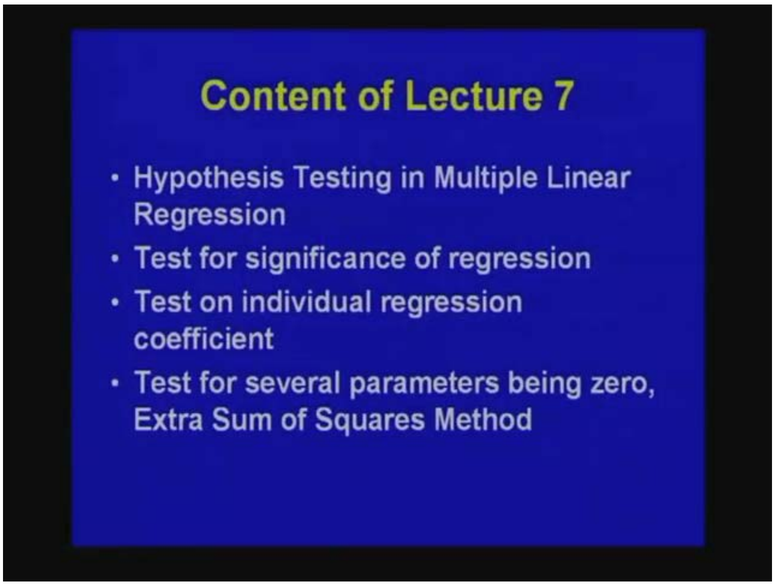
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 7
Multiple Linear Regression (Contd.)

This is my second lecture on Multiple Linear Regression and the content of today's lecture is basically hypothesis testing in multiple linear regression.

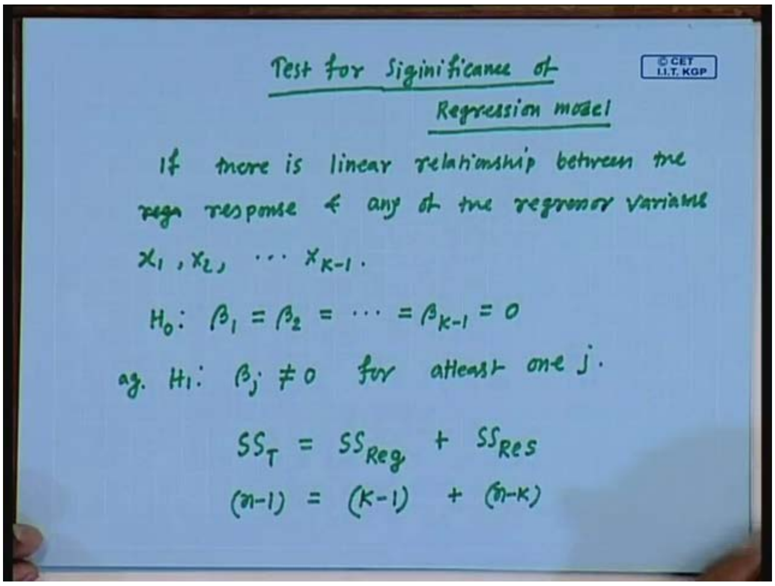
(Refer Slide Time: 01:06)



We will be talking about, how to test the significance of regression model and test on individual regression coefficient and test for several parameter being 0 using the extra sum of squares method. This extra sum of squares method is a very important technique in regression analysis. Anyway, in the last lecture, we have learned how to fit a multiple linear regression model. So, given a set of observations on response variable and say for regressor variable, we know how to estimate the regression coefficients using least squares method.

So, once you have a fitted model, the next important job is to test the significant of the model. So, by testing the significance of a multiple linear regression model, I mean whether there is linear relationship between the response variable and the regressor variable.

(Refer Slide Time: 02:59)

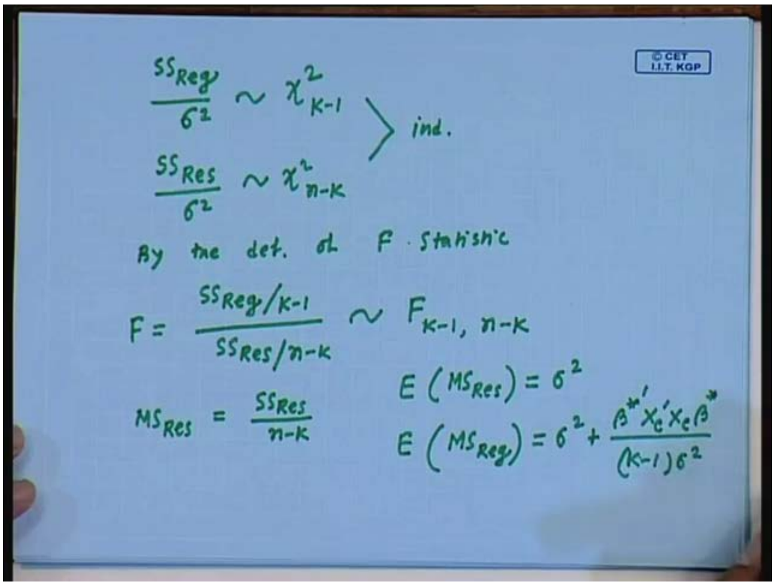


Test for significance of regression model, so as I told that, distinct significance of regression model it means, whether there is a linear relationship between the response and any of the regressive variables. So, this is to test, whether if there is a linear relationship between the response and any of the regressor variables X 1, X 2, X k minus 1.

So, this can be tested by testing the hypothesis that H naught, which is a null hypothesis, by testing this null hypothesis beta 1 equal to beta 2 equal to beta k minus 1 equal to 0 against the alternative hypothesis, that beta j is not equal to 0 for at least one j. So, the null hypothesis says that, there is no linear relationship between the response variable and any of the regressor variable and alternative hypothesis says that, no there is at least one regressor variable, which contribute significantly to the model.

So, this is the hypothesis we want to test, so the rejecting the null hypothesis here implies that, at least one of the regressor variables X 1, X 2, X k minus 1 contributes significantly to the model. So, here beta j is not equal to 0 means, the regressor variable X j contributes significantly to the model. So, to test these hypotheses, we will be taking the ANOVA approach. We know that, the total sum of square SS T, which is equal to SS regression plus residual. Now, we also know that, the SS T has a degree of freedom n minus 1, SS residual has degree of freedom n minus k and SS regression has degree of freedom k minus.

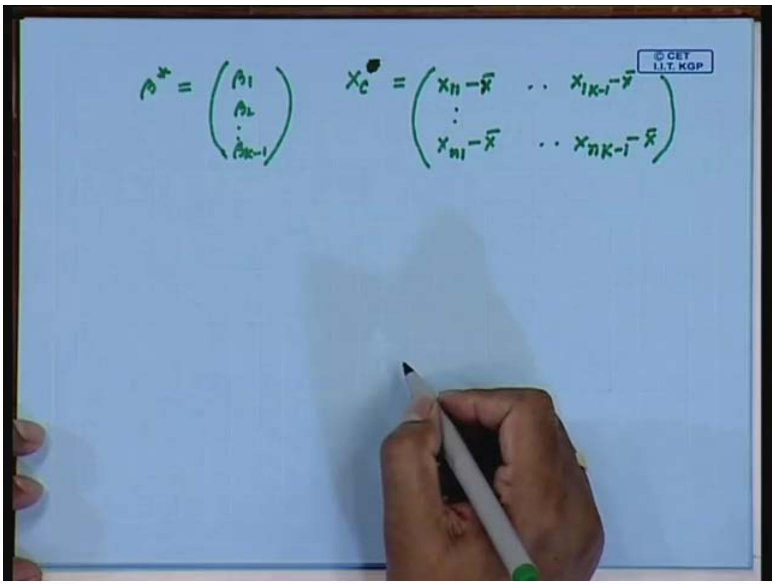
(Refer Slide Time: 08:08)



Now, SS regression by sigma square, this follows chi square distribution with degree of freedom k minus 1 and also we know that, SS residual by sigma square, this follows chi square n minus k and they are independent. Then from the definition of F statistics, by the definition of F statistics, F is equal to SS regression by k minus 1 by SS residual by n minus k. This random variable follows F distribution with degree of freedom k minus 1 n minus k and also it can be, of course we know that, MS residual which is basically SS residual by n minus k, this is unbiased estimator of sigma square.

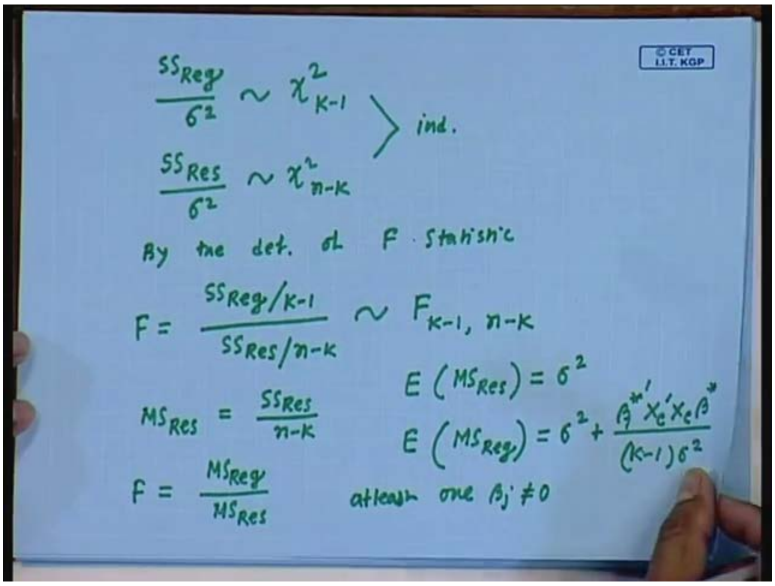
So, we know that, expected value of MS residual is equal to sigma square and also it can be proved that, the expected value of, this one is nothing but MS regression, expected value of MS mean squares regression, this is equal to sigma square plus beta star dashed X c prime X c beta star by k minus 1 sigma square. So, just I need to define, what is this beta star and X c.

(Refer Slide Time: 11:31)



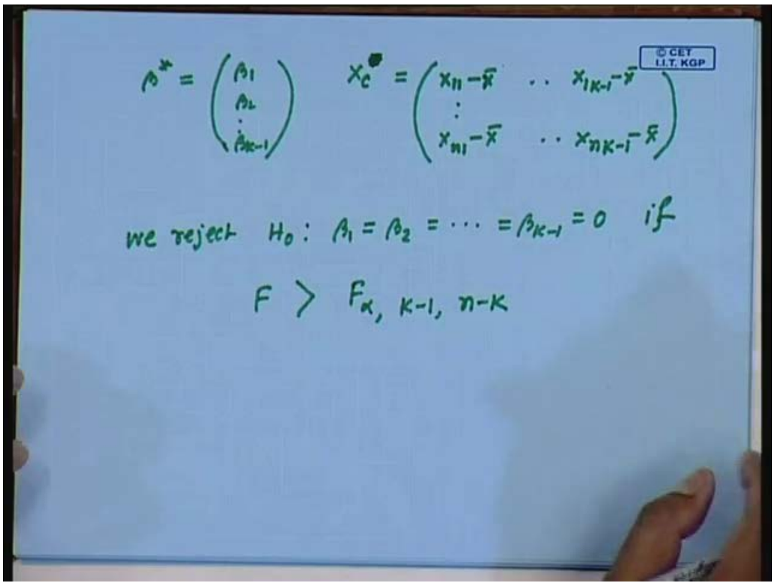
Basically, B star is equal to beta 1, beta 2, beta k minus 1 I mean, the beta vector was like beta naught beta 1 beta 2 beta k minus 1. So, beta star is obtained by excluding beta naught from beta and X c is equal to, it has also been obtained from the X matrix. But, this one is X 11 minus X bar that is, the mean of the observations I mean, all the regressor variables and X 1 k minus 1 minus X bar and X n 1 minus X bar X n k minus 1 minus X bar.

(Refer Slide Time: 12:54)



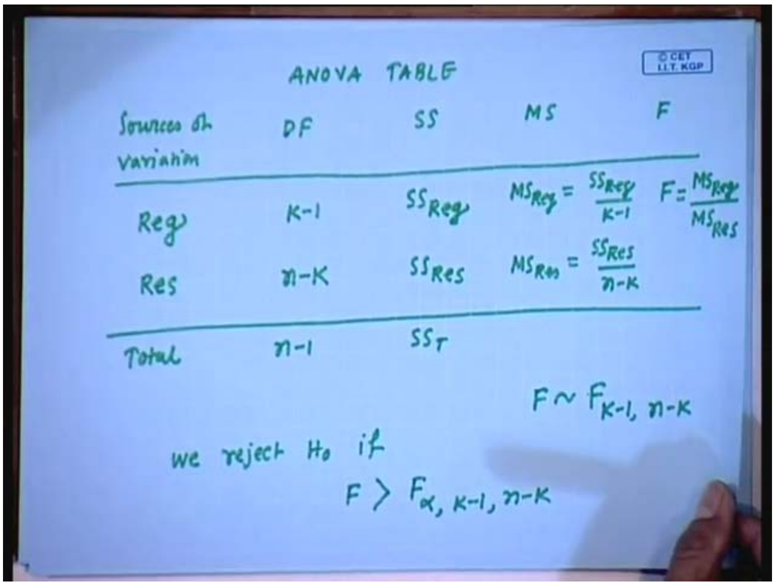
So now, look at these two expected value and these two expected values indicate that, if the observer will value of F is large, because see F is nothing but F is equal to MS regression by MS residual and here is the expected value of MS regression and here is the expected value MS residual. So, if the observed value of F is large then there is at least one beta, which is not equal to j. So, the higher value of F indicates that, at least one beta j is not equal to 0. If all the regression coefficient beta j are equal to 0 then this quantity is going to be 0 and F is going to be equal to 1. So, higher value of the observed value I mean, higher value of observed F indicate that, at least one beta j is not equal 0.

(Refer Slide Time: 14:39)



So, best on this, we reject H naught that is, the null hypothesis that says that, beta 1 equal to beta 2 equal to beta k minus 1 equal to 0. We rejected this null hypothesis if F value is high, high means I mean, F value is greater than F alpha and it has the degree of freedom k minus 1, n minus k, so this value you can get from the statistical table. So, now we just summaries the whole thing using the ANOVA table, so here is the ANOVA table for multiple linear regression.

(Refer Slide Time: 15:37)



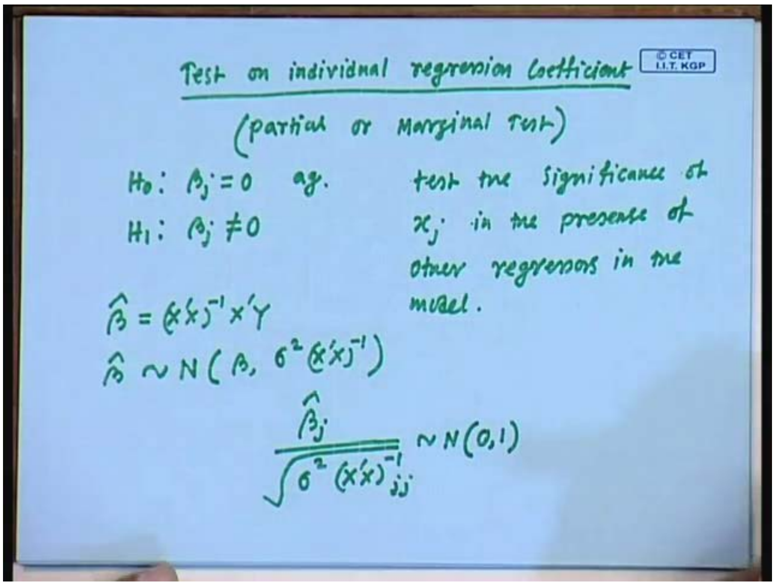
Source of variation, degree of freedom, sum of square, mean square and the F value. Sources of variation, it could be the variation due to, the variation that is explained by the regression, the variation that is remained unexplained that is, SS residual and this the total variation. The degree of freedom for this one is n minus 1, the degree of freedom for regression is k minus 1 and residual is n minus k and this is called SS regression, SS residual, SS T.

So, I have explained all these things, what is SS T, what is SS residual, what is SS regression in the previous lecture. So, MS regression is equal to SS regression by the degree of freedom k minus 1. Similarly, MS residual is equal to SS residual by the degree of freedom n minus k and here you have the F value, which is equal to MS regression by MS residual. And we know that, this F follows F distribution with the degree of freedom k minus 1, n minus k.

And we reject H naught, if F is greater than I mean, observed F is greater than F tabulated F alpha, k minus 1, n minus k. So, next I move to the test on individual regression coefficients, once you determined that your null hypothesis in the previous test is rejected. That means, there is linear relationship between the response variable and the aggressive variable that means, the null hypothesis is rejected means, there is at least one regressor variable, which has the significance contribution to the response variable.

So, once the null hypothesis previous state is rejected, we know that, there is at least one regressor, which has significant contribution to explain the variability in the response variable. Now, the next obvious question is, which regressor variable has significant contribution, so we need to test the regressor coefficient individually.

(Refer Slide Time: 20:28)

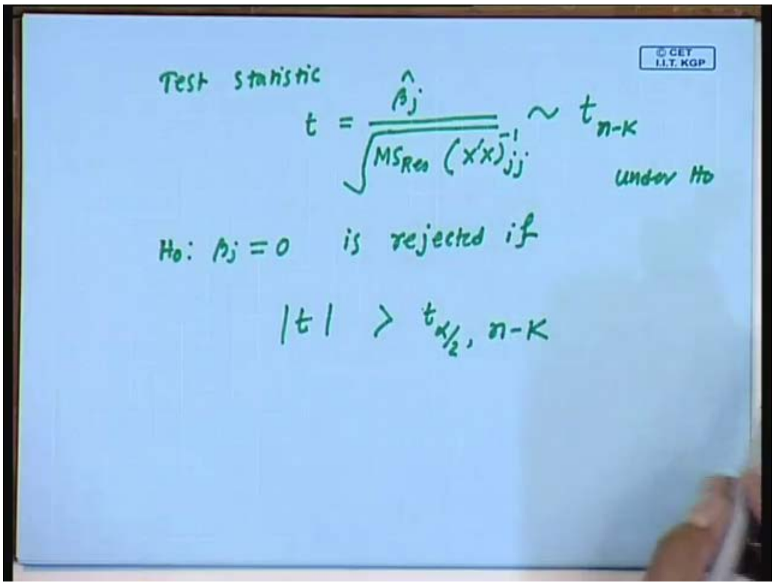


So, test on the individual regression coefficient, so it is also called partial test, partial or marginal test, previous is on global test, forgot to mention that. So here, once you determined that, at least one of the regressor variable is significant, so the next question, which one is significant. So, for this one, we test the hypothesis H naught which says that, beta j equal to 0 against the alternative hypothesis H 1 that is, beta j is not equal to 0.

So, basically this one and test this hypothesis test the significance of x j in the presence of other regressor in the model. So, how to test this hypothesis, that beta j is equal to 0 against the alternative hypothesis, that beta j is not equal to 0. We can go for, of course we know that, the unbiased estimator we got is beta j, which is equal to X prime X inverse X prime Y. And we know that, this beta hat follows normal distribution with mean beta and variance sigma square X prime X inverse, that is what I proved in the last class.

Now, from here, we can conclude that, beta j hat by sigma square, so this is the total I mean, this is the variance covariance matrix. But, here we are only concerned about beta j, so will be taking the j j th element here. So, sigma square X prime X inverse, you just take the j j th element. So, this one is nothing but this follows, you can say that this follows normal distribution with mean 0 and variance sigma square. Now, of course, see sigma square is not known, so if we replace the sigma square by MS residual then this variable or this random variable is going to follow t distribution.

(Refer Slide Time: 25:00)

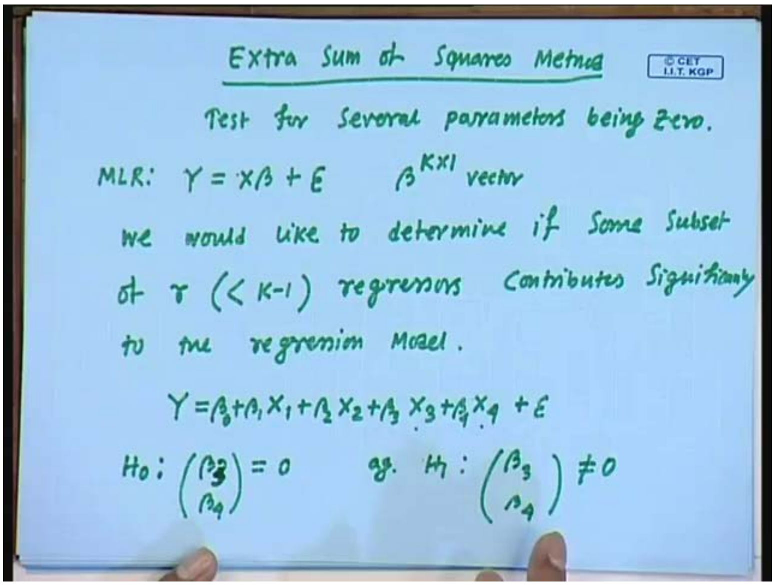


So, the test statistic for this testing is that, t equal to beta j hat by MS residual X prime X inverse the j j th element of this, this follows t distribution with the degree of freedom n minus k, of course this is under H naught. So, I did mistake here, this does not follow normal (0,1), this minus beta j this follows normal (0,1) and under H naught, this beta j is equal to 0. So, under H naught, you can say that, this random variable follows normal (0,1).

And of course, we are going to reject the null hypothesis, so H 0 which says that, beta j is equal to 0 is rejected if t value far from 0, if this t value is greater than t alpha by 2, n minus k. So, this is the critical region to test the beta j equal to 0, so the first step is that, you test whether the fitted model is significant. That means, whether there is any linear relationship between the response variable and any of the regressor variable by using the global test that means, the hypothesis we tested at the beginning.

So, if you see that, yes the test is significant that means, there is significant contribution of at least one regressor then you go further partial test. I mean, once you know that, at least one of the regressor variables has significant contribution to the response variable then you need to be determine, which regressor has the significant contribution. To determine that, you need to go further partial test, so next we will move to, we want to test for several parameters being 0.

(Refer Slide Time: 28:23)



So, this is the technique, here is called extra sum of squares method and this one has very important application in the regressor analysis. What to do here is that, we want test for several parameter being 0. So, what I mean by this is that, suppose you have multiple linear regression model say, Y equal to X beta plus epsilon. Here, this beta is k cross 1 vector, so it involves beta naught and k minus 1 regressor coefficients. Now, what you want is that, we would like to determine if some subset of r, of course r less than k minus 1, k minus 1 is the total number of regressors.

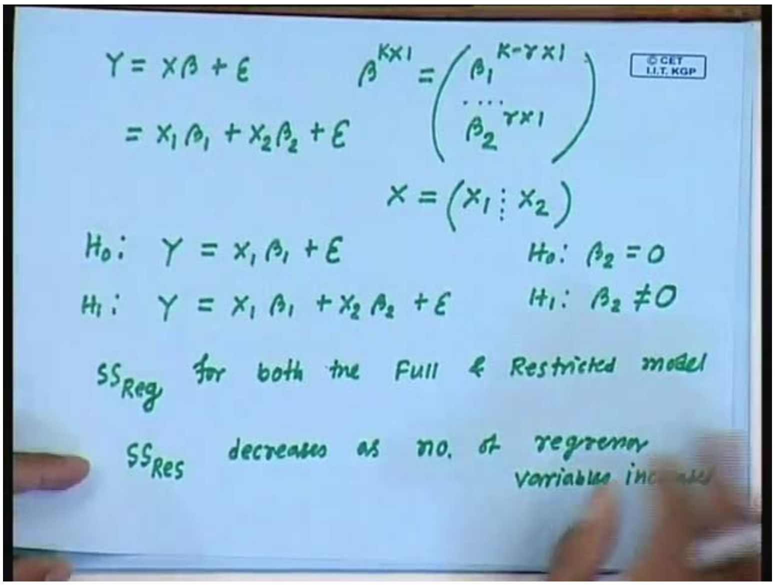
If some subset of r regressors contribute significantly to the regression model, so what I mean by this one is that, suppose you are given a problem. And in that problem, there are 4 regressor variables and 1 response variable. So, you need to fit the model like Y is equal beta naught plus beta 1 X 1 plus beta 2 X 2 plus beta 3 X 3 plus beta 4 X 4 plus epsilon. And after fitting the model, you feel that, some of the regressor variables are not significant, may be you believe that, X 4 and X 3, they are not significant.

So, what you want to test is that, whether beta 3 and beta 4 is equal to 0, again the alternative hypothesis is that, at least one of them is not equal to 0. So, even to test the significance of beta 3 and beta 4 or X 3 and X 4, the regressor variable X 3 and X 4 in the presence of X 1 and X 2. So, extra sum of square technique is use to test such hypothesis. Let me explain again, suppose you have 1 response variable and you have 4 regressor variable X 1, X 2, X 3 and X 4 and you have to fit a model like Y equal beta naught plus beta 1 X 1 plus beta 2 X 2 plus beta 3 X 3 plus beta 4 X 4 plus epsilon.

And after fitting the model of this form in a multiple linear regression model involving 4 regressors, you believe you feel that, X 3 and X 4, they are or for example, may be any subset X 1 and X 4, they are not significant, they do not have significant contribution to explain the variability in Y. So, what do you believe is that, you believe that these two say, X 3 and X 4, they are irrelevant for the response variable Y. So, to test this one significant I mean, to test this, whatever you believe in statistically, you need to test the hypothesis like H naught equal to beta 3 beta 4.

Even to test, this vector is going to be equal to 0 against the alternative hypothesis that (Refer Time: 34:28) beta 3 beta 4, against alternative hypothesis that beta 3 beta 4 is not equal 0. By not equal to 0 means, at least one of them is not equal to 0. So, this is what, we want to test and this type of hypothesis can be tested using extra sum of square method.

(Refer Slide Time: 34:59)



In general, how the model Y equal to X beta plus epsilon and my beta is k cross 1 vector, now I want to split this beta into two parts I mean, one is I call it beta 1 and other one is beta 2. So, this beta 1 is again a vector, it has k minus r cross 1 vector and beta 2 is a r cross 1 vector. And you believe that, the last r regressor variables are not significant for Y, using this partition similarly we can divide the matrix X also. X can be also participant to X 1 X 2 and then you can write this model as X 1 beta 1 plus X 2 beta 2 plus epsilon.

And the hypothesis we want to test is that, you think that this is enough, you think that Y equal to X 1 beta 1 plus epsilon is enough for Y. That means, what I want to mean by this one is that, here in fact analog thing is that, I am testing beta 2 equal to 0, so beta 2 is a vector, which involves r regression coefficient. So, against the alternative hypothesis H 1, that Y equal to X 1 beta 1 plus X 2 beta 2 plus epsilon, which in another word says that, H 1 is that beta 2 is not equal to 0.

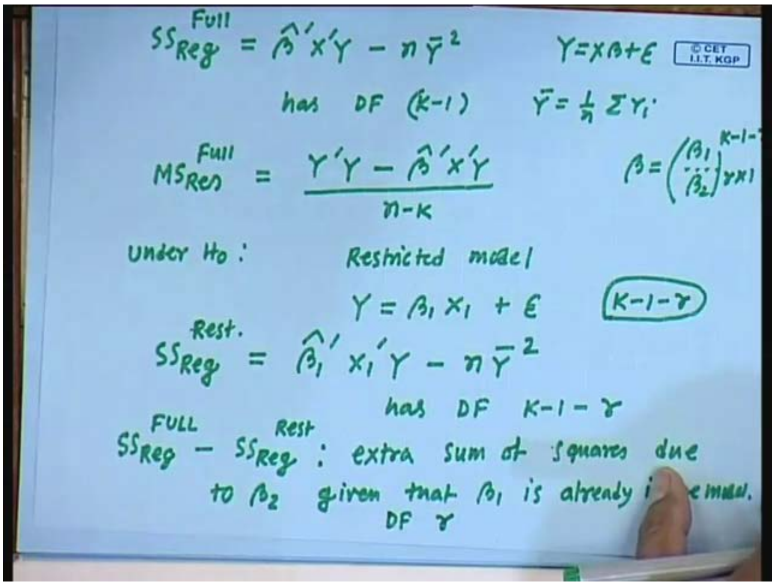
So, what you claim that, this we call this one the restricted model, you feel that this restricted model or the first k minus r regressors are enough to explain the variability in Y, you do not need the last r regressors. And the alternative hypothesis is now the last r regressors, r also significant to explain the variability in Y. So, to test this type of hypothesis, we use the extra sum of square technique. So, what we do here is that, we compute SS regression for both the full and restricted model.

So, this one is the full model, is involved all the regressors and this is the regressor model, which involve only k minus r minus 1 basically, k minus r minus 1 regressors. So, one thing you have to understand that, SS residual, this one always decreases as the number of regressor variable increases. So, this is the very intuitively, of course it is clear, it says that, the SS residual, this thing decreases as you increase the number of regressor variables, whether it is the newly added regressor variable is relevant for the response variable or not, it does not matter.

If you add one more, suppose you have the model with the k regressors, so if you add one more regressors say, if you make k plus 1 regressor variable in the model then SS residual decreases, but if the newly added regressor variable is significant or the very relevant for the model for the response variable then it decreases more, but if it is not that much relevant for the response variable then the SS residual decreases less. The same thing, since SS residual plus SS regression is SS total, which is fixed.

In another word I can say that, SS regression increases, as we increase the number of regressor variables. So, again the same statement, the SS regression increases more, if the newly added regressor variable relevant to the response variable, otherwise it increases less. So, this is the basic idea behind the extra sum of square technique, let me compute SS regression.

(Refer Slide Time: 42:57)



SS regression for the full model first, so SS regression full, we know that SS regression for the full model means, by the full model I mean, Y equal to X beta plus epsilon, so it has all the regressors. So, I know that, the SS regression for the full model is, beta hat prime X prime Y minus n Y bar square, you can refer previous lecture for this one and this has degree of freedom k minus 1.

Of course, Y bar nothing but 1 by n summation Y i and also we know that. MS residual for the full model is equal to Y prime Y, this is basically SS residual Y prime Y beta hat prime X prime Y by the degree of freedom, because I wrote MS residual. So, the degree of freedom is n minus k, this is for the full model. Now, under H naught that is, under the restricted model, so the restricted model says that, Y equal to beta 1 X 1 plus epsilon, so we do not have the last r regressors in this model.

So, under this restricted model, my SS regression, I said restricted, so this is the rotation SS regression under the restricted model, this is going to be beta 1. The same thing, I will just replace beta hat by beta 1 hat and X by X 1, X 1 dashed Y minus n Y bar square.

And this has, here you have not k minus 1 regressors, you have k minus 1 minus r regressors in this model, because we have removed r regressors from this model, so this has degree of freedom k minus 1 minus r.

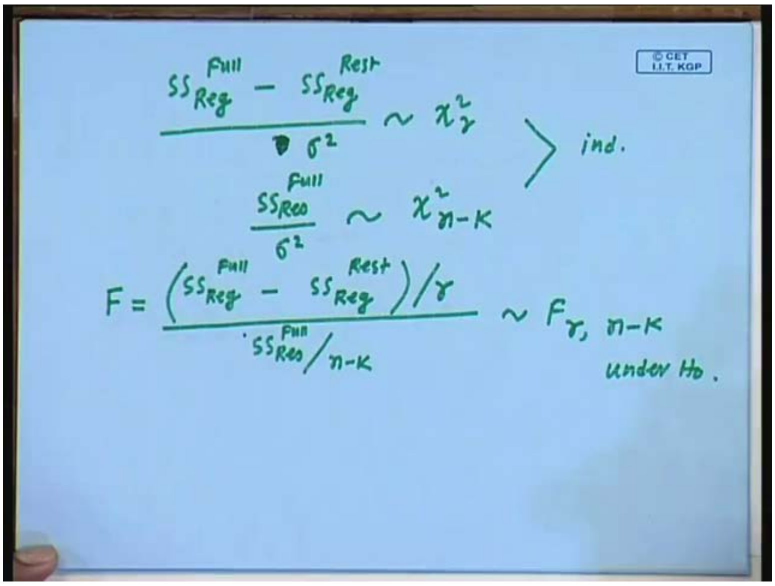
Now, I said that, SS regression increases as the number of regressor variable increases, so that means, the SS regression under the full model is greater than the SS regression under the restricted model. Because, the full model has more regressors variable compared to the restricted model. So, this one you compute SS regression full minus SS regression restricted, this is called the extra sum of square due to beta 2, given that beta 1 is already in the model.

So, this is very important, let me explain little bit, I said that, this is the extra sum of square due to beta 2. What is beta 2, beta 2 is the vector, beta we have splitted into two parts, beta 1 and beta 2. So, beta 2 is the regressor, beta 2 is the r cross 1 vector, so if beta 2 is associated with the regressor coefficient for those r regressor variables. This one is, this is regression for the full model, this is regression for the restricted model. So here, you have all the regressor variable, here you have the first k minus 1 minus r regressor variable.

So, if you subtract these from here, this will give you the extra regressor sum of square, I should say that. Extra sum of square is basically the extra regression sum of square due to the last k minus 1 minus r regressors. So, I hope you understood, so this is called extra sum of square due to the last k minus 1 minus r regressors, given that this one is on order r cross 1, this one is k minus 1 minus r regressors. So, this is that extra sum of square, this is the extra sum of square due to the last r regressors, given that the first k minus 1 minus r regressors are present in the model.

Now, what we can do is that, we can compute the degree of freedom of this one. And the degree of freedom of this one is, see this has degree of freedom k minus 1 and this has the degree of freedom k minus 1 minus r, so this has the degree of freedom r, this minus this one.

(Refer Slide Time: 50:36)

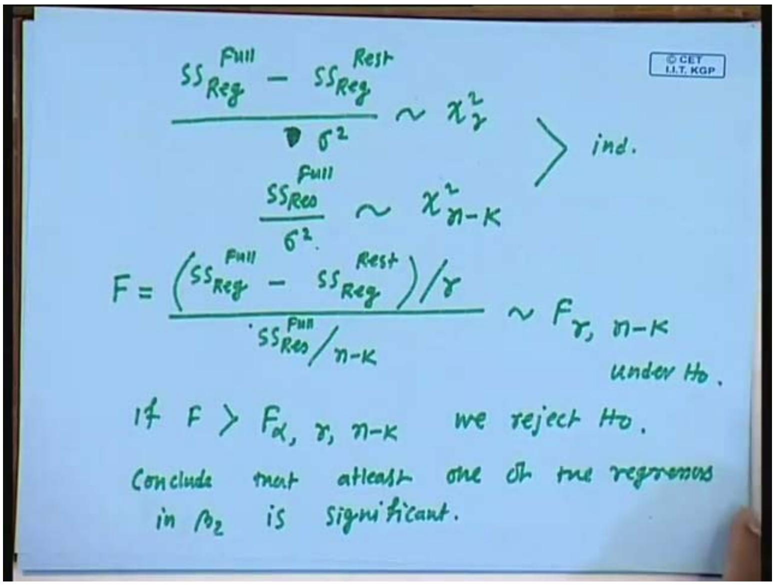


Now, SS regression for the full model minus SS regression for the restricted model, this is the excess sum of square and this has the degree of freedom r, this by sigma square follows chi square r, because this has degree of freedom r. And SS residual for the full model by sigma square, this follows chi square n minus k and you can check that, they are independent. Now, we are in position to compute the F statistic, which is equal to SS, this extra sum of square SS regression for the full model minus SS regression for the restricted model.

You divide this quantity by the degree of freedom r, by the definition of F statistics, so this follows chi square r, so this random variable by r, by this random variable by n minus k. So, SS residual full by n minus k, this thing follows F distribution with the degree of freedom r, n minus k under H naught. So, intuitively it is very clear that, this is the numerator, this portion is the extra regression sum of square due to the last r regressor variable.

And if this quantity is more that means, the last r regressors, they have significant contribution in SS regression. That means, they have significant regression, they have significant contribution to explain variability in Y. So, intuitively it is very clear that, if this quantity is large then we are going to reject the null hypothesis. So, the null hypothesis says that, you go for the restricted model, but alternative hypothesis says that, you go for the full model.

(Refer Slide Time: 53:44)



If this quantity is large then if F is greater than F alpha, r, n minus k then we reject H naught and conclude that, at least one of the regressors in beta 2 is significant. So, I hope you understood extra sum of square, this is very interesting and also very important, that is all for today.

Thank you very much.