you know simple linear regression model or in the multiple linear equation model. We have assumed that the erect on the epsilon has 0 mean, and erectum epsilon has a constant variations, and the erect terms are co related and they are normally distributed. So, what we are doing going to do in this model is that, we will present you know several methods to check the underlining assumptions that we made on the erectomy epsilon.

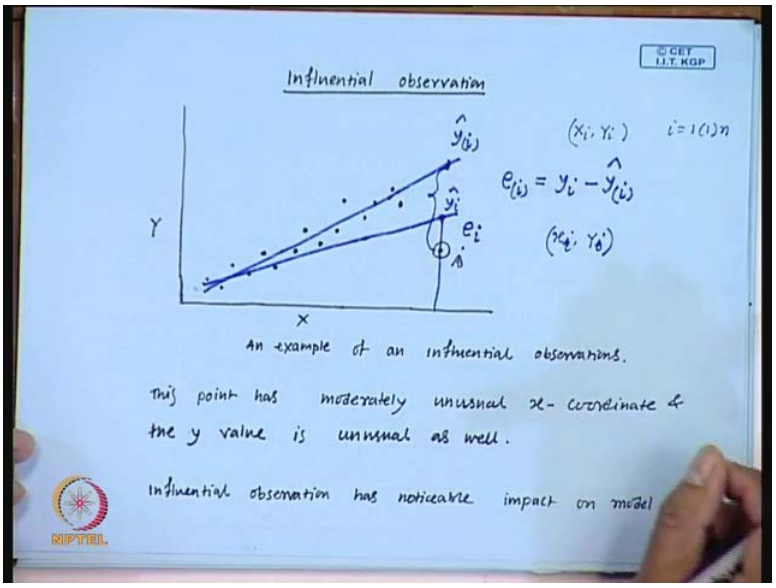
And the methods you know they mostly depend on the primarily depend on the study of residuals, because we think it is convenient to think that the residuals are the realized or observed value of the epsilon. Since, you are going to test some assumption on the erectom epsilon, so the test for that is is based on the residuals, and the lafical analysis the of the residuals are very effective to test the underlining assumptions on epsilon.

(Refer Slide Time: 03:53)



So, now I will be taking about press residual, which is one scaled residual, so I have already talked about regular residual and the other things, today I will be talking about press residual. So, it is beautiful concept here, well the ith press residual denoted by e bracket i is equal to y i minus y i hat, so we know that y i is the ith observed value, and what is y i hat where y i hat bracket i is the fitted value of ith response based on all observations except ith one. So, the basic logic behind this that, so this is not the regular residual this is the ith observation the response, value of the response variable and this is the fitted value of the ith response based on all observations except the ith observation.

(Refer Slide Time: 06:29)



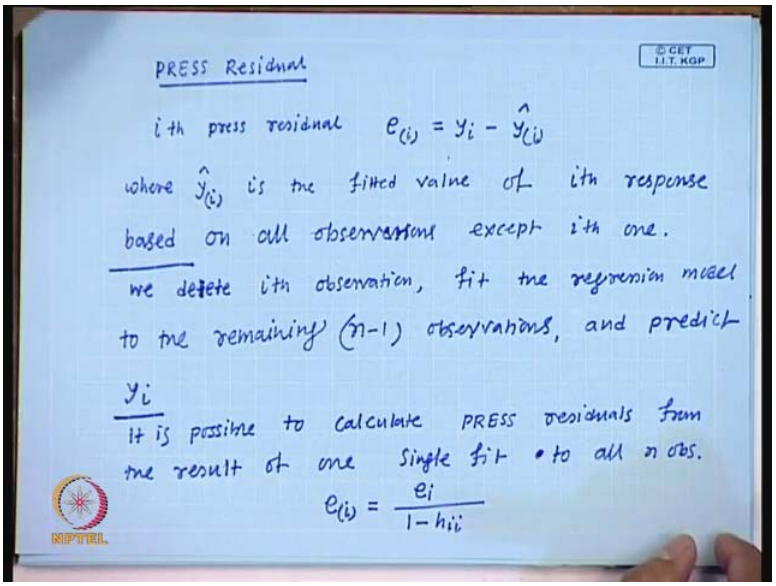
Let me just give idea about this special type of press residual, I will just refer the previous example, this is an example of influential observation. Now, if I fit a model you know based on the observation share, including this influential observation also y fitted model may get influenced by this influential observation. So, my fitted model may look like this, so this is the fitted model, because of this influential observation, and you can see you know the fitted model has been influenced by this influential observation.

And the fitted value here I mean this is the model based on all the observations for the fitted value here is y i hat, so this is my y i hat, and this is I am considering this as you know the ith observation. Then the residual I mean the true value of the ith observation, that is true value y i and the y i hat they are the difference is less. So, the difference is this much, and this is the ith regular residual e i, now if this ith observation is deleted from this data set, suppose this observation is not there in the model in the set of observations.

Then my fitted model will look like this and this model this fitted model is of course, it is not influenced by this influential observation, because I am fitting this model based on all observations except this ith observation. So, this is what I am talking this value you know the this value is y i hat bracket, so this is what I mean you know the fitted value of the ith response variable based on all the observation except the ith observation.

Now, my residual the reference residual which I call here e i bracket is this difference, so e i bracket is equal to y i, so this is y i and this is my y i hat bracket here. So, y i minus y i bracket hat, and you can see that the ith based residual is substantially larger than the regular e i.

(Refer Slide Time: 10:42)



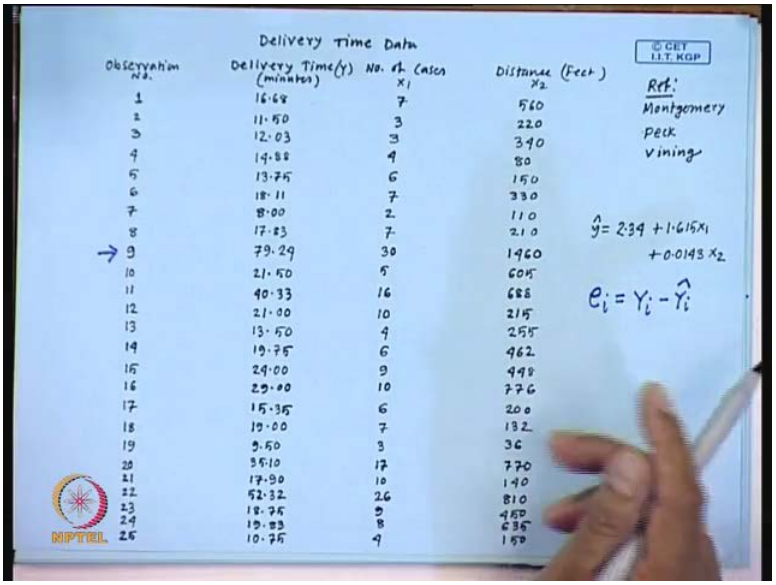
Well, so what we do is that, we delete we delete ith observation, fit the regression model to the remaining n minus 1 observations and predict y i. So, here it may appear that you know to compute the first press residual that is e 1 in bracket. you need to fit model based on all observations except the first observation. So, that is how you get e 1 bracket that is a first press residual, again you know you do not know which one is the influential observation, so you have to repeat this process in time.

So, to get e 2 bracket I mean the second press residual again you have to fit a model to the based on all the observations except the second observations; that means, you know you need to fit model based on n minus 1 observations. That you have to repeat in times, so but what are we going to do, it can be prove that you do not need to repeat this process n times to get the n press residuals, it can be done you know based on one regression fit based on all the observations.

So, here is the technique; however, so it says that it is possible to calculate press residual from the result of one single fit to all n observations, it can be proved that you know e i, i th press residual is equal to al e i by 1 minus h i i. Just recall that you know this e i is the

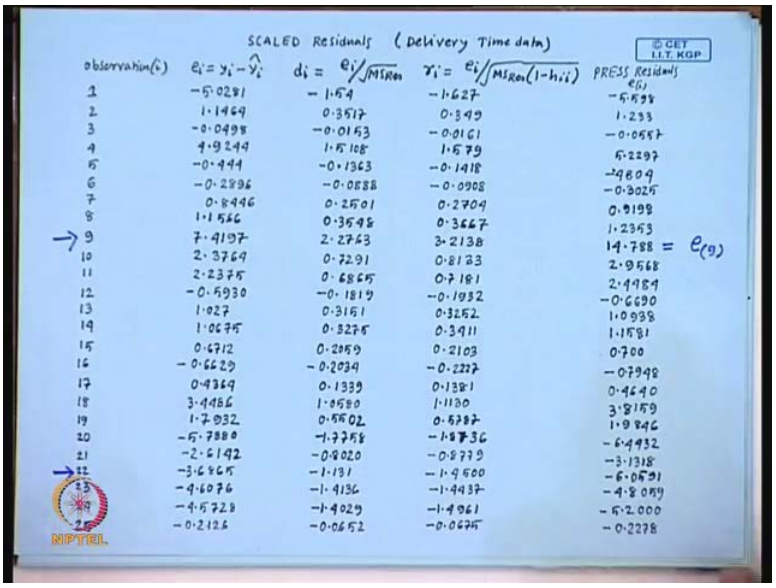
regular residual and h i i is the ith diagonal element of the hack matrices, and this is how you know this quantity they are same, so this is how we calculate the press residuals e 1, e 2, up to e n.

(Refer Slide Time: 15:25)



Now, let me just recall the example, we considered in the last class, so here is an example of multiple linear regression with two regressors x 1 and x 2, and here is this response variable Y. And we suspect that I mean it is likely that the ninth observation is an influential observation or at least it is an average point, because x 1 coordinate is much larger compare compared to the centre of x 1 and similarly x 2 value is much larger.

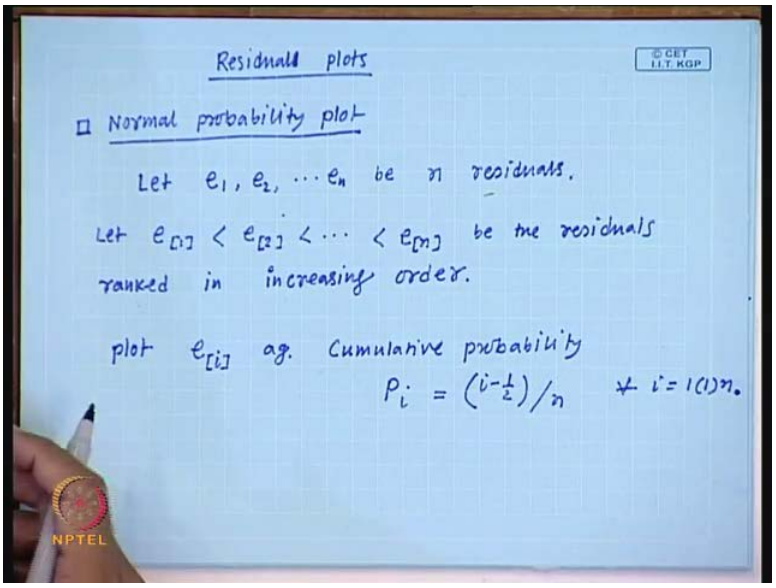
(Refer Slide Time: 16:09)



And, here is the regular residual, this is standardized residual, this is the studentized residual, and this is the press residual. Now, here we check that for the ninth observation the value of the regular residual is 7.41, where as the value the ninth press residual, so this is nothing but this is equal to e 9 press residual, this is 14.788 whereas, the regular residual is 7.4. So, for the point for the ninth observation the press residual value is sustainably larger than the regular residual, now also we can observe that for the 20 second observation, the value of the regular residual is 3.6 whereas, the value of the press residual is minus 6.05.

So, here also the value of the press residual is substantially larger than the value of the regular residual for 22nd observation, so this is what regarding the press residual it appears that, if the ith observation is an influential observation or if it is leverage point. Then there will be substantial difference between the regular residual value and the press residual value for that particular observation. So, next we move to residual plots, so there are several residual plots, so we will be talking about the first one is normal probability plot.

(Refer Slide Time: 18:56)

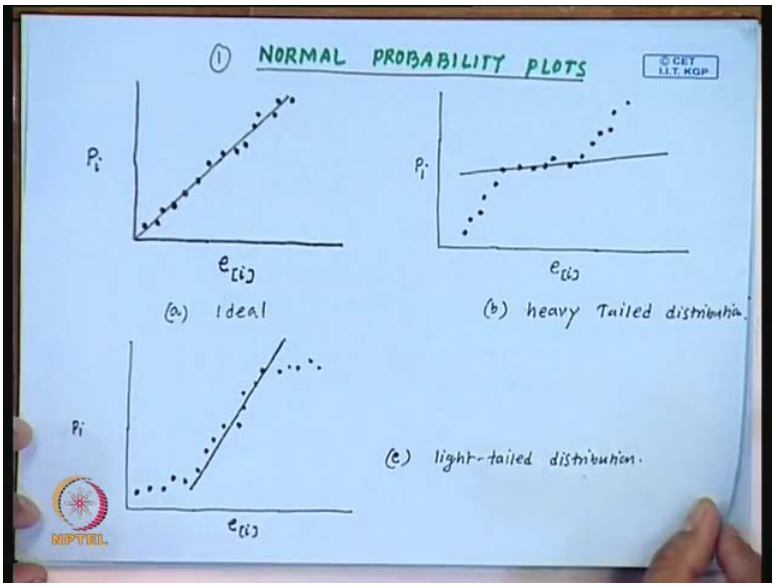


So, residual plots, first we will be talking about normal probability plot, so before I talk about this residual plot you know, I just want to mention again that the objective of this module is to check the underlining assumption on epsilon on the a rectum. So, we will be talking about different methods to check the underlining assumptions, so this is one of them. So, what this normal probability plot does is that it basically, check the assumption is that the epsilon i follows normal 0 sigma square, and they are independent and identical distributed.

So, this normal probability plot what it does is it checks whether the erectum really follow normal distribution or not, so here is the technique to test that. So, what it does is that let e 1, e 2, e n be n residuals; that means, the regular residuals, let e box 1, e box 2, e box n be the residuals ranked in ranked in increasing order. So, given a set of data you know you can fit the model, whether it is simple linear regression or multiple linear regression.

And then you can get the residual values, you just rank them I mean you arrange them in increasing order. Then what this normal probability plot does is that it plots e i against the cumulative probability p i, which is equal to i minus half by n and you do it for all i, i is from 1 to n. So, very simple technique you get the residuals first you arrange them in increasing order, and then you plot e i against p i here is the p i, i minus half by n.

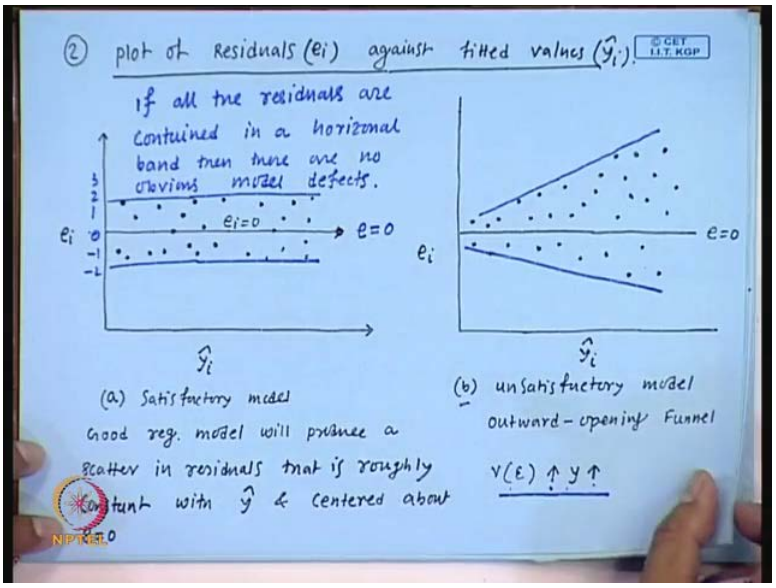
(Refer Slide Time: 23:36)



So, just now I have different types of normal probability plots, that may happen this is, so figure one shows the ideal situation, here you can see that all the points lie on straight line. So, if in your case, if you find that your plot I mean all the points are on the on a straight line, then you can assume that the error distribution is normal. Now, if here in the figure b, you can see that it is not really I mean the points are not on a single straight line, and here it this type of situation occurs.

Then we can say that the distribution is heavily tailed, that is you know it is really not true that the erect distribution is normal. And figure c indicates that the distribution of the erectum is lightly tailed. So, there is little deviation from the I mean it is not reasonable to assume, if such situation occur than it is not advice able to assume that the erect distribution, is normal. So, this is what regarding the residual plot, so residual plot basically, if test the normality assumption of the erectums epsilons. Next, we will be talking about one more plot, that is it is a plot of residual term against the fitted observation.

(Refer Slide Time: 26:26)



So, here is the plot of residual e i against the fitted value y i hat, so given a set from data you know, you fit the model first and then you can get the fitted value. Once, you have the fitted value you can compute the residual, and then you plot the residual against the fitted value. Well now, we need to understand I mean of if the plot if the pattern looks like this what does it indicated, so if the plot of residual against the fitted value looks like this one.

Then you can conclude that the this is a good regression model, and the good regression model will produce scatter in residuals that is roughly constant with y hat and centered about e equal to 0. Then I mean what I am trying to say here is that if all the residuals are contained in a horizontal band. So, here is the horizontal band, here you can see that all the points are contained inside this horizontal band, then there are no obvious model defects.

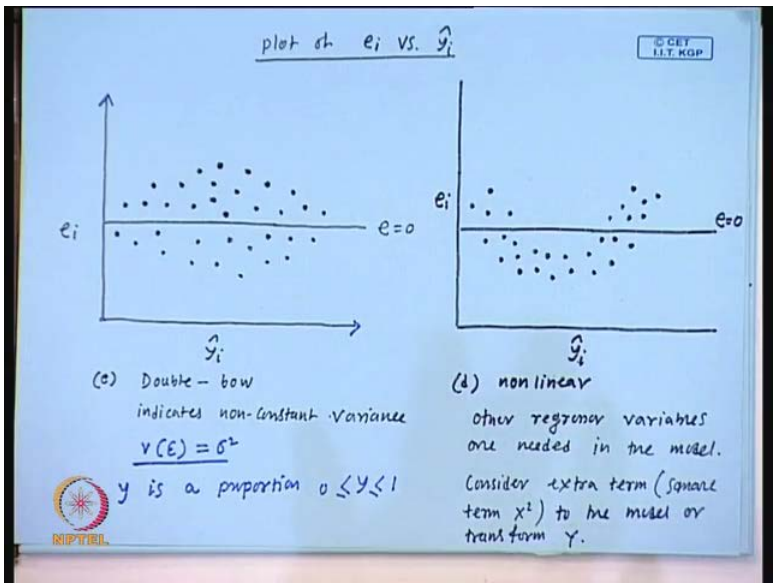
That means, if what I want to say here is that you know, if all the residuals are contained in a horizontal band centered about e equal to 0. This is the line e equal to 0, because residual value could be this is 0 minus 1 minus 2 1 2 3, like that some of the residuals are positive some are negative you have observed that before also. So, what I want to say is that if the residuals are contained in a horizontal band, then there is no obvious defect with the model; that means, it is a good fit.

So, this is the only situation we say that, the fitted model is satisfactory, so there is no problem with the fitted model. Now, look at the second case this is this figure b indicates, and how to outward opening funnel pattern, so here you can see that this e i values increase, I mean e i increases with the value with y i hat. So, this is called outward open funnel pattern, what does this indicates that the indicates non constant variants of it is, we assume the see we have assumed that the variants of the epsilon is constant that is sigma square.

But, if your residual I mean if your plot, which plot I mean this plot particular like epsilon verses epsilon sorry, e i against y i hat. If the pattern is like you know outward open funnel, that indicates that variants of epsilon i or the variants of epsilon increases as a y increases. So, this one is sort of this is the indication of a non constant variants, so it cannot if this occurs you know it is not advisable to assume that the variants of epsilon is sigma square.

So, next we talk about some more pattern, the next pattern is one more thing just I forgot to mention here, instead of outward open funnel heat could be inward open funnel also. That means, it will look like this in that case also indicates that, non constant variants of epsilon, in that case variants of epsilon decreases as y increases. So, both the I mean it could be outward open funnel, it could be inward open funnel, so in both the case indicate the non constant variants of epsilon.

(Refer Slide Time: 33:30)

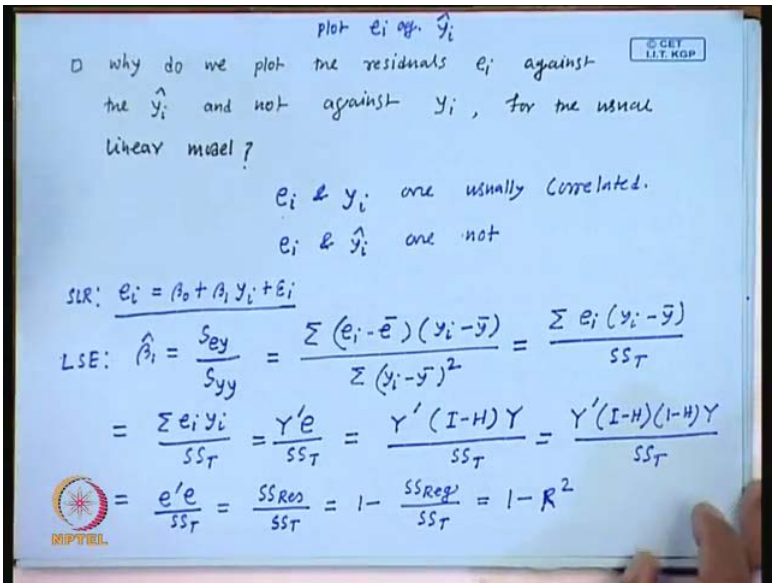


So, next we will be talking about one more I mean some more pattern, like a this is one more pattern here this is the line e equal to 0 y i hat, in this direction, and this is called double bow. And this one also indicates non constant variants; that means, variants of epsilon, we cannot assume that this is equal to sigma square, so this pattern also violet this assumption. Well, and this type of pattern often occurs when y is a proportion, response variable is a proportion and y lies between 0 to 1, also this is in this type of situations we often get double bow pattern.

Now, the last one is that you know the figure d shows the non-linear pattern here, this non-linear pattern indicates that you know other regressor variables are needed in the model. So, this indicates that the relationship between y and the regressor variable is not linear, we need to introduce some non-linear term; that means, consider extra term like square term x square.

I mean the relation is not just the linear relation like y equal to beta naught plus beta 1 x you need to introduce some square term the higher order term, or you terms you take a transformation of the response variable y. I mean maybe we will talk regarding this issues later on again, but what this non-linear pattern indicates that the relationship between the response variable in the regressor variable is not linear. We need to introduce some other terms like, other regressors. For example, we need to take x square x q in the model, or you may need to take a transformation of the response variable y like log y or something 1 by y something like that.

(Refer Slide Time: 36:43)



Now, there is a question you may you might be wondering you know why do, what we did is that we have just checked about another plot of e i against y i hat, so why it is you know why plotting e i against y i hat, why not e i against y i. So, this is the question why do we plot the residual e i against y i hat and not against y i for the usual liner model, it is not, so easy to answer this question like you can think of it, e i y i hat. The answer to this question is that you know e i and y i we do not constraint the plot e i against y i, because this two are usually correlated.

What I mean by this I will talk about that whereas, e i and y i hat are not correlated, there is no relationship between e i and y i hat, what i going to prove is that there is a linear relationship between e i and y i. Suppose, the relationship is the form of the relation is e i I said linear relationships e i equal to some beta naught plus beta 1 y i plus epsilon i am just a simple linear relation between the residual and the observed value y i.

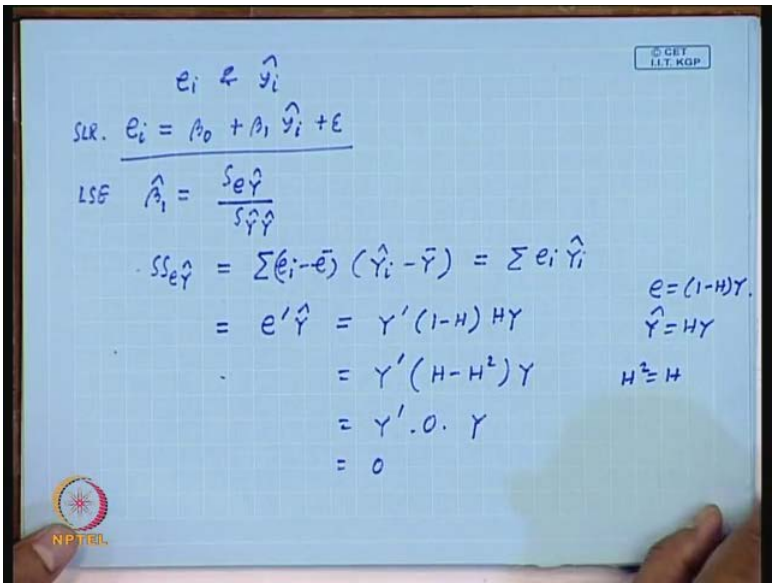
Now, if you we know the same technique this is this is nothing but the simple linear regression model between the residual and y i, you can check that beta 1 hat is nothing but S e y by S y y. So, what is e s y this is the list square estimate you can check with my first model beta 1 hat is nothing but this one this is called the list square estimate. So, this one is nothing but notation only this is nothing but e i minus e bar into y i minus y bar by submission y i minus y bar whole square.

So, what am I trying to prove that there is a relationship between y i and e i of and I am trying to find out that relationship, what type of relationship they have, if it is linier then what is the value of the coefficient. So, this one is nothing but summation e i y i minus y bar, you can check that it is not difficult and this one is nothing but the S S T, and again you know y bar into e i that is sum over e i is going to be 0, so this is equal to summation e i y i by S S T very simple.

Now, in matrix rotation this can be written as e prime y or y prime e same thing by S S T, now what is e in terms of h rotation Y prime and e we know e is I minus H Y, by S S T. And we know what we know I minus since H is idem important matrix I minus H is also idem important matrix, so I minus H can be replaced by I minus H square. So, this is equal to Y prime I minus H into Y minus H, because I minus H into Y minus H equal to Y minus H, as I minus H is idem important matrix into Y by S S T well, so this is nothing but see this is e this is e prime this is e prime e by S S T.

So, e prime is e is nothing but S S residual this is nothing but S S residual by S S T which is nothing but 1 minus S S regression by S S T which is nothing but 1 minus r square. So, this r square is the you can recall r square is the coefficient of multiple determination well, so the relationship between, so that what we proved is that there is a linear relationship between the residual and the observed value. And the coefficient value I mean the slope is equal to 1 minus r square, well let me check with y i hat whether there is linear relationship between e i and y i hat, we can prove that there is that the slope is 0 in that case.

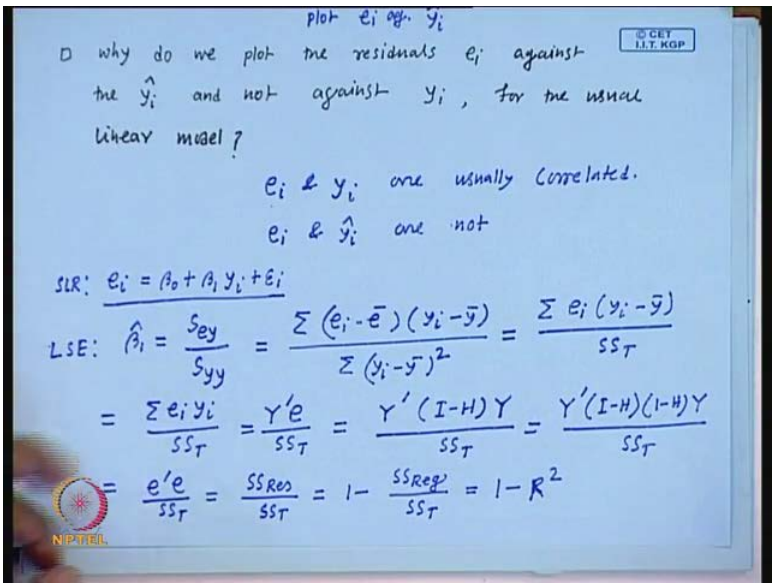
(Refer Slide Time: 43:58)



So, let me check with e i and y i hat, so e i suppose there is a linear relation between e i and y i hat and the relation is e i equal to beta naught plus beta 1 y i hat plus epsilon like that the simple linear regression. And then by list square estimate we can check that beta hat is equal to S e Y hat by S Y hat Y hat, I do not care about this denominator, let me proceed with S S e y hat.

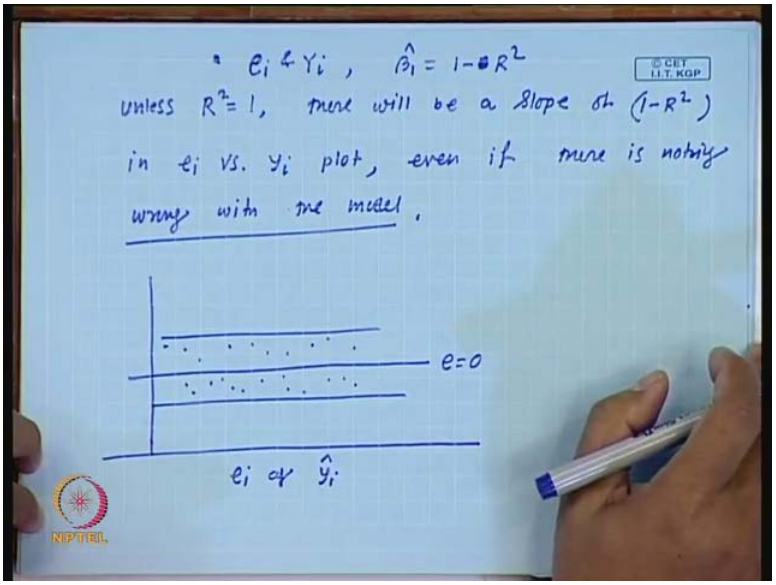
If I can prove that this is equal to 0, then my slope is going to be equal to 0; that means, there is no linear relationship between e i and y i hat. So, this is going to be equal to e i minus e bar into y i hat minus y bar, and you can check that this is nothing but summation e i y i hat. So, in matrix notation this is equal to e prime y hat, so what is e prime, e prime is equal to Y prime 1 minus H, because e equal to 1 minus H, Y and Y hat is equal to H Y. So, H Y here now Y prime H minus H square into Y, now see H is idem important matrix, so H square is going to be equal to H, so this is y prime 0 basically, y this is going to be equal to 0. So, there is no, so this proves that there is no linear relationship if between the residual and the fitted values.

(Refer Slide Time: 46:54)



Well so what I want to conclude here is that, unless in case of e i and y i we have observed that beta 1 hat is equal to 1 minus r square.

(Refer Slide Time: 47:04)



So, in the case of e i and y i beta 1 hat is equal to 1 minus r square, so unless square is equal to 1, there is a positive slope, there will be a slope of 1 minus r square, in e i verses y i plot. Even if there is nothing wrong, with the model, so if you can recall that in e i

verses in e i against y i hat plot, I said that if all the residuals are contained in a horizontal band, centered around e equal to 0.

Then the corresponding fitted model is perfect, but here since there is a theoretical relationship between the residual and y i are it is very likely that the residuals will not be contained within horizontal band centered at I mean e equal to 0. There will be slope of one minus r square when r square is not equal to 0, so that is why you know it is very difficult to conclude anything, if we plot the residual against y i instead of plotting the residual against y i hat. So, that is why the reason you know, and since there is no relationship between e i and y i hat, they are not correlated there is no linear relationship between them, that is why we prefer plotting the residual against y i hat, so that is all for today.

Thank you.