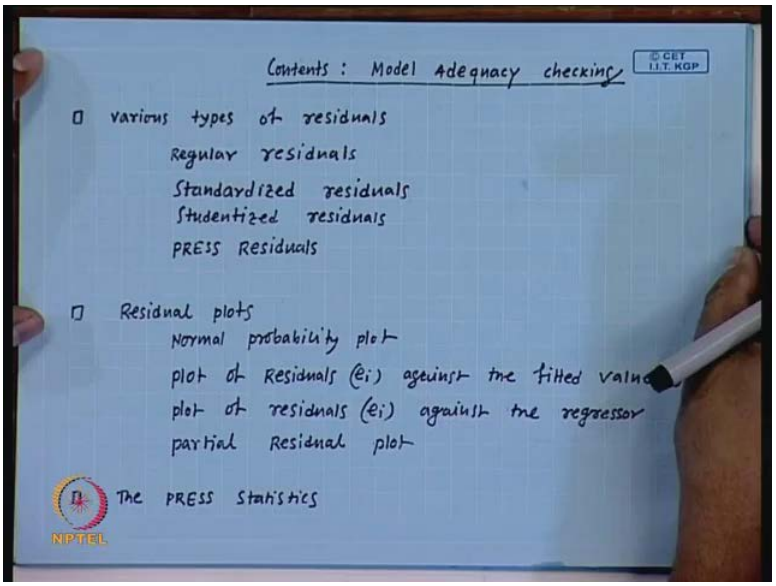
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

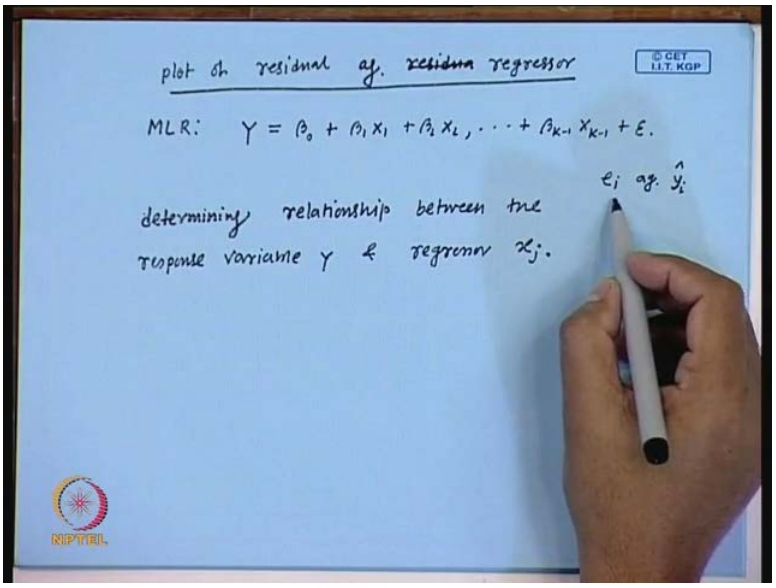
Lecture - 19
Model Adequacy Checking (Contd.)

(Refer Slide Time: 00:27)



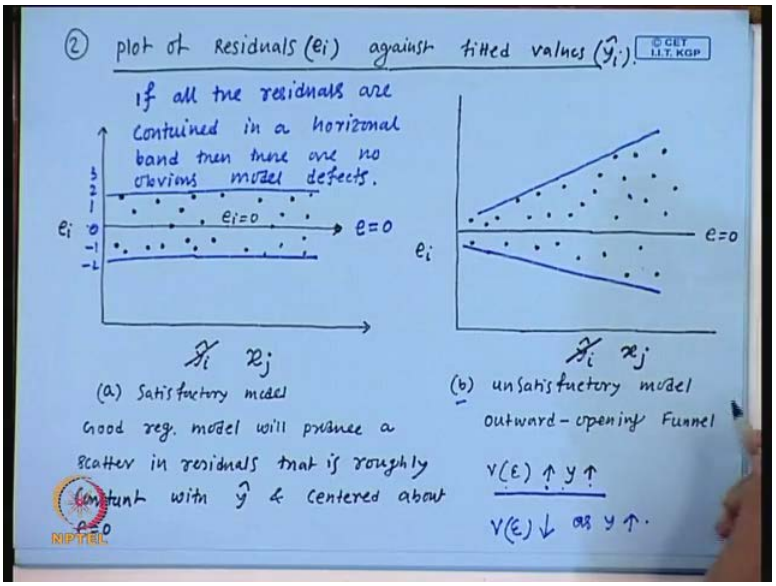
This is my third lecture in module 5 that is the Model Adequacy Checking. Here is the content of this module, we have already talked about various types of residuals like, regular residuals, standardized residual, studentized residual and press residuals. And also we have talked about two residual plots like, normal probability plot and plot of residual against the fitted value y i hat. So, what we are going to do today is that, we will be talking about plot of residual against the regressor and we will be talking about the partial residual plot. And finally we will be talking about the press statistics, so first we will be talking about the plot of residual against the regressor.

(Refer Slide Time: 01:43)



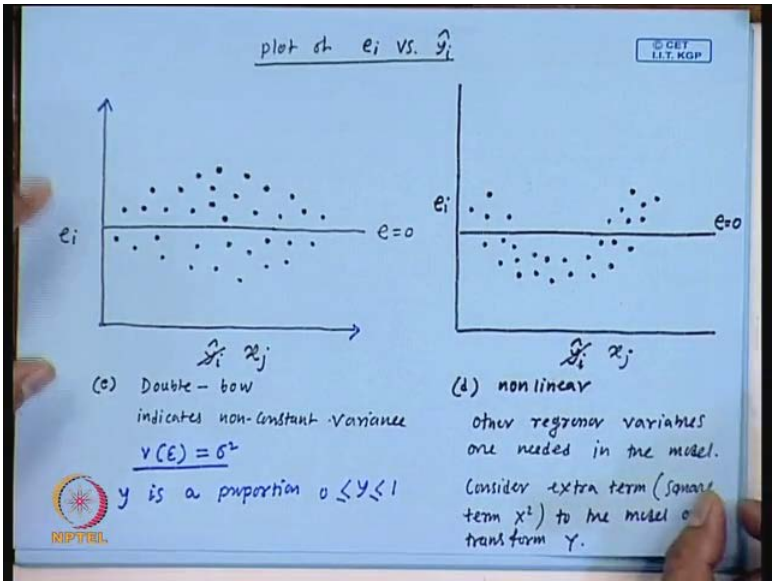
So, here we are in the multiple linear regression model setup, multiple linear regression and we have the response variable y and several regressor variables, like x 1 x 2 may be beta k minus 1 x k minus 1 plus epsilon. So, like residual against y i hat plot, here will be plotting residual against the regressor x 1 and then again residual against x 2 like that, and this particular plot, plot of residual against regressor is important in determining relationship between response variable y and regressor x j. And also this plotter, very similar to the plot of the pattern of the plots, here in the case of residual against regressor are very similar to the pattern of plot, in case of residual versus y i hat.

(Refer Slide Time: 04:49)



So, in this case also here, instead of y i hat we will plot some the regressor x j, and here also the residuals containing in a horizontal band is desirable, and this is treated as this sort of horizontal band containing all the residuals indicates a satisfactory model. And the open funnel like, whether it is similarly here also, instead of y i hat will be plotting in this horizontal axis x j. So, here also the outward open funnel indicates that the variance of, indicates non constant variance like whether it is outward open funnel or inward open funnel that indicates non constant variance.

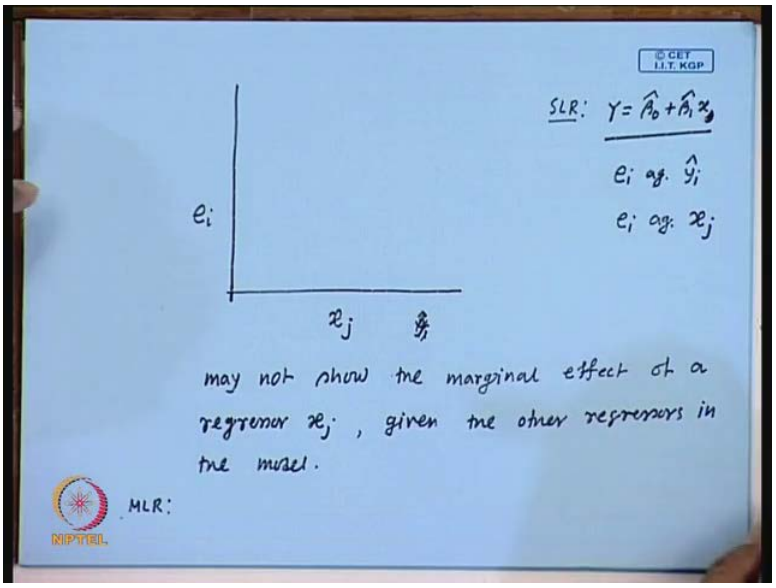
(Refer Slide Time: 06:30)



Similarly, in the case of double bow also for example, x j here and x j here, so all this things, this double bow also indicate non constant variance and this type of non-linear pattern indicates that the assumed relationship between y, the response variable y and x j is not correct. So, that means, we have to consider the extra terms like may be the higher order regressor like for example, x j square or x j q or some transformation on x j for example, may be 1 by x j or may be log x j, so this is what the non-linear curve indicates.

So, what we learned from this residual against the regressor plot, this is important to determine the relationship between the response variable and particular regressor x j. And the limitation of this plot is that, it may not show the marginal effect of x j, a particular regressor x j on the response variable, because what we are doing is that, we are plotting residual against a particular regressor and let me tell little bit more about this one.

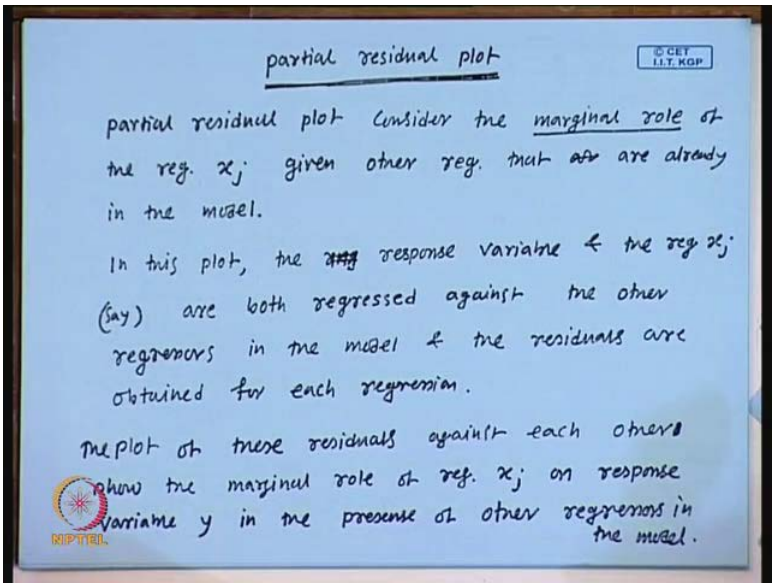
(Refer Slide Time: 08:50)



So, what we do here is that, we are plotting e i against a particular regressor x j, so what I told that the limitation of this plot is that, it may not show the marginal effect you need to understand this part, this plot it may not show the marginal effect of regressor x j, given the other regressor in the model. We are talking about multiple linear regression, if it is simple linear regression and there is only one regressor, so y equal to beta not plus beta 1 x, so in case of simple linear regression there is no difference, between the plot of e i and y i hat and the plot of e i against x j.

So, those two cases are same, and in case of simple linear regression, the plot of e i against y i hat is same as the plot of e I, the residual against x because there is only one regressor and they are linearly related. So, why whether you plot against x or y it does not it does not matter, but in case of multiple linear regression there is a difference, these two cases are not same, because there are, so many regressors more than two regressors are there. So, what I told here is that the limitation of this e i against x j plot is that, that my not show the marginal effect of x j, given the other regressors in the model. So, that is why we need to go for partial residual plot, so I will try to explain the partial residual plot now, this is an improvement of residual against regressor plot.

(Refer Slide Time: 12:25)



So, this is called partial residual plot, so what we do here is that, here in the partial residual plot consider the marginal role of the regressor x j, given other regressor that are already in the model. Let me just give the outline of this plot, what does this partial residual plot does and then I will try to explain the logic behind this plot. So, what it does is that, in this plot the response variable and the regressor, say x j are both regressed against the other regressor in the model.

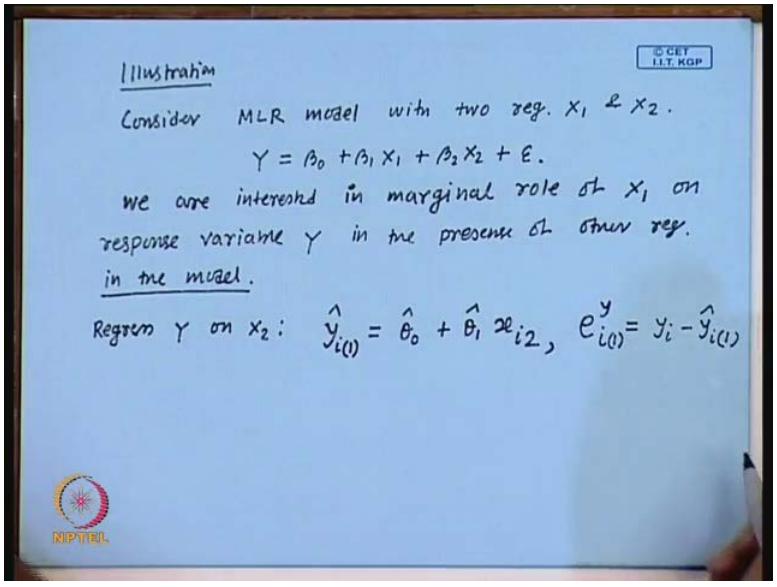
And the residuals are obtained for each regression, so what I told is that, this partial residual plot consider the marginal role of, what is this marginal role of the regressor x j, given the other regressor that are already in the model. So, since we are talking about multiple linear regression here, we need to consider that the marginal role of one regressor in the presence of other regressors in the model. So, the technique is that first both response variable y and particularly regressor x j, both are regressed against the other regressors in the model.

So, here the response variable y is regressed on all the regressors except x j and similarly, x j is also regressed on the remaining regressors and then the plot of, so we will get two residuals from these two regressions fit. And the plot of this residuals then we plot, so there are two residuals, there is two regression model we are trying to fit, one is we are trying to regress the responsive variable on all the regressors except x j, so this is one regression model.

And other one is that x j is regressed on the remaining regressors, so this is the second regression model, from this two once you have the fitted model for this two regression model, you will get the corresponding residual values. And then you plot those two residuals to get this marginal residual plot, the plot of these two residuals against each others. So, the marginal role of regressor x j on response variable y in the presence of other regressor in the model, this partial regression plot is little difficult concept, I will try my best to explain it.

Just now I will give one example, which consists of two regressors; that means, I will take an example of multiple linear regression models with two regressors, and I will try to explain the technique first. And then after that I will try to give the logic behind the idea behind the partial residual plot, I said that this partial residual plot considered the marginal role of x j. So, you need to understand what I mean by the marginal role of x j on the response variables, in the presence of other regressors in the model, so there are several regressors, I told. Let me just first give an example to illustrate the technique what I explained here, and then I will be talking about the logic behind this partial residual plot, so the example here let me illustrate the technique first.

(Refer Slide Time: 21:19)



So, consider multiple linear regression model with two regressors x 1 and x 2, so my model is of this form y equal to beta not plus beta 1 x 1 plus beta 2 x 2 plus epsilon. Suppose, for example, first I am interested to know the marginal role of x 1 on the

response variable y in the presence of x 2, suppose we are interested in marginal role of x 1 on response variable y, in the presence of other regressor in the model. So, according to the technique what I will do is that, there are two regressors x 1 and x 2 and I am interested to see the marginal role of x 1 on the response variable.

So, first what we will do is that, my aim is to eliminate the effect of x 2, see there are two regressors x 1 and x 2, I am interested to see the marginal role of x 1 on y, so what I will do is that, I will just first I will eliminate the effect of x 2 from the response variable y. And for that what I have to do is that, I have to fit a model between I will regress y on x 2 first, and from there the residual will give me the part of variability in y, which is not explained by x 2.

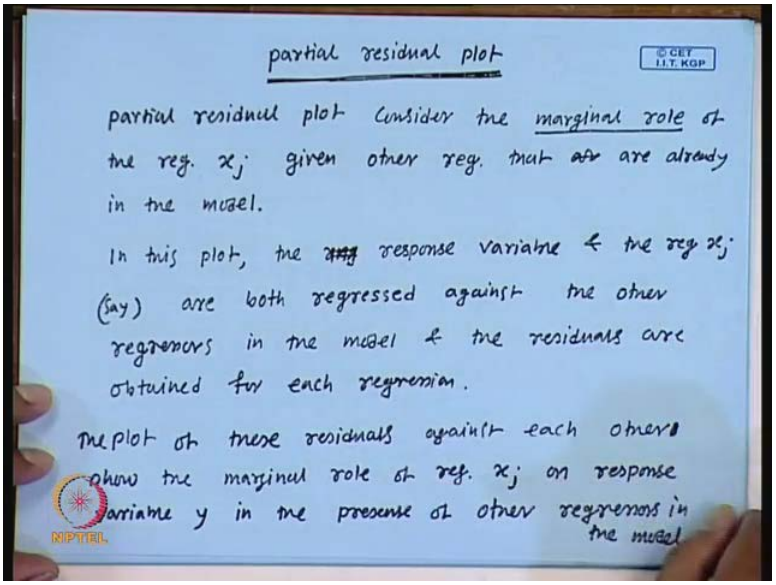
And then I expect x 1 to I want to see how much of that residual x 1 can explain, so this is the idea, I am interested in the marginal role of, I repeat I am interested in the marginal role of x 1 on the response variable y. So, first what I will do is that, I will eliminate the part of variability that can be explained by x 2; that means, I am just eliminating the effect of x 2 from y. So, that remaining part I want x 1 to explain, and I will see how much the remaining variability can be explained by x 1. So, that is what the basic idea behind this partial residual plot.

So, to eliminate the effect of x 2 from y or to note the part of variability, which cannot be, which is not been explained by x 2, first I need to regress y on x 2, and I will calculate the residual that is the residual part is the part of variability, that is not explained by the regressor x 2. So, first what I will do is that, I will regress y on x 2 that is same as saying that, in general in case of many regressors regress y on all regressors except x 1.

If you are interested in finding the marginal role of x 1, since here are here we have only two regressors I am regressing y on x 2 only, suppose my fitted model is y i hat equal to theta not hat plus theta 1 hat x i 2. So, this is the fitted model between the response variable and x 2, and I want to introduce one notation here, I will put bracket 1; that means, that response variable has been regressed on all the regressors except x, this is the meaning of this. So, the regressor variable is has been regressed on all the regressors except x 1.

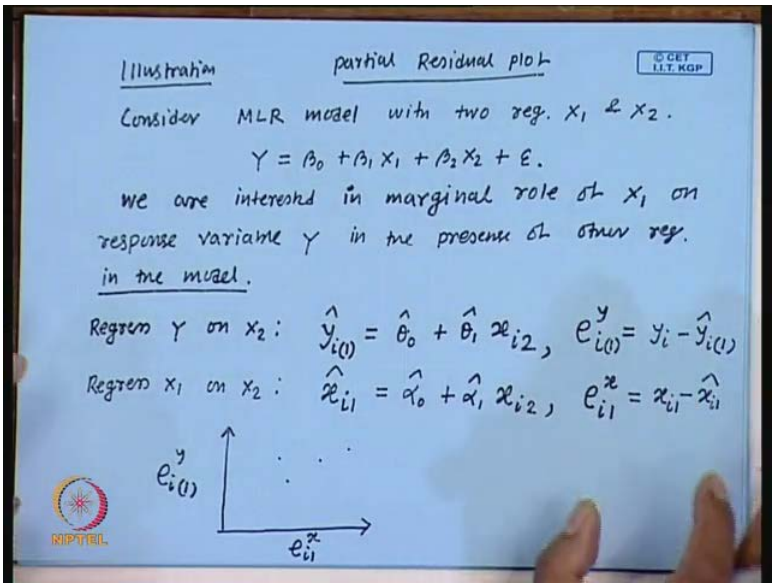
And once we have this fitted model, we can compute the part of variability in y, which has not been explained by the second regressor, so that is the residual e i is generally we write y i minus y i hat. But, here I will write y i minus this is the observed value and this is the fitted, when y is regressed on x 2, this fitted value is y i hat 1 and I will put 1 bracket one here. To denote that this is the residual and also I will put y here, little bit complicated notation, but this is the residual obtained from this regression.

(Refer Slide Time: 29:23)



And also according to my partial residual plot technique, it says that in this plot, the response variable and the regressor x j, suppose I am interested to find the marginal role of x j on the response variable. So, I will regress response variable and the regressor x j on the remaining regressors.

(Refer Slide Time: 29:44)



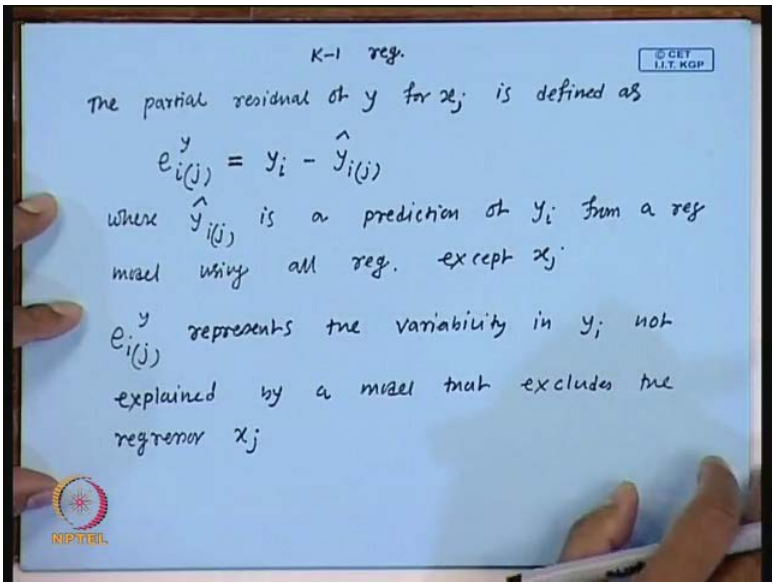
So, here also I will regress x 1 on x 2, so regress x 1 on x 2 and suppose my fitted model is x i 1 equal to alpha not hat, simple linear regression model alpha 1 hat and I am regressing on x 2, so x i 2 this is my fitted model. And so this basically see how much of the variability in x 1 can be explained by the other regressor, sometime they are not completely independent there might be little bit of dependence between them. So, we just want to eliminate the effect of or the contribution of other regressors, on the regressor we are interested on.

So, here the residual is e i 1 x, I will put here may be little complicated notation x i 1 is the original value of the first regressor, I am talking about the i th observation minus x i 1 hat. And then what this partial residual plot does is that, I am talking about partial residual plot does is that, it plots e i 1 y against e i 1 x, so this plot is called the partial residual plot whatever it might be. We will slowly establish the relation between this two, in case if we are assuming that, the relation is linear then what we can expect from the plot of this, what pattern we expect in the partial residual plot before that, so this is an example I hope you understood the technique at least of the partial residual plot.

So, it says that you first regress both suppose, if I am interested to find the marginal role of x j on the regressor y, then I will regress both y and x j on the remaining regressors. And so there will be two regressions model and you find the residuals corresponding to those two regression models, and plot them to get the partial residual plot. So, this is an

example, which consist of only two regressors, next I will just introduce the notations the same concept only, I will just introduce the notations for in general case, when there will be for example, k minus 1 regressor instead of only two regressors, so everything and just some more notation here, in case of say for example.

(Refer Slide Time: 34:14)

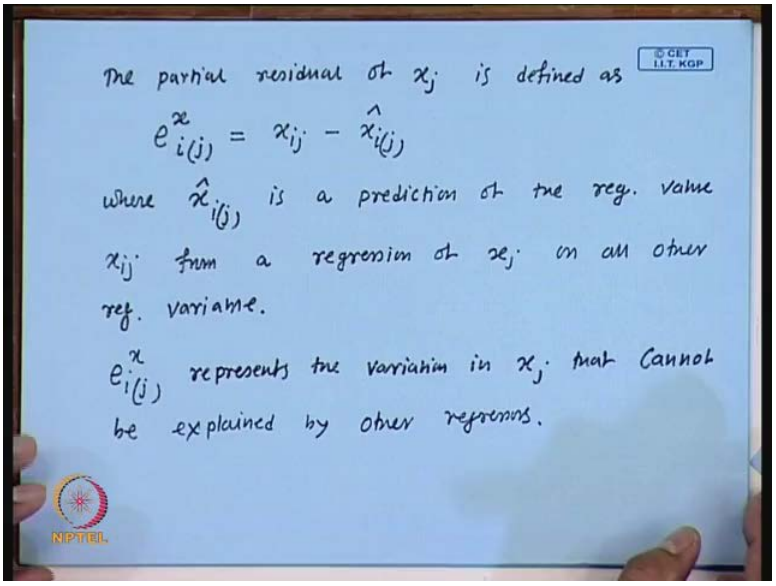


There are k minus 1 regressor instead of only two regressors, so the partial residual of y for x j is defined as e i j y, the same notation here, i for the i th observation is equal to y i minus y i j, I hope you understand the notation. So, this is the fitted value when the response variable has regressed on all the regressors except x j, so in my previous example this is what, this is nothing but the partial residual of y for x 1, this is the partial residual of y for x 1 and this is the partial residual of y for x j. So, where y i j hat is a prediction of y i from a regression model using all regressors except x j, so I hope it is clear now.

And this e i j y represents the variability in y i that is not explained by a model that excludes the regressor x j, so this is the part of variability, which is part of variability in y i which is not been explained by all the regressors except x j. So, in other words also, what does this mean is that, this is sort of the effect of all the other regressors except x j has been removed from the response variable y. So, we want to see how much of this variability which is remained unexplained, e will be how much of that variability, how much of this variability can be explained by x j.

So, that so the dependents of the, so this complicated this residual e i j y is the part of variability in the response variable y, which is not been explained by all the regressors except x j. That means, that residual represent sort of the effect of all the other regressors, except x j has been removed from the response variable y. And we want to see how much of this variability can be explained by x j alone, so that is the marginal role of x j to explain the variability in the response variable y. So, this is what we want to mean by the marginal role of x j, so this is the notation corresponds to y.

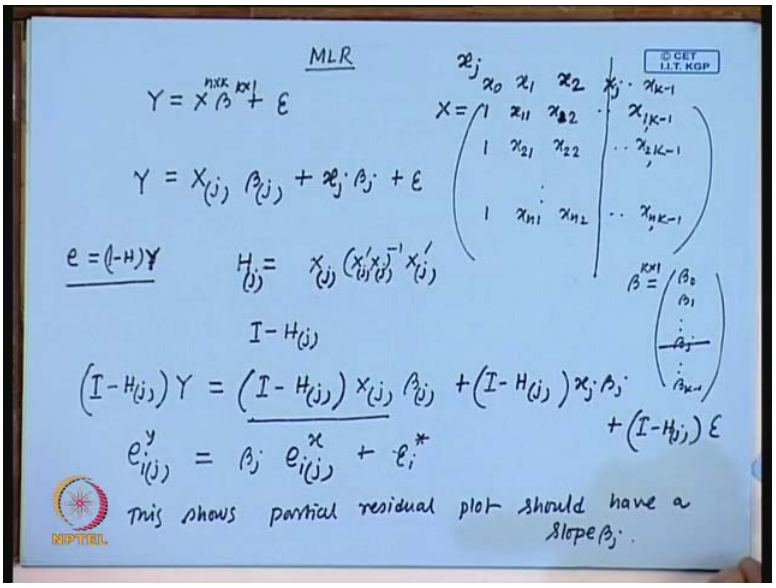
(Refer Slide Time: 40:37)



Now, the partial residual of x j similarly, x j is regressed from the remaining regressors, so partial residual of x j is defined as e i for i th observation j for j th regressor we put the notation x here, so this is x i j minus x i j hat. So, this is the fitted value obtained when the j th regressor is regressed on the remaining regressors, where x i j hat is a prediction of the regressor value x i j from regression of x j on all other regressor variable. So, similarly this e i j x represent, the variation in x j that cannot be explained by other regressor, so this is quite routine thing.

And now we have understood the technique of partial residual plot and we know little, we have some idea about the logic behind this partial residual plot, why it is so and next say what we do is that, we fit two regression model and we get two residuals. And now our aim is to find out the relation between these two residuals, whether these two residuals are linearly related or there is some other relation between these two residuals.

(Refer Slide Time: 44:19)



So, little difficult, multiple linear regression model, we are in this setup and we are interested to find the marginal role of x j on the remaining, on response variable y in the presence of other regressors. So, in the multiple linear regression model in matrix notation, the model is y equal to x beta plus epsilon, what is this x, what is this y, y is the vector of y 1, y 2, y n and beta is the vector of. So, this x is a n cross k matrix and beta is k cross 1.

So, beta consist of beta not beta 1 up to beta k minus 1 and epsilon is epsilon 1, epsilon 2 up to epsilon n, what is this x, X is equal to 1, x 1 1 x 1 2 x 1 k and then 1 x 2 1 this is the for the second observation x 2 to x 2 k and 1 x n 1 x n 2 x n k, so these are the rows are corresponds to the observations. So, there are n observations and the columns are corresponds to this is the column associated with the second regressor x 2, so this one is x 1 for example, and you call it say x naught and I should write might be k minus 1 k minus 1 k minus 1 x k minus 1.

So, what I want to do is that, I want to break it into two parts, one is x j is the x matrix except the j th column, so I want to remove the j th regressor from this matrix, so x j is the matrix obtained from x by removing the j th column here. So, similarly b j, b j is also in a before it was beta naught beta 1 up to beta k minus 1, now I am just removing the b j from this vector, so this is called b j.

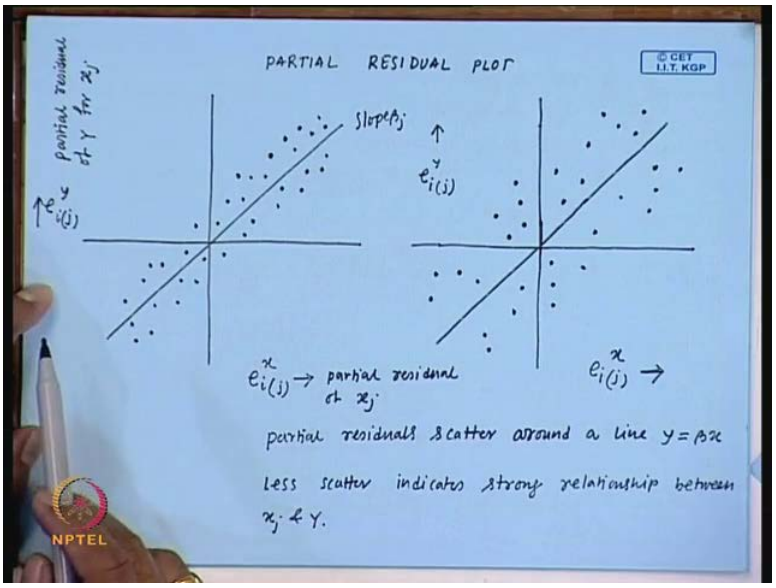
Now, I will add those two things what I have removed from here, I will add here that is x j, so x j is the j th column and b j what I have removed the coefficient from this coefficient vector plus epsilon. I hope you understand it; this is just adjustment once you remove from here and then again you add it here. And also about the hat matrix, if this is the model then the hat matrix is H equal to x (x prime x inverse) x prime, but if I consider only this part, if I just remove the j th regressor from the model. Then that means, (Refer Time: 48:26) x j here, x j here, j here, j here then this is called H j, so this is my H j.

Now, what I will do is that I will multiply this, so this is the hat (Refer Time: 48:44) matrix I will just multiply I minus H j left multiply this matrix in this equation, so what I will get is that, I will get I minus H j little difficult, but I hope if you concentrate you can understand. So, this is I will have left multiply this matrix the left hand side, so again (I minus H j) x j beta j, so see what I told the beta j, beta is consist of originally beta naught beta 1 beta j and up to beta k minus 1. And what I did is that this beta j is nothing but this beta bracket is nothing but the same vector we just remove the j th 1.

So, it will become before it was, for example it was k cross 1, now it will become k minus 1 cross 1 anyway, so plus I minus H j just this is the equation x j beta j plus I minus H j epsilon. So, what is that we know that in matrix notation e equal to I minus H y, then this one is the residual when y is regressed on all regressor except x j. So, then this one is nothing but e i j y which is equal to you can check that, this quantity is equal to 0, this is 0, because if you multiply x j here, this two will cancel out, then x j minus x j basically.

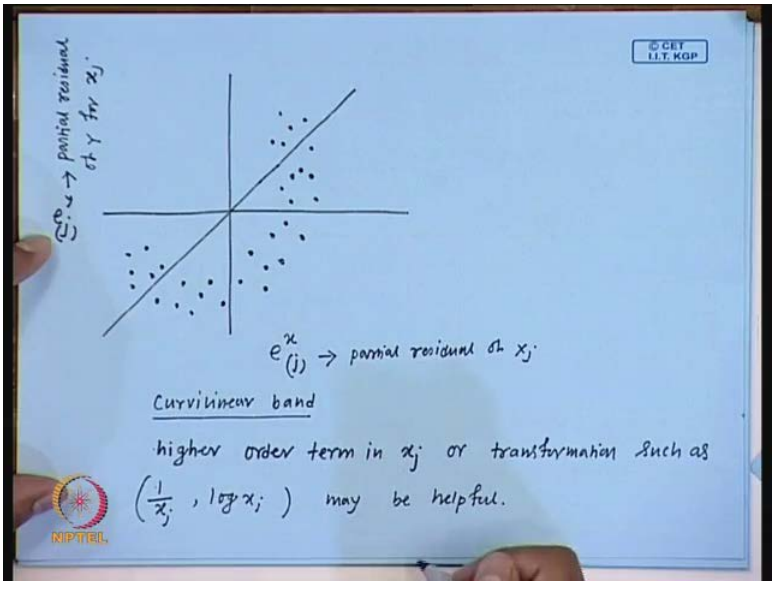
And then this one is nothing but according to our notation this is nothing but beta j into this is sort of when the residual obtained, when x j is regressed on the remaining regressors, so this is e i j x and I call it epsilon i star. So, this shows that, so this is the relation between if j th regressor has a linear relationship with the response variable, then the two residuals will have also the linear relationship with the same regression coefficient. So, this shows partial residual plot should have a slope beta j, so what we understood is that, we know how to compute two residuals. And we have understood that the relationship between these two residuals are also linear state line fit between these two residuals, with the slope beta j, now I will just talk about the different pattern of the partial residual plot.

(Refer Slide Time: 53:40)



So, here you can see the partial residual of y is along the particle axis and the partial residual of x j along with the x axis, and similarly here you can check that the partial residuals are scattered around a line y equal to beta j x in both the cases. And if this is the pattern obtained for your example, then you can conclude that there is a strong linear relationship and also it says that, the less scattered indicates strong linear relationship between x j and y, so this is one type. So, the less scattered indicates strong linear relationship between the x j and y, so this is how the marginal role of x j in 1 y.

(Refer Slide Time: 55:01)



And the other pattern is, if the pattern is like this in curvilinear band, then this indicates that the higher order term in x j or transformations such as, 1 by x j or log x j may be helpful. That means, this indicates that the x j is not linearly related to y either you have to go for the higher order term for x j, or you can you have to go for some transformation on x j, like log x j or 1 by x j. So, this is how we find the marginal role of regression variable on the response variable o i, and I think I have to stop now.

Thank you.