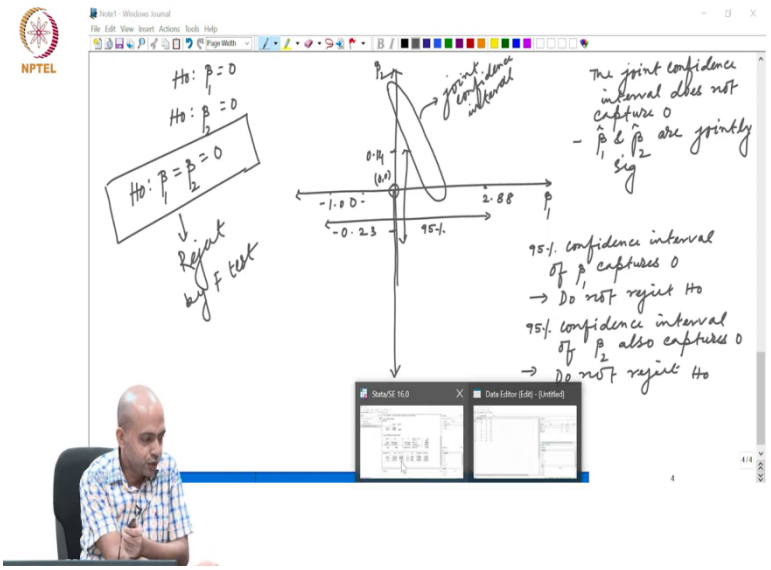
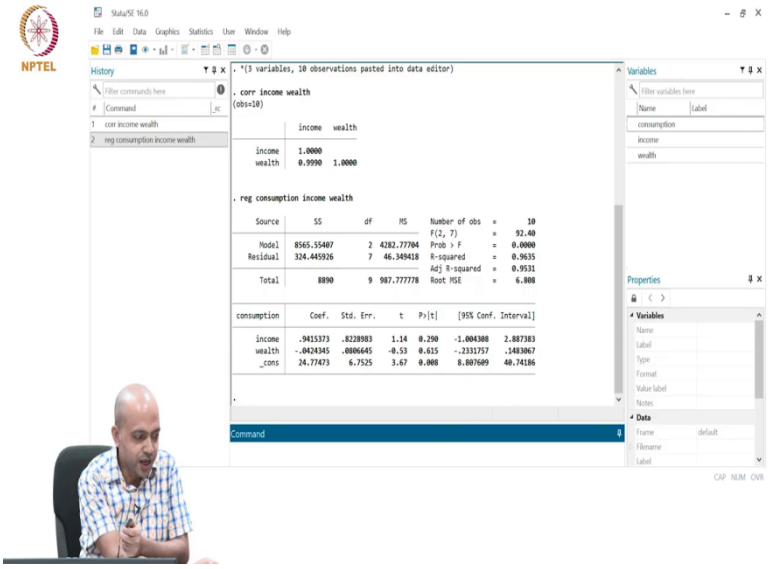
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 47
Relaxing the assumptions of CLRM-Multicollinearity and Autocorrelation Part – 4

(Refer Slide Time: 0:15)

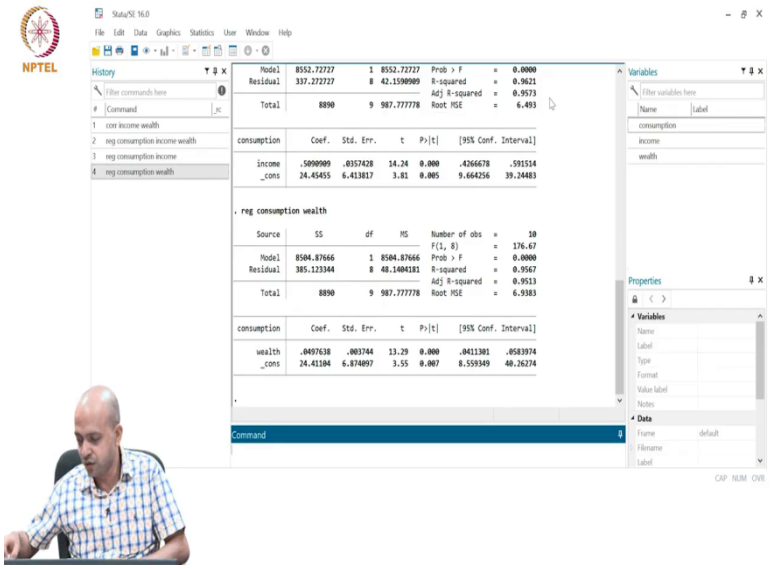




So, the joint confidence interval does not capture 0. So, beta 1 and beta 2 hat are jointly significant and that is why we got overall significance. So, that means, the joint significance if you test that means, beta 1 equals to beta 2 equals to 0 this test if you formulate this null hypothesis, they that you reject. That you can reject by F test. Why? Because look at F value.

Your F value shows 92.40 and it is highly significant. So, that is the reason we can say that the joint confidence interval looks like this and that does not capture the 0 value. That is why we are now able to justify why it is not possible to get a 1 to 1 mapping from t to F. That means, even though variables are individually insignificant by your t statistics, they are highly significant combinedly by F test because you are rejecting this hypothesis which is beta 1 equals to beta 2 equals to 0. So, in presence of multi collinearity, instead of using individual t statistic for inference making, we should actually look at the overall significance of the model by F statistic.

(Refer Slide Time: 2:56)

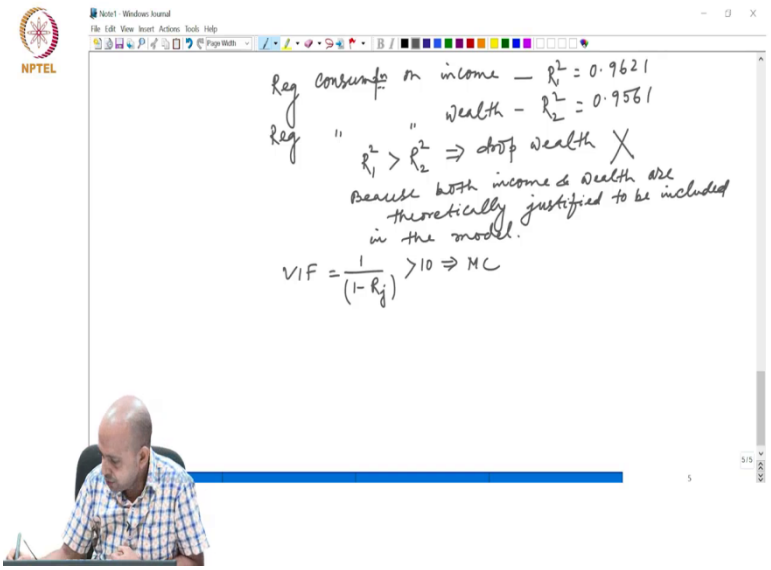


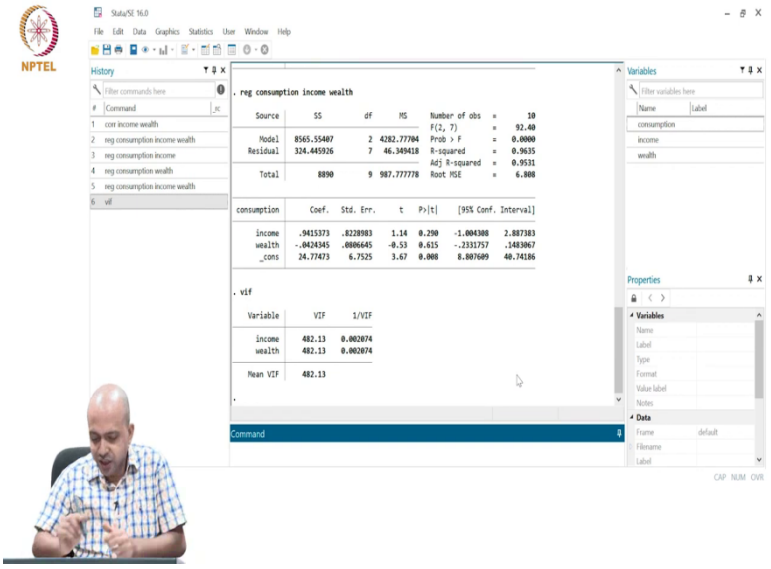
And then when you go back to your data once again suppose you are thinking that let me regress consumption on income alone. Now see the variable which was insignificant earlier is now highly significant. Similarly, and what is the R square? The R square is 0.9621 and if you regress consumption on wealth, see the wealth is also significant at one percent level because your p value is this. And what is the R square? 0.9567.

Now you might be thinking as a novice econometrician, generally what they do? They try to construct two different regression models. One with income another one is wealth using income as explanatory variable first, they will note down the R square and the R square is 0.9 621. Then they will regress consumption on wealth and note down the R square-0.9567. Since R square is high in the first case, they will drop the wealth variable, thinking that that is the best model.

So, what they do? They try to maximise, get the maximum R square from the model, they take the model which keeps maximum R square and then they drop the other variable. But then what I said, this approach is totally wrong because both income and wealth they are theoretically justified to be included in the model. They are theoretically justified to be included in the model. So, that is why even though the first model gives more R square when income is used as explanatory variable.

(Refer Slide Time: 5:48)





So, that means, what you do you regress consumption on income and what is the R square? R square equals to 0.96, so 0.9621 and then you regress consumption on wealth and your R square equals to 0.9561. Let us say this is R1, this is R2. So, R1 square is greater than R2 square and then you may end up with drawing this type of inference that drop wealth. But this

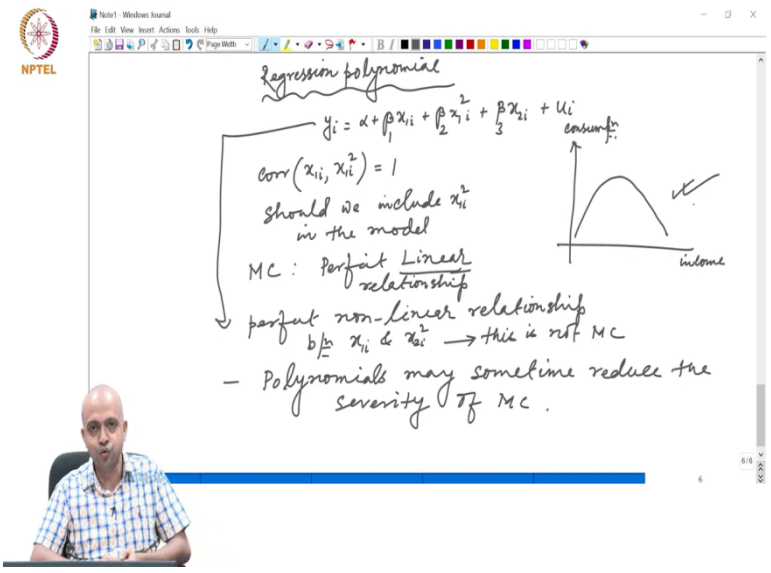
approach is wrong. Why? Because both income and wealth are theoretically justified to be included in the model; so, you have to increase your sample size, there is no other alternative in this particular case.

And as I said how will you detect multicollinearity? From the result what you can see that there is high R square with if you regress this reg consumption and then income and wealth so that means there is a high R square 0.9635 but all the variables are insignificant with unexpected sign that itself gives you some kind of indication that your data is actually suffering from multicollinearity problem.

After estimating the model if you VIF command, variance inflating factor and what I said, that variance inflating factor, the rule of thumb says that if it is greater than 10 then that gives you the multicollinearity problem, that particular variable. Look at the VIF here, both income and wealth, the VIF is 482. So, you can understand the severity of the multicollinearity here. But the rule of thumb says only greater than 10, here it is 482.

So, from this what we can say that VIF actually equals to 1 by 1 minus R square j should, if it is greater than 10, then that shows multicollinearity and here it is 482, so from this also we can understand the severity of multicollinearity. Now, at the end we will conclude our session by giving another example. Suppose in your income consumption example we are interested to see is there any non-linearity between income and consumption that means, it may so happen that as income grows, consumption will initially increase but after that it may go down.

(Refer Slide Time: 10:59)



So, that means, let us say that this is I would say that regression with polynomial that means, we are doing this type of regression. y i equals to alpha plus beta 1 x 1 i plus beta 2 x 2 i square plus beta 3 x 3 i square, x 3 i, this is let us not x 2 i, rather this is x 1 i square and this is only I have two variable income and wealth, x 2 i and plus u i. So, I am hypothesizing a an inverted u shape relationship between this is let us say income and this is let us say consumption. So, as income increases, consumption will increase initially and then it will come down.

Because our micro-economic theory says the individual may suffer from diminishing marginal utility. As income increases, I will at the initial stage I will keep on increasing the consumption of that a particular unit but after that it may goes down, it may go down. So, let us say that that is the assumption what we make. So, when you hypothesize a non-linear relationship either inverted u shaped u shaped.

In this example we are just making an assumption that this might be the relationship. So, what you have to do? You have to include the variable in its level form, you have to include the variable in its quadratic form. So, that is why I have included x 1 i and x 1 i square. Now, if you look at the pair wise correlation, what would be the pair wise correlation between x1i and x1i square? You can understand since x1i square is the square of x1i there is perfect collinearity, so this is almost; this is equals to 1, perfect collinearity between x1i square and x 1i.

Then the question is should we include x1i square in the model? Because if you include, then there is multicollinearity problem, is not it. So, that means, even though our objective is to check some kind of non-linearity between consumption and income, the moment you include you may think that we should not include the x1i square in the model because of high degree of correlation. Then what is the solution? The solution is we have to think carefully.

What does multicollinearity indicate? Multicollinearity says perfect linear relationship but it does not say that we should not have non-linear relationship also. Here what we are having in this particular case, in this model, we are getting perfect non-linear relationship between x1i and x2i square. So, that means this is actually not multicollinearity. So, that is why never ever hesitate to include square of a variable in your model if your objective is to check non-linearity between the dependent variable and any of your independent variable.

Because perfect non-linear relationship is not multicollinearity. Multicollinearity does not rule out the possibility of having this type of perfect non-linear relationship and that is why

when you include this type of model, when you include the consumption income square in your model, then there is absolutely no problem in estimation. There is no problem in estimation because this is after all a non-linear relationship.

Rather, sometimes econometrician they say that if you include polynomial in your model, the severity of multicollinearity actually goes down. That is some suggestion. So, polynomials may sometime reduce the severity of multicollinearity. This is also another solution to multicollinearity. No problem of including multinomials in the model fearing about multicollinearity if your objective is to check the non-linearity among variables.

So, with this we are closing our discussion on multi collinearity. So, the dataset what I have used for empirical demonstration is table 9 point 5 from your textbook income and consumption, Gujarati's textbook. So, basically what we have discussed? We discussed about what is multicollinearity? We said that this is a perfect linear relationship among variables and then what are the consequences of multicollinearity? We said that even though unbiasedness and consistency property are still maintained in the presence of multicollinearity, the efficiency property that means, minimum variance property is disturbed.

And why this is so? Because multicollinearity generates a variance inflating factor; so, your variance get inflated, standard error get inflated as a result of which t statistic defined as beta hat divided by standard error of beta hat gets inflated; when standard error gets inflated then t statistics goes down and variables, explanatory variables appear to be artificially insignificant. So, we should not relay much on t statistics in presence of MC because you get overall significance of the model by the F statistic. That is the consequence.

And the detection also high degree of high association, a high value of R square with many insignificant variable that is the classic symptom of multicollinearity also by which you can detect the presence of this problem and then we have also discussed about several remedial measures. We said that transformation of the variable by taking first difference or in ratio form as one solution or you may increase the sample size because multicollinearity is a sampling problem.

If you feel that you can increase the sample size, then you have to go for that and lastly we also talk about the importance of dropping the variable. Actually when the variables are theoretically justified we cannot drop any of the variables and include the other one depending on which particular variable gives you maximum R square, that approach is wrong

when the variables are theoretically justified. So, what we need to do actually? We need to increase the sample size.

And lastly we talk about regression with polynomials. So, that means, if your objective is to check non-linearity among dependent and any of the independent variables, you must include the polynomial without fearing about multicollinearity problem because multicollinearity does not say that perfect non-linear relationship should not be there, it only says perfect linear relationship. It is not about non-linear relationship that is why we should include that and that polynomials may sometime reduce the severity of multicollinearity.

So, with this we are closing our discussion on multicollinearity and in our next class we will discuss about relaxing some other assumption out of those ten and we will see what are the consequences, again how to detect that problem and the remedial measures that we will discuss in our next class.

Thank you.