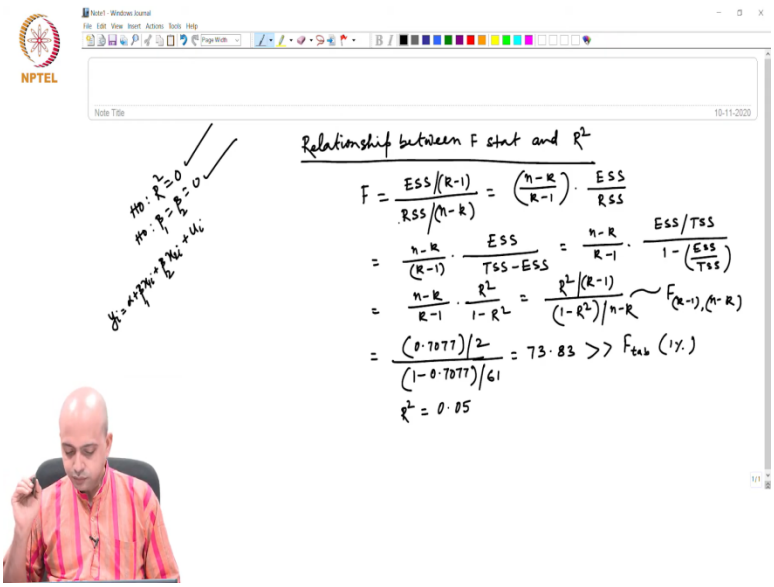
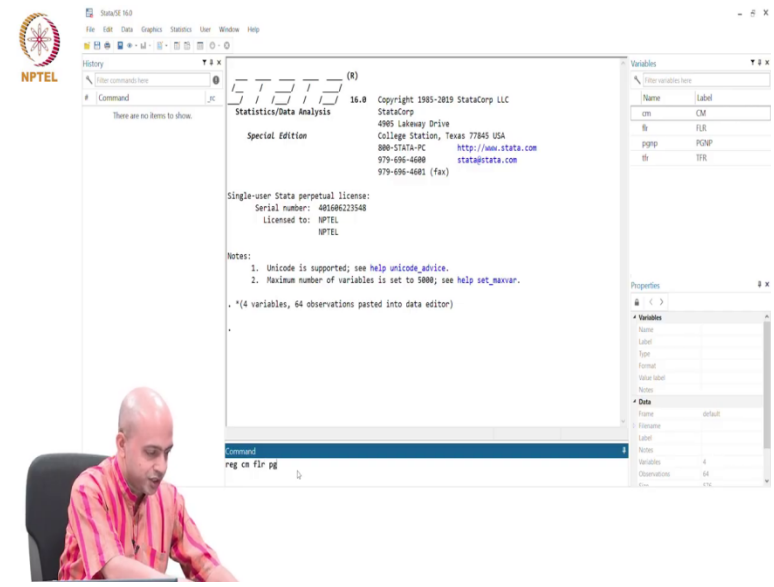
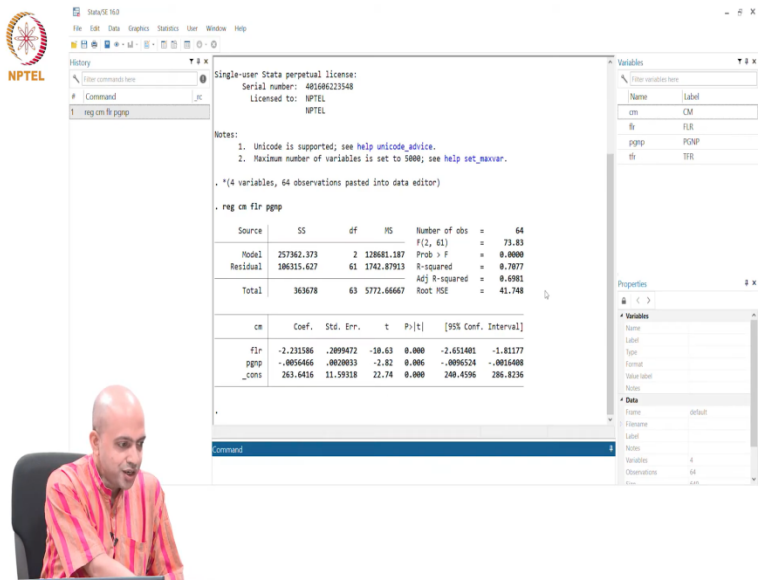
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras
Lecture 24
Multiple linear regression model and application of F Statistics Part - 3

(Refer Slide Time: 00:14)







So, welcome. We were discussing about the importance of F statistic in the context of multiple linear regression model. And then we said that F statistic is basically, helps testing the overall significance of the model which is not possible using the individual t statistic. Now, today we will see another application, another important application of the F statistic. So, let us try to understand the relationship between F statistic and the goodness of fit measure R square.

So, relationship between F statistic and R square. Now, F statistic if you recall, how we have defined F statistic? F statistic was defined as ESS by its corresponding degrees of freedom which is k minus 1. When you have k number of parameters to be estimated from the model, divided by RSS with its corresponding degrees of freedom which is n minus k. This is how we have defined our F statistic.

Now, we can do some algebraic manipulation and this will turn out to be (n-k) by (k-1) into ESS by RSS, equals to (n-k) into (k-1) into ESS, we can write as TSS minus, sorry this ESS is fine. The numerator is ESS. And this RSS we can write as TSS minus ESS. And then, if we divide the numerator and denominator by TSS, then what will happen? We will get (n-k) into (k-1), then we are dividing the numerator and the denominator by TSS. This would become (1-ESS) by TSS.

Now, ESS by TSS is R square. So, this would become (n-k) by (k-1) into R square divided by (1-R square) equals to, (R square/ k-1) into (1-R square)/ (n-k). Now, we have arrived at a relationship between F and R square and this particular measure, R square by k minus 1 divided

by 1 minus R square by n minus k will follow an F distribution with k minus 1 degrees of freedom for the numerator and n minus k degrees of freedom for the denominator.

Now, the question is, why we have derived the relationship between F and R square. From this expression, what we can understand, higher the value of R square, higher would be the value of F. And then, the calculated value of F would be greater than the tabulated one. That means this relationship between F and R square is quite useful to test one important hypothesis. What is that hypothesis we can say?

The hypothesis says, let us say our null hypothesis is R square equals to 0. That means the model does not have any significant explanatory power. If R square is 0, that means your model is not able to explain a significant portion of the total variation in your dependent variable, that is why this is the null hypothesis. Alternatively, we can write this hypothesis as, let us say that our model was yi equals to alpha plus beta1 x1i plus beta2 x2i plus ui, then alternatively, we can write the null hypothesis as beta1 equals to beta2 equals to 0.

So, that means the same overall significance of the model what we have derived earlier using the relationship between F and R square, what we can do? We can test either the significance of the R square itself and then significance of overall significance of the model. So, that is the beauty of this relationship between F and R square. You can test the statistical significance of even the R square also.

For example, what we are, as an example, what we are testing yesterday, the model, our model was child mortality rate was explained by female literacy rate, FLR and PGNP. So, if we write reg CM FLR PGNP, then this is our model. And what is the R square? The R square value is 0.7077. Now, if we put R square equals to 0.7077 here, this will come as 0.7077, then divided by k minus 1. What is the k here? k equals to 3, you have two explanatory variable and 1 constant one.

So, 3 minus 1 equals to 2. And then, 1 minus 0.7077 and that will be divided by n minus k. n equals to 64, k equals to 3, so this is 61. And if you calculate this, then this value will turn out to be, you will see that 73.87. And this calculated value you have to compare with the tabulated value at 2 and 61 degrees of freedom. And if you compare you will see that your calculated value is much greater than F tabulated, even at 1 percent level of significance.

So, that means we can reject our null that R square equals to 0 and also we can reject our null that beta1 equals to beta 2 equals to 0. Now, why this is useful? This is useful because sometimes when you are working with cross sectional data, your R square may turn out to be let us say, 0.05. So, you do not know whether this R square is a good fit or a bad fit. Now, people may ask you that you have estimated a model but your R square is only 0.05. How will you counter that?

Now, in that context, what you should do actually, you should try to find out the statistical significance of R square, because this R square you got out of only 1 sample from the n number of possibilities from the given population. So, if you put the R square value in this F expression, you can get one F statistic and then, that you can compare with the tabulated value and say even if the R square value is 0.05 mathematically, it is statistically still significant.

So, that means when you are working with cross-sectional data, it is quite likely that because of this heterogeneity across so many individuals, so many entities, your R square may turn out to be a low value. That is possible. So, what we should check actually in cross-sectional data, that whether the variables are individually significant, whether they are giving expected sign. And also, the statistical significance of R square.

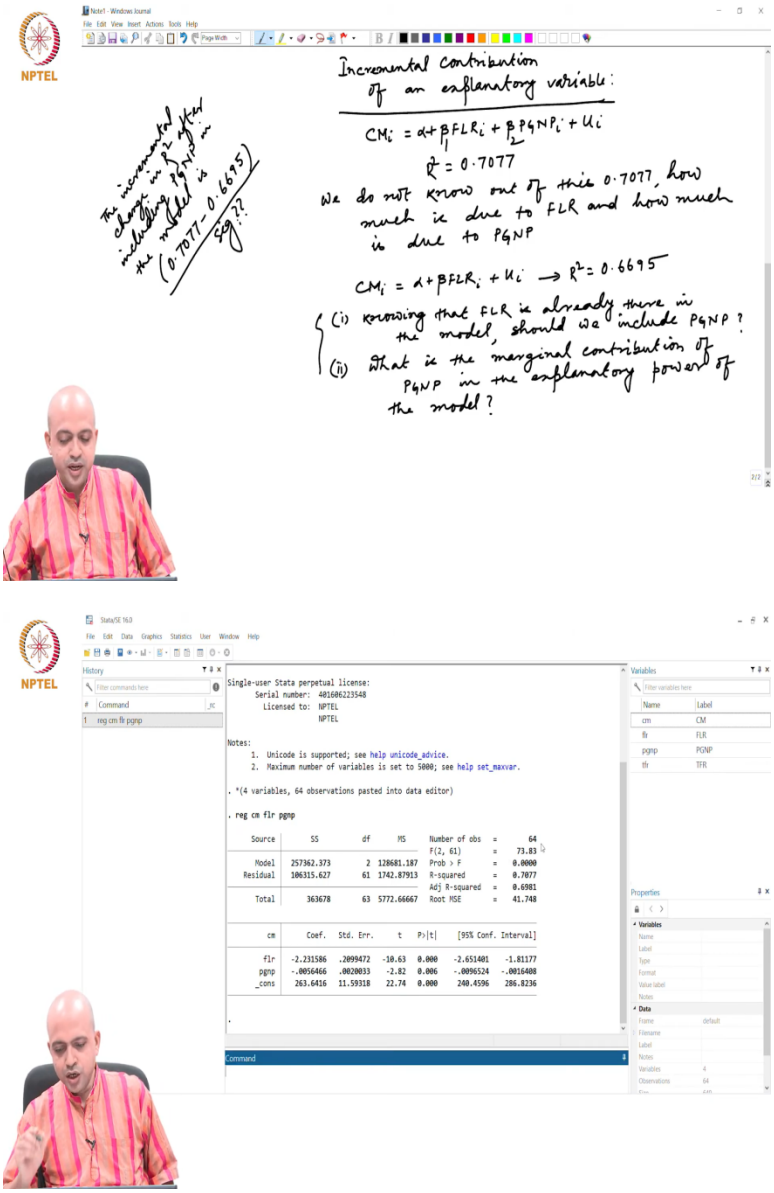
And since the significance of R square we are testing through F statistic, the moment you see that your model is overall significant, that means this F value, 73.87 and corresponding p value is 0.0007, so that means we will immediately understand, that means our R square also is significant.

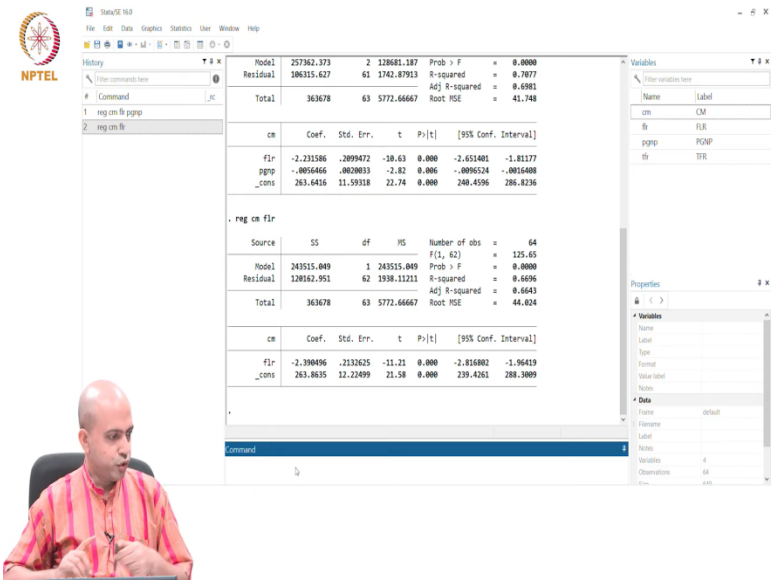
So, this relationship between F and R square will tell you two things; whether your model is overall significant and whether R square is also significant. That is quite logical, right? Because R square is nothing but the explanatory power of your model. How much the total variation in the dependent variable is explained by the explanatory variable you have included in your model. And that is nothing but the overall significance.

So, that means overall significance of the model and goodness of fit, they are like same thing measured by two different things, one is by F and another is by R square. That is why we could derive a nice relationship between F and R square. So, as long as your F statistic is significant, your model is overall significant. You will understand the R square is also significant. So, that means even though the R square value turns out to be low mathematically, by looking at the F

statistic and its significance, you can claim that my R square is statistically significant. So, that is what you can do by deriving the relationship between F and R square which you should keep in mind.

(Refer Slide Time: 12:10)





Then another important use of F statistic that we can derive using the F statistic, let us discuss. This is called incremental contribution of an explanatory variable. So, our model was CM child mortality rate we were trying to explain by beta1 FLR plus beta2 PGNP. Now, if you estimate this model, what is your R square? Your R square is 0.7077. Now, we do not know out of this 0.7077, how much is due to FLR and how much is due to PGNP.

So, that means we cannot allocate R square among different explanatory variables, that is what we learnt. Now, suppose we are including the variable sequentially. So, that means first we will estimate CM equals to alpha plus beta1 FLR plus ui, this is the model. So, we will run this model and then we will note down the R square. So, reg CM and then FLR, this is the model and your R square is 0.6696. So, from this model, R square is equal to 0.6695, this is the R square.

Now, many a times in applied research, applied empirical research, a researcher find a problem with knowing that FLR variable is already there in the model, should we include PGNP also? Because as we said yesterday, one extra explanatory variable if you add, that explanatory power of the model will definitely increase, but how much? And whether that incremental change in explanatory variable is more than the cost we incur. What is the cost? Cost in terms of lower degrees of freedom.

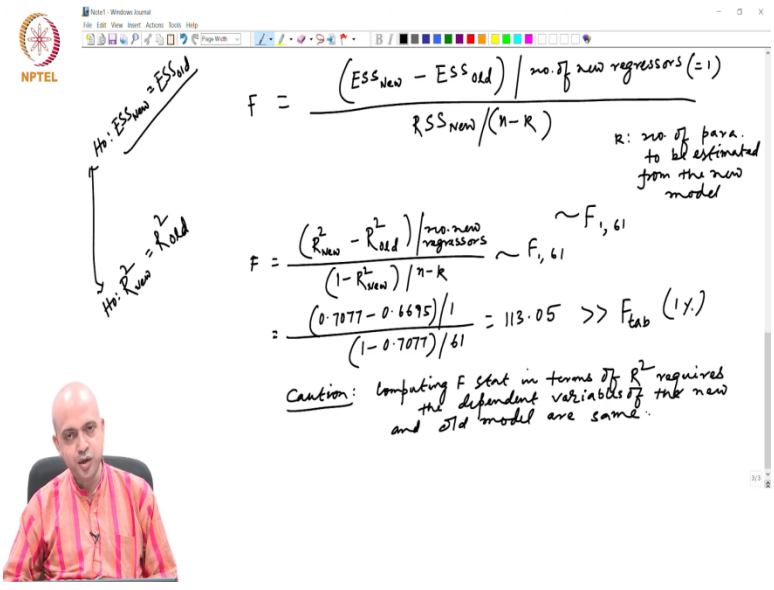
So, that means the question that I am trying to understand, knowing that FLR is already there in the model, should we include PGNP also? So, that means this is the quantity, this is the question; knowing that FLR is already there in the model, should we include PGNP? This is the question.

That means indirectly we can say that what is the marginal contribution of PGNP in the explanatory power of the model? That is why, what we want to know. These are the things we want to know.

So, now if we run this model, the R square is 0.6695 and if we add PGNP, then the R square increases from 0.6695 to 0.7077. So, what is the incremental change in R square? The incremental change in R square after including PGNP in the model, is 0.7077 minus 0.6695, this is the incremental change. Now, the question is whether this value is statistically significant or not, that is what we want to know. Is this clear?

So, when you have only FLR in the model, your R square is 0.6695. Now, knowing the fact that FLR is already there in the model, we are trying to add PGNP. And when PGNP is added, the R square improves from 0.6695 to 0.7077. So, that means marginal contribution of PGNP in the explanatory variable, explanatory power of the model is 0.7077 minus 0.6695. So, the question here is whether the difference between these two is statistically significant or not, how will you do that? That is also possible to get using an F statistic.

(Refer Slide Time: 21:04)



How will you do that? So, if you add extra variable, then what will happen? Your explanatory power of the model will increase. Let us say that this is denoted by ESS new compared to the ESS what you were having earlier. Let us say that is ESS old. And then, you divide this by number of new regressors. So what I am saying number of new regressors.

And then, your dependent, your numerator should be, what should be your denominator? Dominator should be your RSS new divided by n minus k, where k is actually number of parameters to be estimated in the new model. And this is how the F statistic is defined. Here, what we should write? k is basically number of parameters to be estimated from the new model.

Now, what is the logic? And this will follow an F distribution with here, the number of new regression equals to 1, so we will follow 1 and n minus k is 61 because k equals to 3. What is the logic of this? Here, the null hypothesis what I am saying that ESS new equals to ESS old. That means, what is our claim? Our claim is that PGNP brings significant contribution in terms of explanatory power of the model.

That is our claim. Knowing the fact that FLR is already there in the model, we are trying to add PGNP. Why we are trying to add PGNP? Because we hypothesize that PGNP brings significant explanatory power in the model. If that is your claim, that means you are saying that new explanatory power which is ESS new, is significantly higher than the old explanatory power which is ESS old. And if you nullify that claim, that means ESS new should be equals to ESS old. That is my null hypothesis.

But here, you look at the way we have constructed the test statistic, it is difference between ESS new and ESS old. So, that means higher the difference between ESS new and ESS old, higher would be the value of F. And higher the value of calculated F, higher would be the probability that it is actually greater than the tabulated value. And higher is the probability that F is greater than the tabulated one, we can reject our null hypothesis and we can say that yes, this ESS new is actually significantly higher than ESS old, that means we should add PGNP also in our model.

That is the logic. And this same expression we can write in terms of R square also, and then, in terms of R square, this would become R square new minus R square old divided by number of new regressors divided by 1 minus R square new divided by n minus k, that also follows F distribution with 1 and 61 degrees of freedom.

Now, why I am defining this in terms of R square? Because ESS, value of the ESS is huge, you will have difficulty in calculation. That is why we are putting the same expression in terms of R square also, easy to calculate. So, what is your R square new? R square new is 0.7077 minus

0.6695 divided by 1. And then, 1 minus 0.7077 divided by 61 which is actually equals to, what should be the value of this? The value should be equals to 113.05.

And this is your calculated F value. Since the value is 113, even without looking at the table also, we can understand that yes, the PGNP is actually statistically significant. So, that means my here, the null hypothesis is R square new equals to R square old, same thing. So, ESS new equals to ESS old is equivalent to saying R square new equals to R square old. So, you can since the value is higher, 113 you can say that this would be significantly greater than the F tabulated even at 1 percent level of significance.

So, this way what we can do? We can actually test the marginal contribution of a variable, when you are to decide whether a new variable should be added or not? As I said earlier, the new variable will come, what would be the benefit of adding a new variable? The benefit is in terms of extra explanatory power in the model which is in terms of the ESS or R square. And what is the cost? Cost is basically losing degrees of freedom.

That is why see, in this model here, in the denominator, we have n minus k, the expression is adjusted for degrees of freedom. The more variable you have, k would be more and you have to divide this by more 60, I mean divide it by n minus k. So, that is some way or the other, you are actually adjusting with degrees of freedom. So, once you get your F value, then compare this F value with a tabulated one and then you can say that yes, my calculated value is greater than the tabulated one and so I can reject my null hypothesis. Either ESS new equals to ESS old or R square new equals to R square old.

But one thing you have to keep in mind, whenever you are constructing F statistic, in terms of R square, one cautionary note, so I am giving you one cautionary note. Computing F statistic in terms of R square requires the dependent variable of the new and old model are same. So, whenever you are computing F statistic in terms of R square, because R square, here we are actually taking the difference, we are comparing R square new with R square old, so that means what is R square?

R square is that percentage of your explanatory, dependent variable which is explained by the independent variables. Now, this type of comparison is meaningful only when your R square new is equals to R square old. So, that means when your explanatory variable is different, in one case

it is CM and let us say in another model it is log of CM, then we cannot actually compare the two R square. Because the R square by definition, it says what is the percentage of total variation in y, dependent variable, explained by your model.

If the dependent variable itself is different, then we cannot compare 2 models in terms of their R square. So, whenever you are computing any statistic, using R square, a cautionary note is that we must ensure that the dependent variables of two competing models are same. If they are not same, then we need to define the F statistic either in terms of ESS or RSS. Because ESS and RSS is always comparable, not the R square, that is one cautionary note we have to keep in mind while computing the F statistic in terms of R square.

So, with this we are closing our discussion today. So, in our next class, we will again discuss the other hypothesis testing in term, using the F statistic. F statistic has several alternative applications in the area of hypothesis testing, particularly in multiple linear regression models and those important hypothesis testing we will discuss in our next class. Thank you.