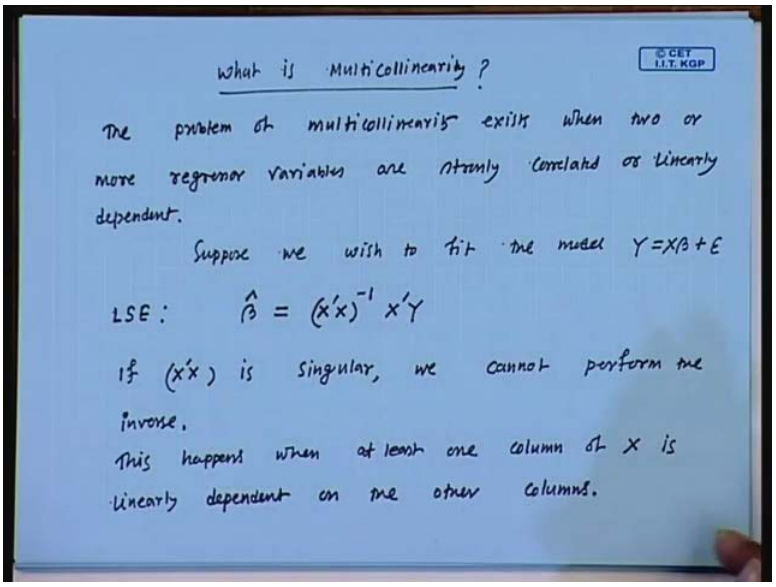
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 14
Multicollinearity

This is my first lecture on Multicollinearity, and today we learn what is multi multicollinearity and the effects of multicollinearity or the problems due to multicollinearity.

(Refer Slide Time: 01:14)

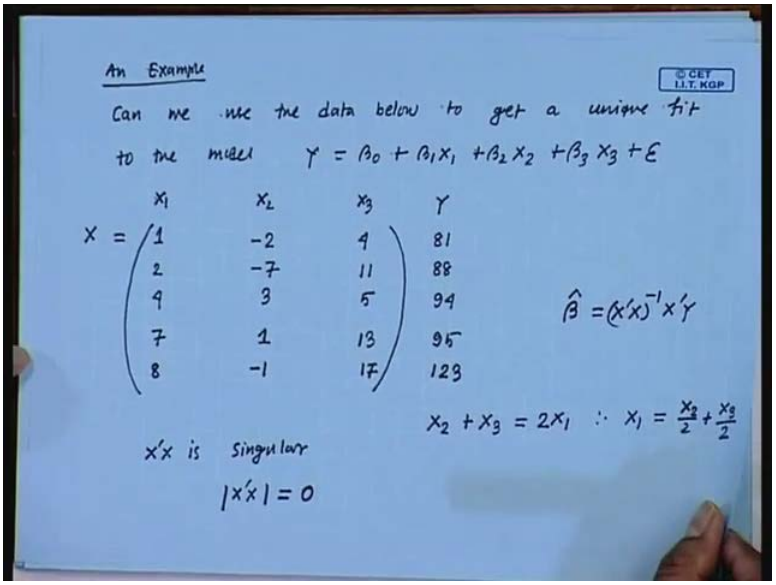


So, what is multicollinearity? The problem of multicollinearity exists when two or more regressors variable are strongly correlated or linearly dependent. Suppose, we wish to fit the model Y equal to X beta plus epsilon, so we assuming that, we considering that, this is the matrix form of multiple linear regression and there are k minus 1 regressors and one response variable. So, we know that, the least square estimates of beta is beta hat equal to X prime X inverse X prime Y.

So now, if X prime X is singular then, we cannot perform the inverse, X prime X inverse and singular means, the determinant is equal to 0, X prime X determent determinant is equal to 0. And this happens, when atleast 1 column of X is linearly dependent on the other columns. So that means, the i th column of this matrix X stands for the i th regressor variable. So, the meaning of this one is that, if one regressor can be I mean, one

regressor is linearly dependent on the other regressors then, X prime X is singular and we cannot compute the inverse of the X prime X matrix, let me give one example for this one.

(Refer Slide Time: 05:58)



An example, can we use the data below to get a unique fit to the model Y equal to beta naught plus beta 1 X 1 plus beta 2 X 2 plus beta 3 X 3 plus epsilon. And the data carries data X 1 X 2 X 3 and the response variable Y, so 1 minus 2 4 81 2 minus 7 11 88 4 3 5 94 7 1 13 95 8 minus 1 17 1 23. So, here in this example, we have three regressor variables X 1 X 2 X 3 and one response variable, and here is the data. Now, the question is, whether we can use the data, use this data to get a unique fit to this model.

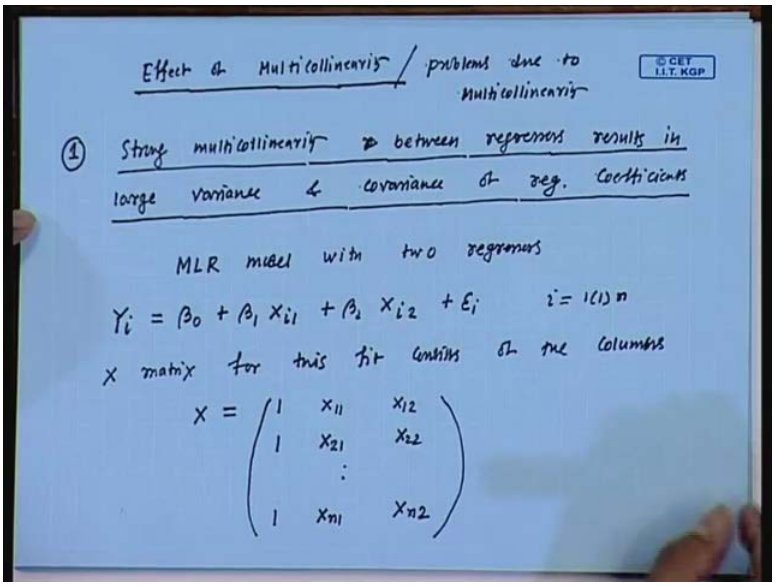
Of course, we can get a unique fit that means, we can estimate the regression coefficients using the least square estimate that is, beta hat equal to X prime X inverse X prime Y, provided X prime X is not singular. Now here, if you observed, it is not difficult to check that, this three regressors are not independent. Here you check that, X 2 plus X 3, take this first observation, so X 2 value is minus 2, X 3 value is 4 and X 1 value is equal to 1, you can check that, X 2 plus X 3, that is equal to 2, so X 2 plus X 3 is twice of X 1.

Take the second observation, X 2 plus X 3 is equal to 4, which is two times of X 1, third observation also X 2 plus X 3 equal to 8, which is 2 X 1. So here, the relation between the regressors is like X 2 plus X 3 is equal to twice X 1 or same as writing that, X 1 can be expressed in terms of using, X 1 depends on X 2 and X 3, so X 1 is X 2 by 2 plus X 3

by 2. So here, really we cannot estimate the regression coefficients, because X prime X is is singular here, so here because of this fact, they are not independent.

So, that is why, X prime X is singular, so X prime X determinant is, here it is exactly equal to 0, determinant is equal to 0. So, this is an example here, which illustrate the definition of multicollinearity, here I mean, in this particular data, the problem of multicollinearity exists, because the columns of X, so the X is nothing but the X matrix is nothing but this one. So, here the one column for example, the first column can be I mean, it linearly depend can be written as the linear combination of X 2 and X 3. So, that is why, X prime X is singular and the problem of multicollinearity exists here. So, next we will be talking about the effect of multicollinearity.

(Refer Slide Time: 12:18)

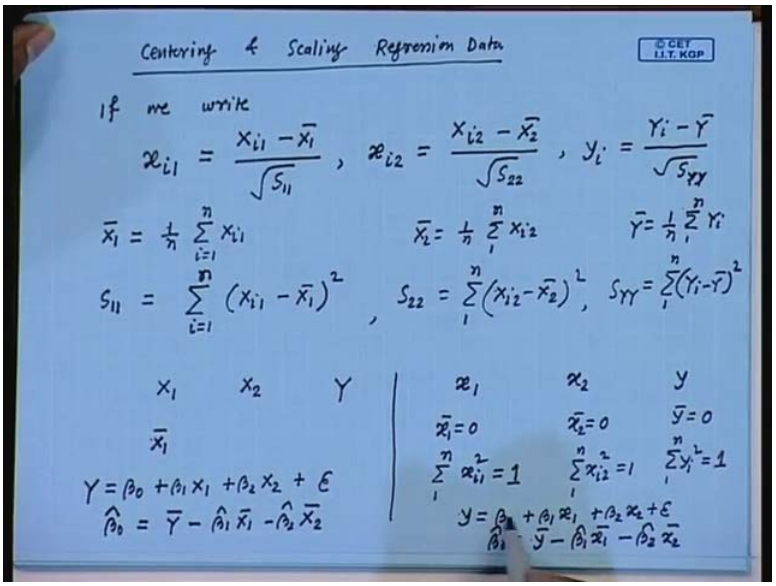


Multicollinearity, the first or I can also say that, the problems due to multicollinearity, so the first one is, it says the strong multicollinearity between regressors, results in large variance and covariance of regression coefficients. So now, I am going to illustrate, what I mean by this, this says that, if there exist strong multicollinearity in the data then, that results in large variance and covariance of regression coefficients. So, let me consider multiple linear regression model with two regressors.

So, and my model here is Y i equal to beta naught plus beta 1 X i 1 plus beta 2 X i 2 plus epsilon i, so this i stands for the i th observation, so i is from 1 to n. Now, the X matrix for this fit consists of the columns, so here is the X matrix, X matrix for this model is, we

know that, this is 1 X 1 1 X 1 2 1 X 2 1 X 2 2 like that 1, X n 1 X n 2. Now, we will talk about the centering and scaling of this regression data.

(Refer Slide Time: 16:30)



Centering and scaling of regression data, so what we do is that, if we replace, if we write X i 1 is equal to X i 1 minus X 1 bar by X 1 1 square root of that. So, let me tell, what is this notation, X 1 bar of course, it is the mean associated with the first regressor. So, this is equal to 1 by n summation X i 1, i equal to 1 to n and S 1 1 is equal to summation X i 1 minus X 1 bar square, sum over i equal to 1 to n. So, let me write also, say x small x i 2 equal to capital X i 2 minus X 2 bar by S 2 2 square root and y i small y i is equal to capital Y i minus Y bar by root S y y.

So, this is the mean associated with the second regressor, so X 2 bar is equal to 1 by n summation X i 2, i is from 1 to n and S 2 2 is equal to X i 2 minus X 2 bar whole square 1 to n. And here also, Y bar is equal to summation Y i, 1 to n and S y y is equal to summation Y i minus Y bar whole square. So, what I did here is that, my original data is X 1 X 2 and Y and then, what I am doing is that, I am replacing this X 1 by small x 1, so small x 1 small x 2, just notation only and then, small y.

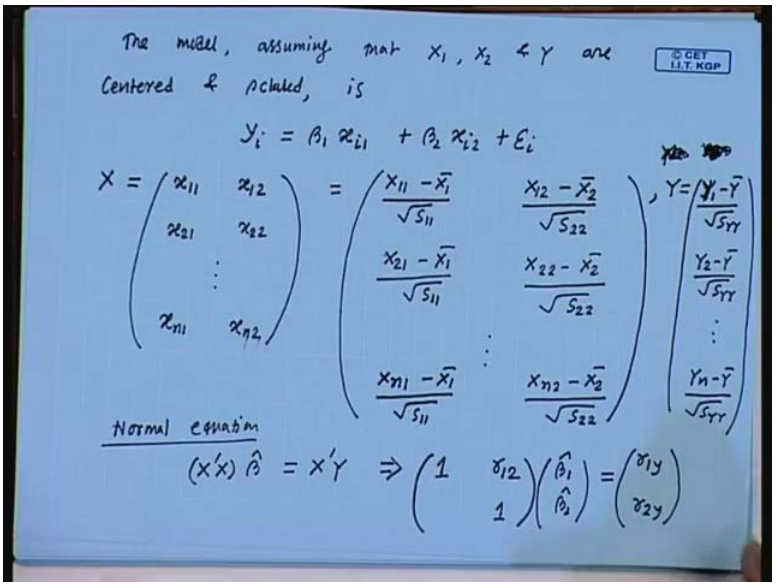
Now, one thing we need to observe is that, the mean of the original observation is I mean, the mean associated with the first regressor for the given data is of course, it is always X 1 bar. But here, if you take the mean of the transformed data, now this the

mean of small x i 1 is always equal to 0. So, here x 1 bar is equal to 0, similarly x 2 bar is also equal to 0 and y bar is also equal to 0.

And the other thing to observe here is that, summation x i 1 square 1 to n, this is equal to 1 and this is also true for the second regressor like summation x i 2 square 1 to n, is equal to 1 and summation y i square 1 to n, is equal to 1. Now, the model for the original data is, we wanted to fit this model Y equal to beta naught plus beta 1 X 1 plus beta 2 X 2 plus epsilon, this transformation is called centering and scaling of regression data. And here, for this model we know that, the least square estimates gives beta naught hat is equal to Y bar minus beta 1 hat X 1 bar minus beta 2 hat X 2 bar.

But, if you fit the same model here, for the transformed data also, if you fit the model like y equal to beta naught plus beta 1 x 1 plus beta 2 x 2 plus epsilon. For the transformed data I am fitting the same model then, it is not difficult to check that, that beta naught hat is going to be 0 here, because see beta naught hat is equal to y bar minus beta 1 hat x 1 bar minus beta 2 hat x 2 bar. But, all this you see, y bar is equal to 0, x bar is equal to 0, x 1 bar is equal to 0 and x 2 bar is also equal to 0. So here, this intercept for this transformed data is always going to be equal to 0, so that is why, for this scaled and centered data, we will omit the intercept beta naught from the model.

(Refer Slide Time: 25:07)



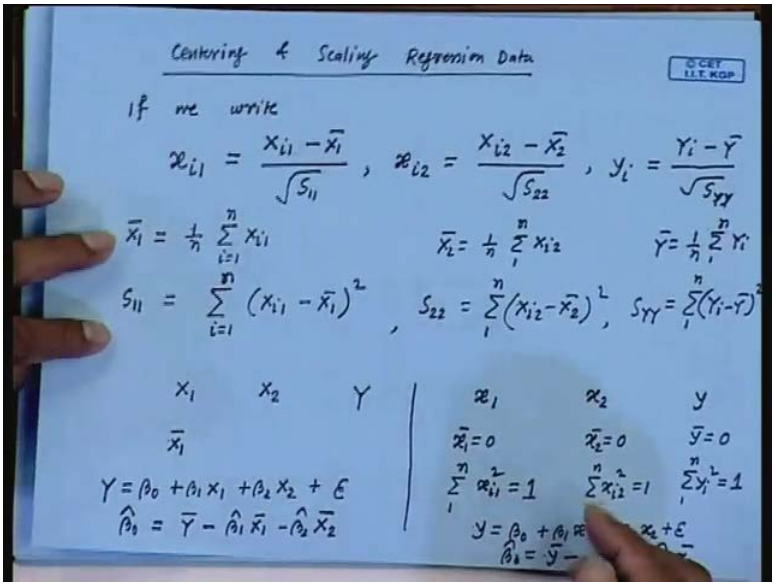
So, the model assuming that, X 1 X 2 and Y are centered and scaled is just y i equal to beta 1 x i 1 plus beta 2 x i 2 plus epsilon i. So, we omitted the intercept beta naught from

the model, because for the centered or the scaled data, we just checked that, intercept beta naught is always going to be equal to 0. So, next we will write down the X matrix for this model, so the X matrix is equal to, basically it is x 1 1 x 1 2, I am writing small that means, this is for the transformed data, x 2 1 x 2 2 and x n 1 x n 2. See that, column 1 1 1 1 is not there, that column is associated with the intercept beta naught.

So, this one is basically, x 1 1 is nothing but X 1 1 minus X 1 bar by root of S 1 1, x 1 2 is capital X 1 2 minus X 2 bar by root of S 2 2, and similarly X 2 1 minus X 1 bar by root of S 1 1, X 2 2 minus X 2 bar S 2 2. And we have X n 1 minus X 1 bar by root of S 1 1 X n 2 minus X 2 bar by root over of S 2 2. So, this is the X matrix for the transformed data you can say, now what is the normal equation associated with this model. The normal equation in general, basically I am trying to find the least square estimate of two regression coefficients, beta 1 and beta 2.

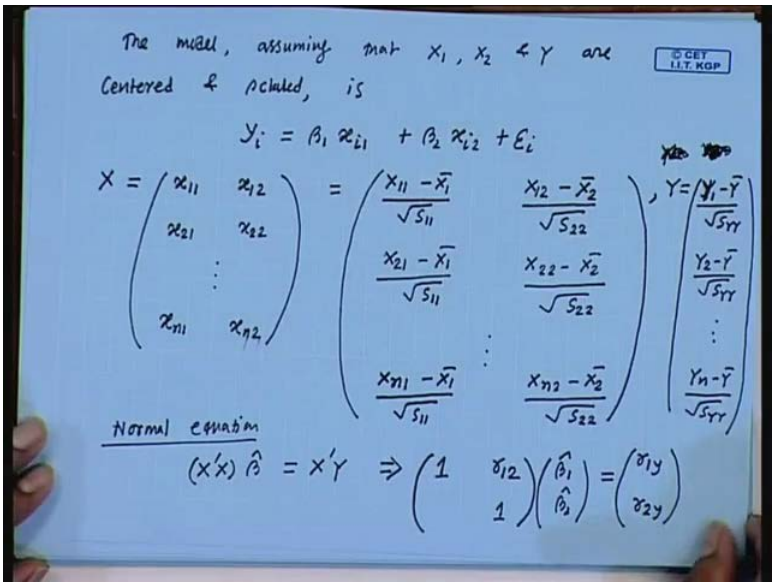
So, the normal equation is X prime X beta hat equal to X prime Y, you know what is Y, Y is, in matrix form Y equal to capital Y 1 minus Y bar by root of S y y, Y 2 minus Y bar by root over of S y y. Similarly Y n minus Y bar by root over of S y y, so this is what the y matrix is. So, here is the normal equation, now what is X prime X, this one is, this implies say, X prime X is, you can check that, the X prime X is 1 and then, r 1 2, this r 1 2 is nothing but the sample correlation between X 1 and X 2 and 1 here.

(Refer Slide Time: 31:07)



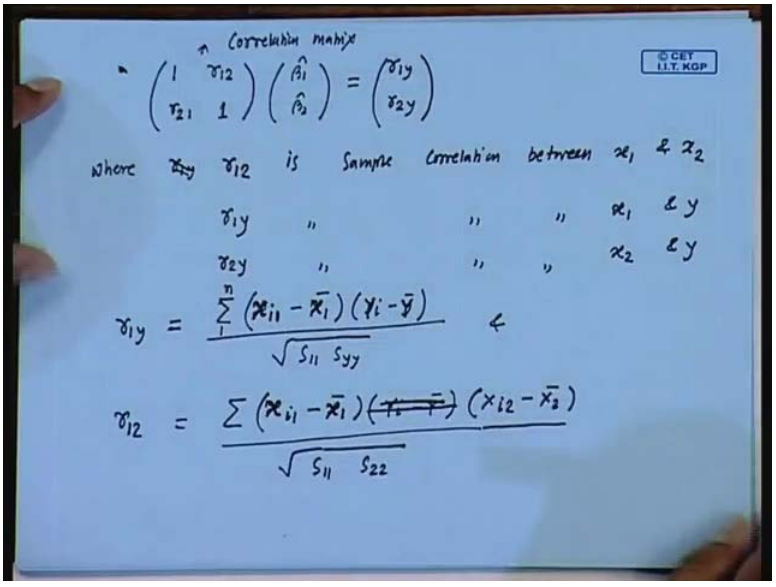
Because of the fact that, this sum x i 1 square is equal to 1 and sum x i 2 square is equal to 1, that is why these two elements are equal to 1.

(Refer Slide Time: 31:19)



And beta hat, here this matrix beta hat has two elements, beta 1 hat beta 2 hat, which is equal to r 1 y r 2 y. This r 1 y is the sample correlation between x 1 and the response variable y and r 2 y is the sample correlation between the second regressor and the response variable y.

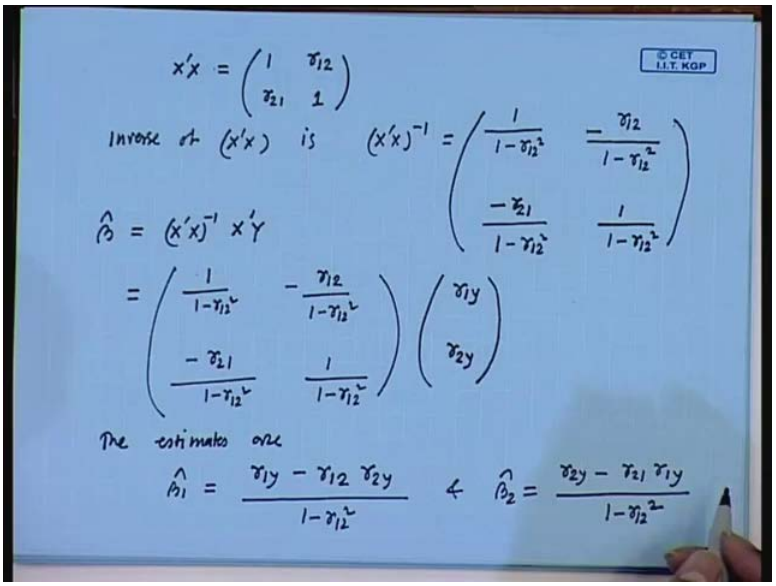
(Refer Slide Time: 32:18)



So, the normal equation we have is, r is 1 r 1 2 r 2 1 r 1 2 and r 2 1 are same, 1, beta 1 hat beta two hat, is equal to r 1 y r 2 y, where I already mentioned that, r 1 y or r 1 2 is sample correlation between x 1 and x 2, and r 1 y is the sample correlation between x 1 and y. Similarly, r 2 y is the sample correlation between x 2 and y, so what is this formula of r 1 y, r 1 y is equal to sum x i 1 minus x 1 bar into y i minus y bar, 1 to n, root over of S 1 1 S y y and r 1 2 is equal to sum x i 1 minus x 1 bar y i minus y bar, sorry this is equal to x i 2 minus x 2 bar, S 1 1 S 2 2.

So, may be more precisely, I should replace them by capital X capital Y, so you can check with the original X matrix now. This is the X matrix for the transformed data and then, you can check, why X prime X is equal to this, you know what is r 1 2 and what is r 1 y r 2 y, all these things. So, you can check that, why this element is r 1 2 that is, the sample correlation between the regressor x 1 and x 2. Now, to get beta 1 hat and beta 2 hat, we need to compute the inverse of this matrix, this matrix is called the correlation matrix now. This is the X prime X matrix, this one is, now it is called the correlation matrix.

(Refer Slide Time: 36:35)



So, what we have is that, we have X prime X is equal to 1 r 1 2 r 2 1 1 here and you can check that, the inverse of X prime X is X prime X inverse, which is equal to 1 by 1 minus r 1 2 square minus r 1 2 by 1 minus r 1 2 square minus r 2 1 1 minus r 1 2 square 1 by 1 minus r 1 2 square, it is not difficult to check that, this is the inverse of X prime X.

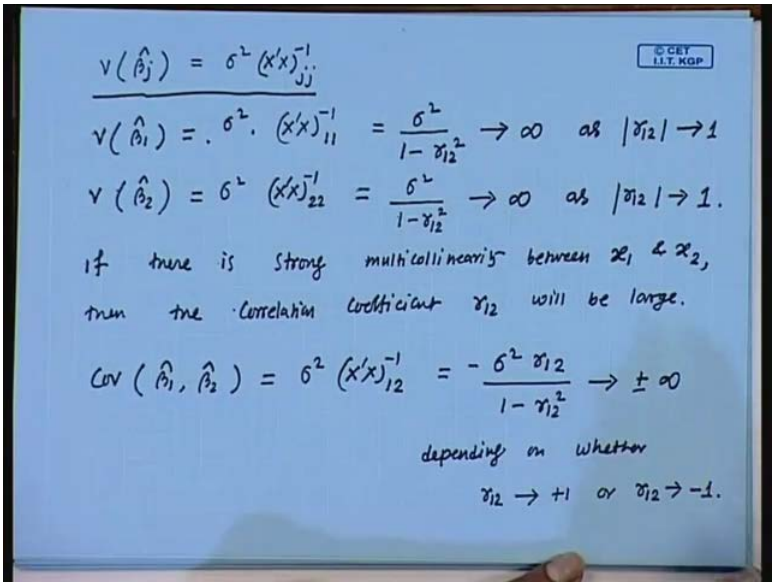
So, beta hat, the least square estimator of the regression coefficient beta hat is equal to X prime X inverse X prime Y.

So, which is equal to 1 by 1 minus r 1 2 square minus r 1 2 1 minus r 1 2 square minus r 2 1 1 minus r 1 2 square 1 by 1 minus r 1 2 square, X prime X inverse. X prime Y is equal to r 1 y r 2 y, so from here, the estimates are beta 1 hat equal to r 1 y minus r 1 2 r 2 y by 1 minus r 1 2 square. And beta 2 hat, so this is basically the least square estimate for the transformed data, in terms of the sample correlation coefficient and beta 2 hat is r 2 y minus r 2 1 r 1 y by 1 minus r 1 2 square.

So, what I said the problem with, if there is a problem of multicollinearity in the data then, that results in large variance and covariance of correlation coefficients. So, right now, we have the least square estimator of the regression coefficient beta 1 and beta 2. Next we will check, what is the variance of beta 1 and we are going to prove that, the variance of beta 1 and also the variance of beta 2, that tends to tends to infinity, as r 1 2 tends to 1.

That means, when there is a strong multicollinearity between the regressor x 1 and x 2 then, the variance of beta 1 hat and the variance of beta 2 hat are going to be infinity, so that is what, we are going to prove now. So, next we need to find the variance of beta 1 hat and variance of beta 2 hat in terms of the sample correlation coefficient r.

(Refer Slide Time: 41:14)



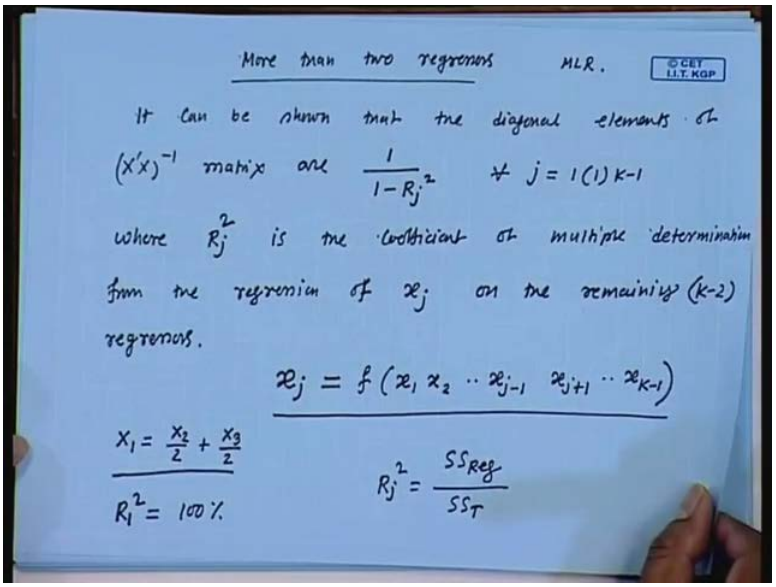
So, we know that, variance of beta j hat in general, is equal to sigma square X prime X inverse j j th element of this X prime X inverse. So here, from this formula, variance of beta 1 hat is equal to sigma square into X prime X inverse 1 1 th element. So, what is X prime X inverse 1 1 th element, the 1 1 th element is 1 by 1 minus r 1 2 square, so this one is equal to sigma square by 1 minus r 1 2 square. Now, so this is the variance of beta 1 hat and similarly, the variance of beta 2 hat is equal to sigma square X prime X inverse 2 2 th element, which is also equal to, 2 2 th element is also 1 by 1 minus r 1 2 square.

So here, the variance of beta 2 hat is also sigma square by 1 minus r 1 2 square, now if there is strong multicollinearity between the regressor x 1 and x 2 then, the correlation coefficient r 1 2 will be large. So, large means, the modulus value can be equal to 1, so when the modulus of r 1 2 tends to 1, the variance of beta 1 hat tends to infinity, so this tends to infinity, as r 1 2 tends to 1. So, this will tends to 1, when the regressor x 1 and x 2 are strongly correlated or you can say, when there is strong multicollinearity between x 1 and x 2 then, r 1 2 tends to 1.

Similarly, this quantity is also, the variance of beta 2 also tends to infinity, as r 1 2 tends to 1. So, also I said that, strong multicollinearity results in large variance and covariance, so what is a covariance of beta 1 hat and beta 2 hat, this is equal to sigma square X prime X inverse 1 2 th element. So, what is 1 2 th element or 2 1 th element that is, minus r 1 2 by 1 minus r 1 2 square, so this one is going to be sigma square r 1 2 minus 1 minus r 1 2 square.

And this also tends to plus minus infinity depending on, whether r 1 2 tends to plus 1 or r 1 2 tends to minus 1. So, this is the proof of I mean, we using two regressors in the model, in the multiple linear equation model, we illustrated how the strong multicollinearity results in large variance and covariance of the regressor coefficients, so the same. So, this illustration is using the two regressors in the multiple linear regression model, but this is also true, if you have more than two regressors in the model, so that we are going to mention now.

(Refer Slide Time: 47:15)



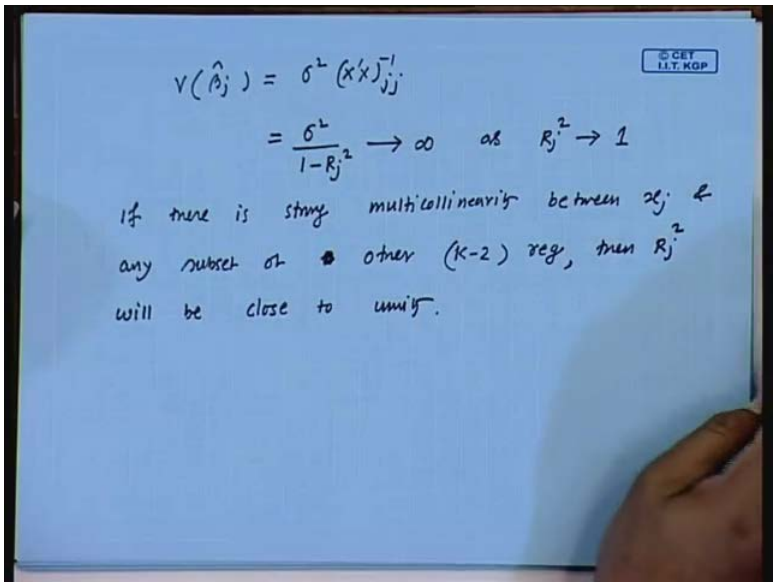
Suppose, you have more than two regressors in the multiple linear regression model then, it can be proved that, it can be shown that, the diagonal elements of X prime X inverse matrix are 1 by 1 minus R j square, for all j from 1 to k minus 1. So, I am talking about multiple linear regression model with k minus 1 regressors and of course, I need to define, what is this R j square. So here, before we had R 1 2 square, now r 1 2 has been replaced by R j, where R j is the coefficient of multiple determination from the regression of x j on the remaining k minus 2 regressors.

I think I need to explain this, here R j square is the coefficient of multiple determination, we know what is coefficient of multiple determination in multiple linear regression model that is, R square is equal to S S regression by S S total. And this parameter measures the proportion of variability in the response variable, that is explained by the regressor variable. But here, what I mean by R j square is that, so that is, we know what is R square, the coefficient of multiple determination, when we fit a regression model between y and the, between the response variable and the k minus 1 regressor variables. But here, this R j square is the coefficient of multiple determination from the regression of x j on the remaining k minus 1 regressors. Tthat means, here the multiple linear regression model is inbetween x j and remaining k minus 1 regressors. So here, x j is expressed in terms of the remaining regressors say, x 1 x 2 x j minus 1 x j plus 1 x k minus 1.

So, here the model is inbetween, trying to express x j in terms of the remaining regressors. So, if you can recall my first example today, there we have checked that, the X 1, if I am recalling correctly, X 1 is equal to X 2 by 2 plus X 3 by 2 and this one is true for all the observations we had, so this is what we want. And the coefficient of multiple determination for this example, this one is basically R one square and here, R 1 square is 100 percent, because X 2 and X 3 can explain 100 percent of the variability in X 1.

So, this R j, now I hope you understood that what is R j square, R j square is the coefficient of multiple determination for this regression model. So, you fit the model, you express x j in terms of the remaining regressors and then, you compute S S regression, for this model you compute S S total and you will get R j square.

(Refer Slide Time: 53:50)



So, since this is the diagonal element, now I can say that, when you have more than two regressors in the multiple linear regression model, the variance of beta j hat is equal to sigma square X prime X inverse j j th element and the j j th element is equal to sigma square by 1 minus R j square. So, this tends to infinity, as R j square tends to 1, R j square tends to 1 I mean, sometime we write, most of the time we write this in terms of percentage, so R j square is 100 percent means that, R j square tends to 1.

So, R j square tends 1 means, this will happen, if there is strong multicollinearity between x j and any subset of other k minus 2 regressors then, R j square will be close to unity. So, if there is a strong multicollinearity between the regressor x j and any subset of

the other remaining k minus 2 regressors then, R j square will be close to unity and the variance of beta j hat will be tends I mean, will tend to infinity.

So, this proves, just today we could manage to talk about only one effect of multicollinearity that means, only one problem due to multicollinearity. That says that, strong multicollinearity results in large variance and covariance of the regression coefficients. And we illustrated this point, both for the multiple linear regression model using two regressor variable in the model and then, in general. Next class, we will be talking about some other, I know there are several effects of multicollinearity in multiple linear regression model. So, we will be talking about those problems in the next class.

Thank you very much.