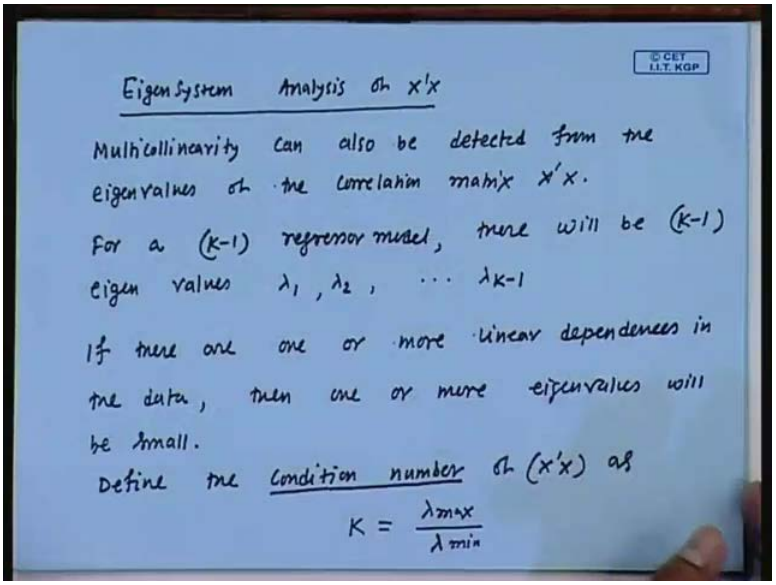
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 16
Multicollinearity (Contd.)

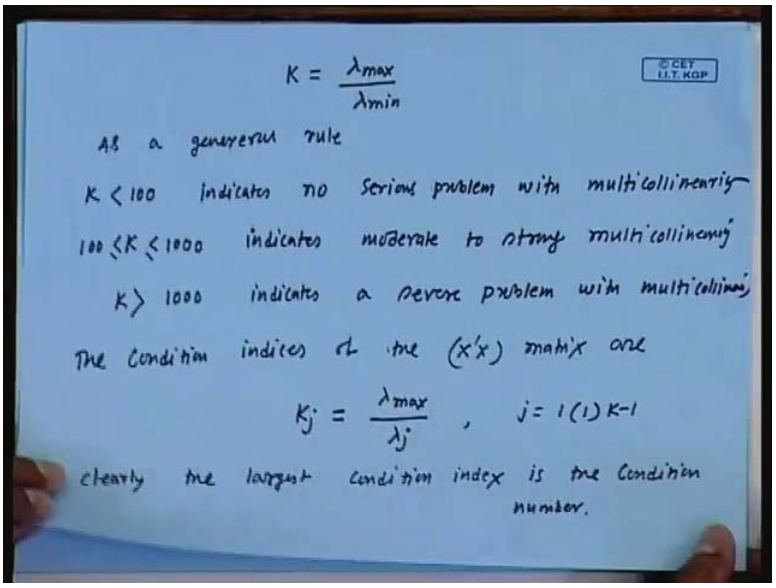
Hi, this is my 3rd lecture on Multicollinearity. The problem of multicollinearity exists when two or more regression variables are dependent or you can say that when two or more random regressive variables are linearly dependent. So, in the last class, we talked about different technique techniques to detect multicollinearity, we learned about examination of correlation matrix, and also we learned about you know Eigen system analysis of x prime index matrix or the correlation matrix.

(Refer Slide Time: 01:59)



So, first I will recall the Eigen system analysis we learned in the previous class, well what we will do here is that we first compute the K minus 1, Eigen values of the x prime x matrix, and then we compute the condition number which is K denoted by K equal to lambda max by lambda minimum.

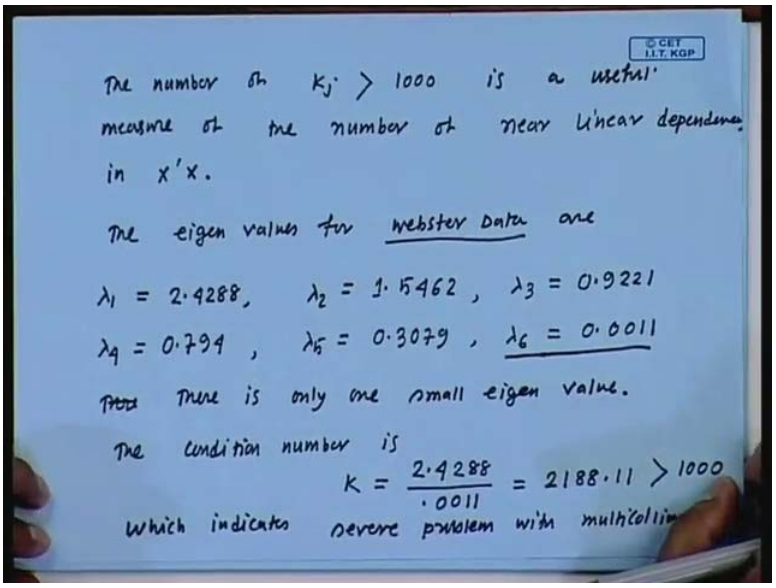
(Refer Slide Time: 02:36)



And the larger value of K indicates the severe problem with multicollinearity as a general rule you know this is what we talked in the last class, if K is less than 100 then that indicates no serious problem with multicollinearity. If K is between 100 and 1000 then that indicates moderate to strong multicollinearity, but if K is greater than 1000 then that indicates a severe problem with multicollinearity.

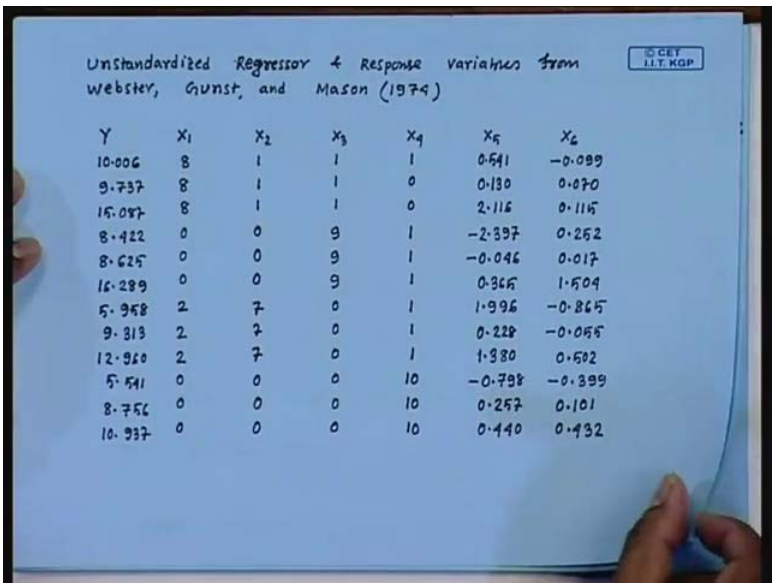
So, this is how we you know condition number is a used to detect the multicollinearity, the advantage of this Eigen system analysis is that it not only detect multicollinearity, it can measure the number of linear dependencies in the correlation matrix. And also, you know it can determine or it can identify the nature of linear dependencies between the regresses, so for that you know what we do is that we compute the condition indices.

(Refer Slide Time: 04:40)



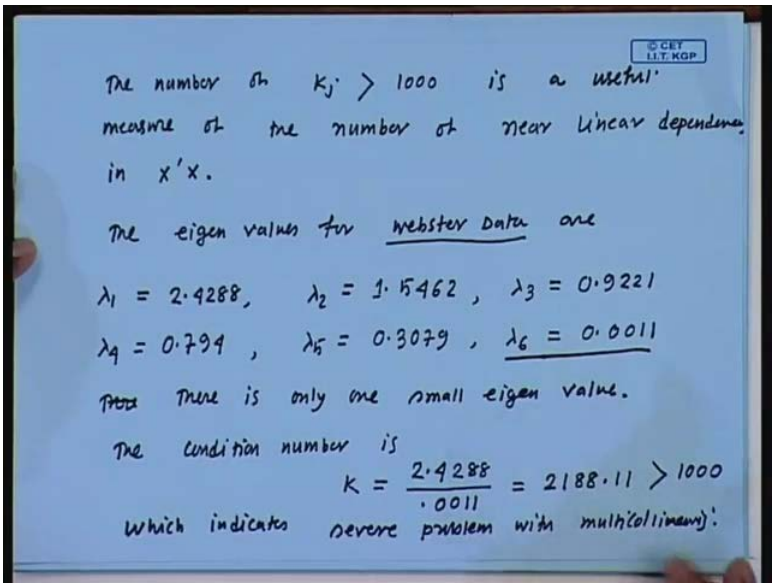
Here, is the condition indices K j, K j is associated with the jth regressor, so the K j is lambda maximum by lambda j, and the number of K j greater than 1000 is a useful measure of the number of linear dependencies in x prime x.

(Refer Slide Time: 05:05)



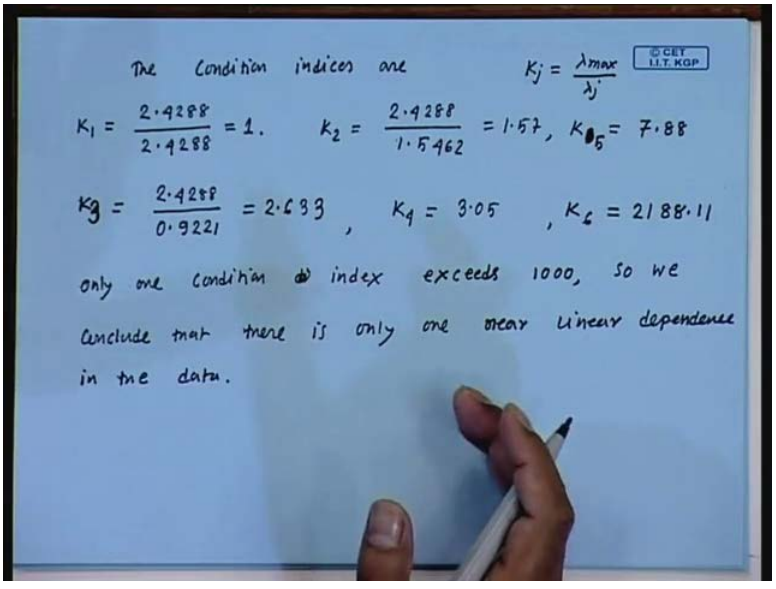
And then I illustrated this result using the webster data, this is the webster data we talked about in the previous class, we have 6 regressors response variable right and I referred this data as you know webster data.

(Refer Slide Time: 05:29)



Now, for the Eigen values of x prime x matrix of the correlation matrix for the webster data is here, you know lambda 1 lambda 2 lambda 3 lambda 4 lambda 5 lambda 6 and the smallest Eigen value is this one, which is closed to 0. Now, we compute the condition number, so the condition number is 2188 which is greater than 1000, so this condition number indicates the presence of severe multicollinearity in the webster data. So, next what we will do is that, we will compute the condition indices also since we know the Eigen values we can compute the condition indices.

(Refer Slide Time: 06:40)

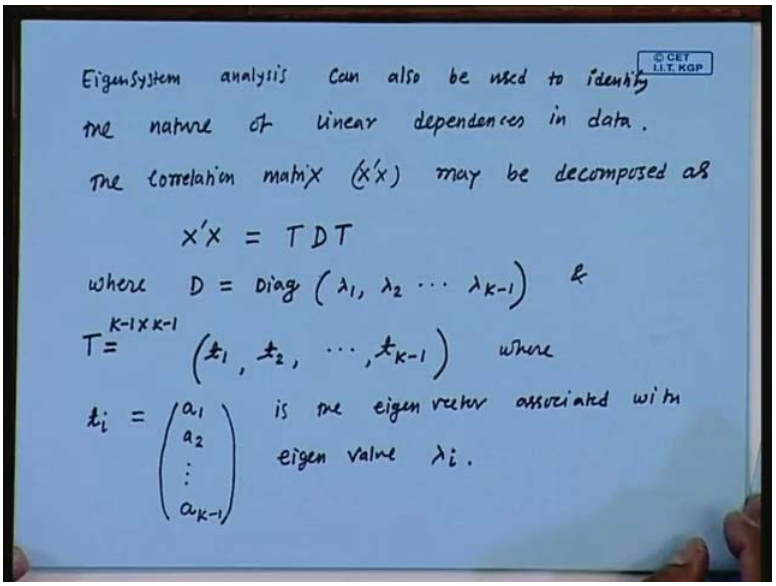


So, the condition indices are what is condition indices K j, K j is equal to lambda max by lambda j, so K 1 is lambda max is 2.4288 by lambda 1 is basically maximum, so lambda 1 is 2.428 which is equal to 1. The K 2 is lambda max which is equal to 2.4288 by lambda 2 which is equal to 1.5462 equal to 1.57. Similarly, you compute K 3 which is equal to lambda max by lambda 3 is equal to 7.88, K 4 is lambda max by lambda 4 that is 2.4288 by lambda 4 is 0.9221 which is equal to 2.633.

So, I did this a mistake here, this is K 3 this is K 5, any way you can compute for this K 3 and then K 4 is equal to 3.05, K 5 is equal to this quantity and K 6 is equal to 2188.11. Now, we know that here only one condition index exceed 1000, that is K 6, so we conclude that there is only one near linear dependents in the data, because you know we I mentioned before that the number of K j greater than 1000 that measures the number of linear dependencies in the data.

And since, here only one K j that is K 6 is greater than 1000, that is why the number of linear dependents or dependencies is equal to one only. Well, now what will this technique has lot of advantages like you know it not only detect the presence of multicollinearity it can measure the number of linear dependencies in the data and also it can identify the nature of the linear dependencies.

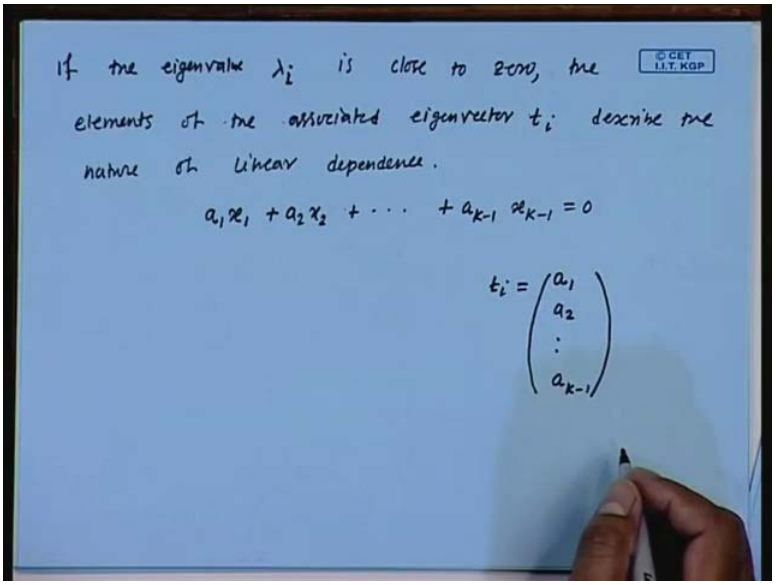
(Refer Slide Time: 11:03)



So, now, we will explain that portion, here Eigen system analysis can also be used to identify the nature of linear dependencies in data. So, let me explain this portion first, the

correlation matrix x prime x may be decomposed as you know x prime x you can write or decompose as x prime x is equal to T D T, where D is a diagonal matrix, whose main diagonal elements are the Eigen values. So, D equal to diagonal lambda 1 lambda 2 lambda K minus 1, and t is equal to t is a K minus 1 cross K minus 1 matrix, and whose the columns of this T matrix their t 1 t 2 t K minus 1. So, here this t i is the Eigen vector associated with lambda i, where t I is equal to a 1 a 2 a K minus 1 is the Eigen vector associated with Eigen value lambda i, so this is the decomposition of the correlation matrix.

(Refer Slide Time: 15:23)

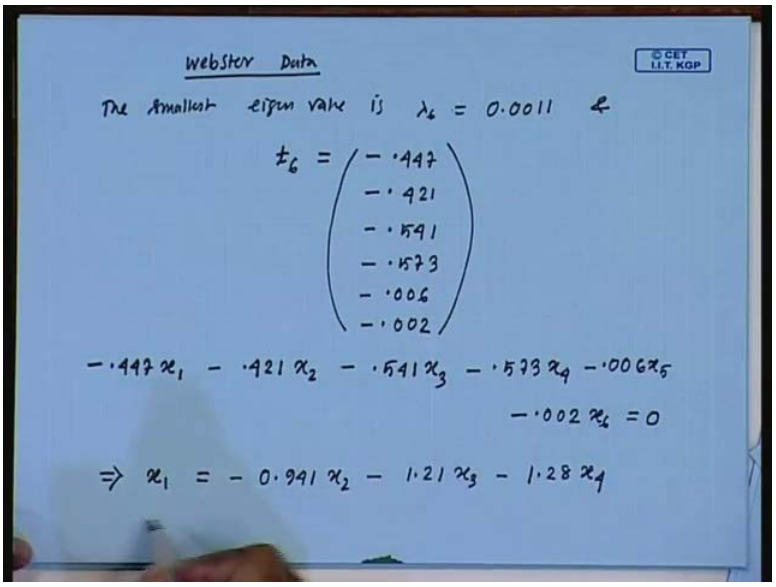


Now, if lambda j, if the Eigen value lambda i is closed to 0, the elements of the associated Eigen vector, that is t i describe the nature of linear dependents. So, the nature of this linear dependents is like a 1 x 1 a 2 x 2 plus a K minus 1 x K minus 1 is equal to 0, so this coefficient of the regressor variables a 1 a 2 a K minus 1. They are basically the elements of t i, this is the Eigen vector associated with lambda i and lambda is very closed to 0, so t i is the Eigen vector associated with lambda i and t i is equal to a 1 a 2 a K minus 1.

So, maybe I will just give the little motivation behind this, you know if lambda i is closed to 0 then the condition index associated with lambda i is large, I mean that would be greater than 1000. And I mean and then you will get one linear dependents between the regressor variables associated with lambda i, so that is why you know as I mentioned

before also you know that the number of condition indices greater than 1000 that measures the number of linear dependencies in the data. And corresponding to each lambda i for which the condition index is greater than 1000, you will get a linear dependents or you will get you can identify linear dependents between the regressor coefficient.

(Refer Slide Time: 19:16)

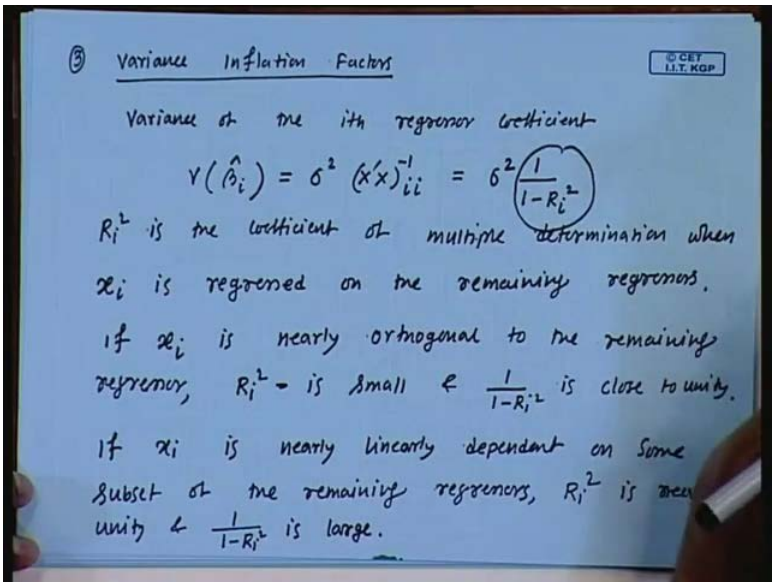


Let me explain I mean illustrate this thing using the webster data, here the smallest Eigen value is lambda 6 which is equal to 0.0011 and the associated Eigen vector that is say t 6 is equal to minus 0.447 minus 0.421 minus 0.541 minus 0.573 minus 0.006 minus 0.002. So, this is the associated Eigen vector corresponds to lambda 6 and then the nature of linear dependents is minus 0.447 x 1 minus 0.421 x 2 minus 0.541 x 3 minus 0.573 x 4 minus 0.006 x 5 minus 0.002 x 6 equal to 0.

And from here you know since these two are very small, we can ignore x 5 and x 6 from this equation and this implies that x 1 is equal to minus 0.941 x 2 minus 1.21 x 3 minus 1.28 x 4. So, this is the linear dependents between the regressor x 1 x 2 and x 3 and x 4, and this linear dependents is associated with lambda 6. And if you have more Eigen value which are closed to 0, corresponds to each lambda i mean which are very small or which are closed to 0, you will get a linear dependents like this. So, what we learned from this Eigen system analysis is that it can detect the presence of multicollinearity, it

can measure the number of linear dependencies in the data, and also it can identify the nature of linear dependencies in the data.

(Refer Slide Time: 23:16)



So, next we move to the variants inflation factor, this is another way to you know to detect the presence of multicollinearity, this is called variance inflation factors. So, first we will recall the variance of ith regressors coefficient, I mean variance of ith regressor coefficient means the variance of the least co estimate of i th regression coefficient beta i hat.

We know that this one is equal to sigma square x prime x inverse the i ith element of x prime x inverse, now you know this can be it can be proved that this i ith element is equal to 1 by 1 minus R i square sigma square here. Now, what is this R i square R i square is the coefficient of multiple determination, when x i is regressed on the remaining regressors.

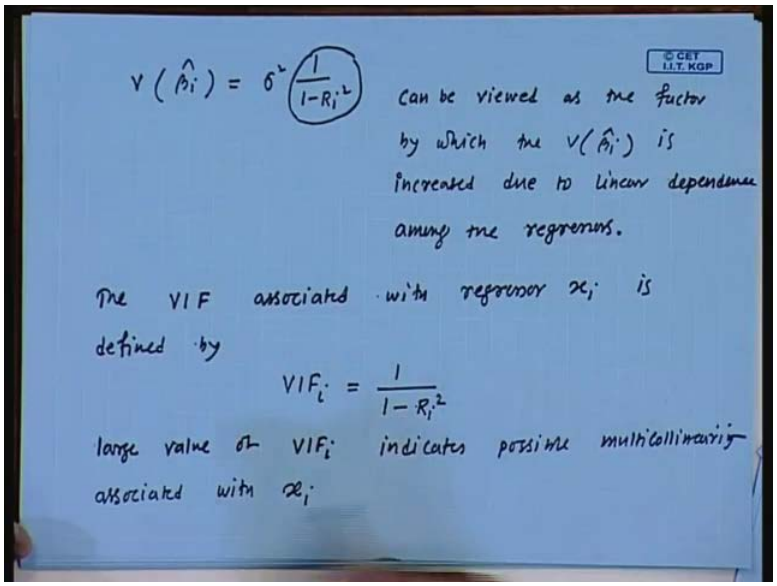
Now, see if the ith regressor x i is nearly orthogonal to the remaining regressor, here you know this nearly orthogonal I mean that if the ith regressor is independent of the remaining regressors. Then R i square is small and 1 by 1 minus R i square is close to unity well, so the meaning of this one is that no x i is nearly orthogonal to the remaining regressors; that means, x i is independent of the remaining regressors. That means, there is no linear dependents associated with x i, I mean x i can cannot be represent in terms of the linear combination of the other regressors.

Then the coefficient of multiple determination when you are regressing x i on the remaining regressors the coefficient multiple determination will be small, and the value of 1 minus 1 by 1 minus R square is close to unity, and the variance of beta i hat is going to be sigma square close to sigma square. Now, if x i is nearly linearly dependent on some subset of the of the remaining regressor, R i square value will be near to close to 1, R i square is near unity and 1 by 1 minus R i square is large.

So, the meaning of this, you know x i is linearly dependent on some subset of the remaining regressors; that means, there is a linear dependents between x i and some subset of the remaining regressors. If some linear dependence is there between x i and some subset of the remaining regressors; that means, x i can be represented in terms of as a linear combination of some subset of the remaining regressors well.

So, the which implies that R i square which is the coefficient of multiple determination when x i is regressed on the remaining regressors is will be large and that will be closed to unity. And that implies that the value of 1 by 1 minus R i square is large, that ultimately you know if x i is there is a linear dependence in the data between the regressors. Then variance of beta i hat is going to be large because this factor is going to be large well, so this factor you know this 1 by 1 minus R i square well let me write here.

(Refer Slide Time: 30:48)

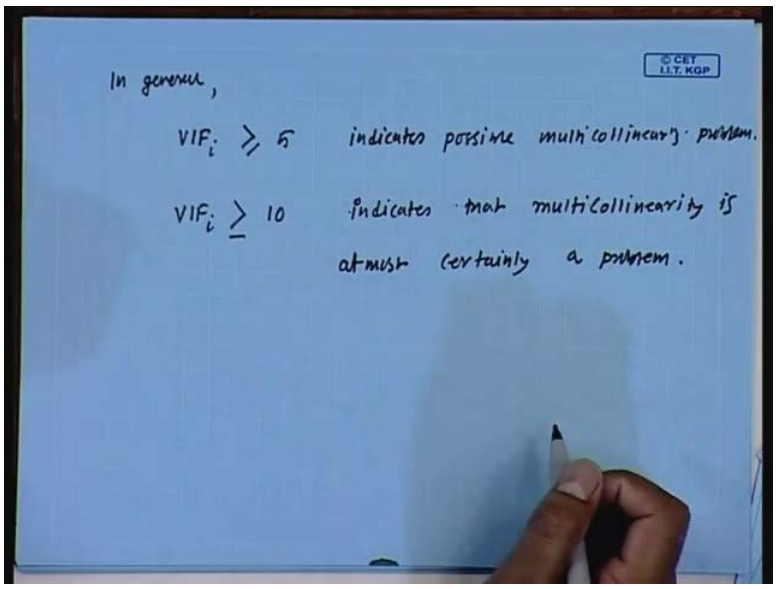


So, variance of beta i hat is equal to sigma square into 1 by 1 minus R i square and this quantity this factor you know this can be viewed as the factor, by which the variance of

beta i hat is increased due to linear dependents among the regressors well. So, if there is a linear dependence between x i and a subset of the remaining regressors, then the value of this is large, then the variance of beta i hat is also large. Now, if x i is independent of the remaining regressors or I say that it is if it is nearly orthogonal to the remaining regressors, then this value this factor is closed to 1 and the variance of beta i hat is almost equal to sigma square.

So, the variance inflation factor VIF associated with regressor x i is defined by VIF i which is equal to 1 by 1 minus R i square and; obviously, large value of VIF i indicates possible multicollinearity associated with x i. The meaning of this one is that you know if multicollinearity associated with x i; that means, this we if this is large; that means, there is a linear dependence between x i and subset of the remaining regressors. Then only the value of this one is going to be large, so the large value of v i f i indicates possible multicollinearity associated with the regressor x i.

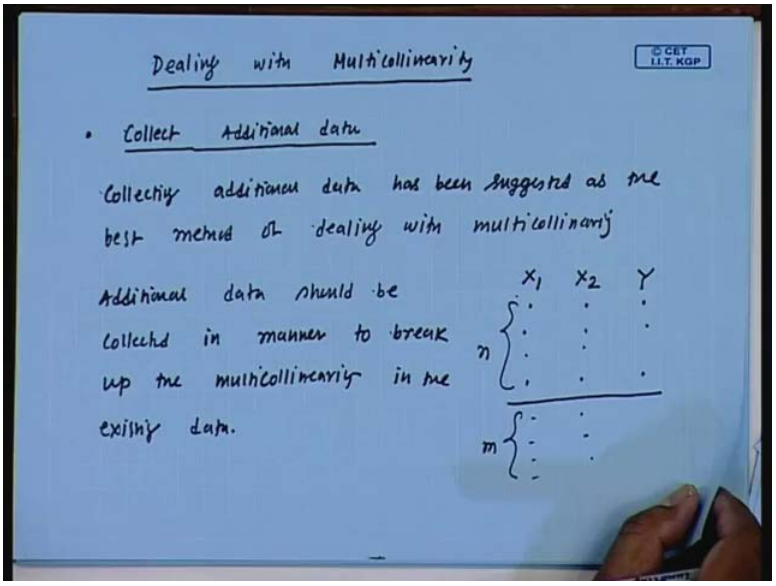
(Refer Slide Time: 34:50)



Now in general you know VIF i greater than equal to 5 indicates possible multicollinearity problem, and VIF i greater than or equal to 10 indicates that multicollinearity is almost certainly a problem. So, this is how the variance inflation factor associated with the ith regressor can be used to detect the multicollinearity, so this technique variance inflation factor can determine can detect the problem of multicollinearity.

But, any I mean of course, it cannot identify the nature of multicollinearity, so Eigen system analysis is a better technique, because it can detect the multicollinearity, it can measure the number of linear dependencies in the data, and also it can identify the nature of linear dependences in the data. So, next we will be talking about you know if you can if you detect that the you know there is multicollinearity in the data, then how to deal with multicollinearity, so we will be talking about several techniques to deal with multicollinearity.

(Refer Slide Time: 37:28)



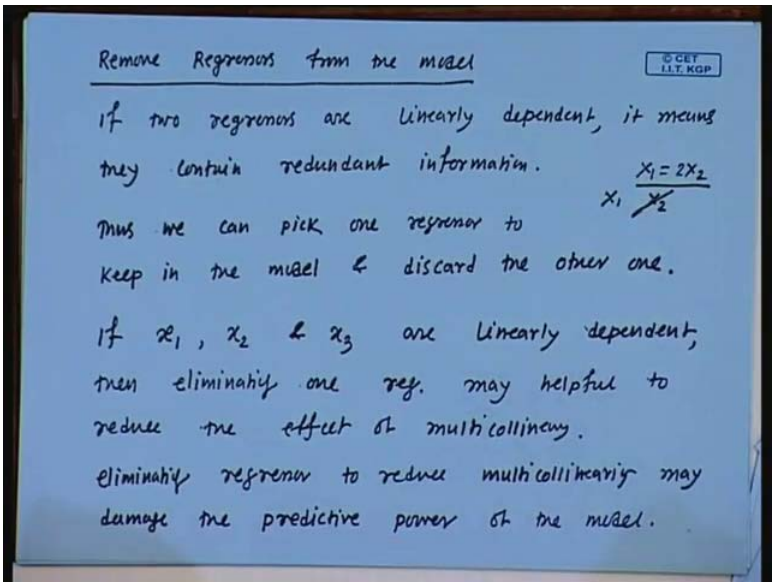
So, dealing with, so first technique is you know collect additional data, well you know collecting traditional data has been suggested, as the best method of dealing with multicollinearity. What it says is that let me illustrate this one it will get ,suppose you have you are in the multiple linear regression model, and suppose you have only two regressors x 1 and x 2 and the response variable Y, and you have end data points.

So, you have some data you have end data, now you have detected that you know multicollinearity exist in this data; that means, here we have only two regressors; that means, x 1 and x 2 they are linearly dependent. So, what we do is that we collect some more data say another m data to break the existing multicollinearity in the present data. So, what we do is that suppose, so this two regressors they are linearly dependent that is why that is the I mean multicollinearity, because of the linear dependence between these two regressors.

Now, this additional data should be collected in a manner to break up the multicollinearity in the existing data, I hope you understood that you know initially you had end data points, and here you know x 1 and x 2 they are linearly dependent. So, you collect another some more additional data say m data points, in such a way you know, when you combine the complete set of data you know n plus m data then x and x 1 and x 2 are not any more linearly dependent.

You have to collect the data in a manner to break up the multicollinearity in the existing data, so this is you know one way to deal with multicollinearity or to break the multicollinearity in the existing data, but in many instance, you know this is not possible in practice.

(Refer Slide Time: 42:38)



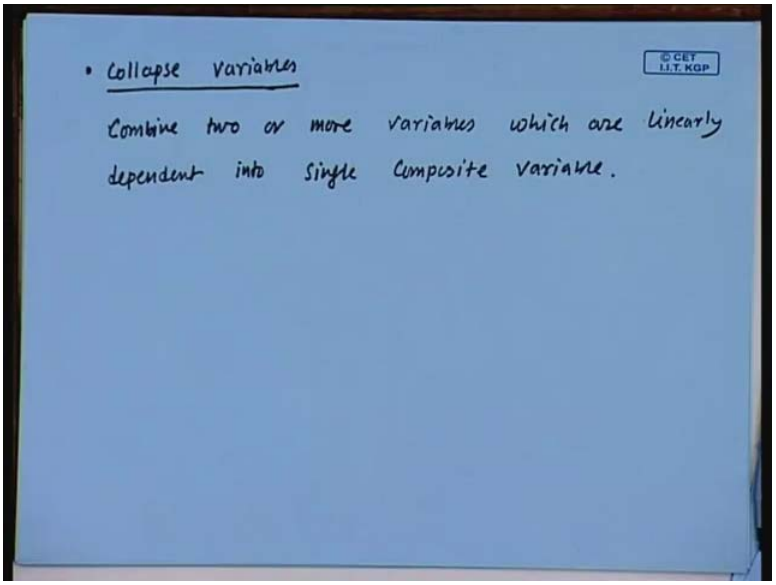
Next, we will talk about one more technique to deal with multicollinearity that is called remove regressors from the model, here if two regressors are linearly dependent, it means the content redundant information. So, what we can do is that you know if two regressors for example, x 1 and x 2 they are linearly dependent may be for example, say x 1 is equal to twice of x 2, but then I mean basically they I mean the information regarding x 2 is redundant.

So, what we can do is that thus we can pick one regresor to keep in the model and discard the other one, so this basically suggest you know, if you have two regressors, which are linearly dependent, and you can remove the other one. You can remove say for

example, x 2 from this model, if say for example, x 1 x 2 and x 3 are linearly dependent, then eliminating one regressor variable, may helpful to reduce the effect of multicollinearity, but the problem is that you know if say for example, you have three regressors, x 1, x 2, x 3 in the model.

And or may be more regressors, but x 1, x 2, x 3, they are linearly independent, then maybe you can remove one regressor for example, x 3 from the model, but you know to reduce the effect of multicollinearity. But, if x 3 might happen that you know x 3 the regressor which you have removed that x 3 might be significant to explain the variability in the response variable. In that case you know this removing one regressor may damage the predictive power of the model, well so that is why it says that you know eliminating regressor to reduce multicollinearity may damage the predictive power of the model.

(Refer Slide Time: 48:20)



So, this is one way you know to deal with the problem of multicollinearity, the other technique is collapse variables, so it says that you know you combine if there are linear dependence between two or more than two regressors, which are linearly dependent into a single composite variable. So, basically this collapse variable it says that if you have linear dependence between say 2 or more than 2 regressors, then you can combine you know those regressors by a composite regressor variable.

So, these are the techniques to deal with multicollinearity, so one is you know collecting more data point, and the other one is removing one regressor from the model and

combining the regressors, which are linearly dependent. So, that is like ways for in this module we have learned what is multicollinearity and you know this multicollinearity is the name of a problem in multiple regression model well.

So, the problem of multicollinearity arises, when two or more regressor variables are linearly dependent and we have learned you know how to detect the multicollinearity, if it exist in the data. And also first you know we learned about how what are the problems due to multicollinearity, and then we have learned you know how to detect multicollinearity, if it exist in the data and also we learned how to deal with multicollinearity, so that is all for today.

Thank you for your attention.