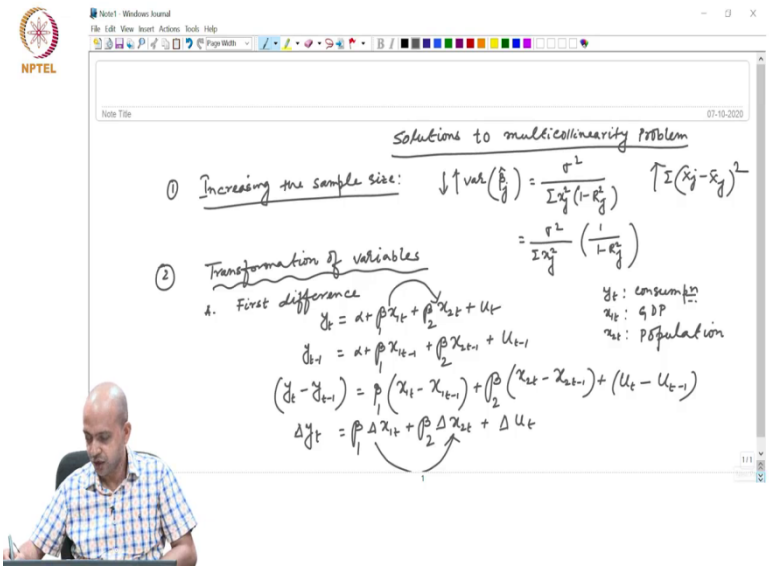
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Relaxing the assumptions of CLRM-Multicollinearity and Autocorrelation Part – 3

(Refer Slide Time: 0:15)



This is what we are discussing and if you recall, at the end we were talking about the solution by increasing your sample size because after all as we discussed yesterday that multi collinearity is mostly a sample problem. And what basically happens if you increase the sample size? If you look at the variance of beta hat j, beta hat j equals to sigma square divided by summation xj square into 1 minus R square j.

So, that means, this we can say that sigma square divided by summation xj square into 1 by 1 minus R square j. There are two components and in presence of multi collinearity what happens? This is actually your variance inflating factor if you recall, 1 by 1 minus R square j. So, this variance inflating factor in presence of multi collinearity will give you some upward shock in the variance.

But if you increase the sample size, the summation xj square which is nothing but sum of Xj minus Xj bar whole square. So, generally, as you increase the sample size, this also increases. And if this increases, that gives you a downward shock to the variance of beta j. So, that means that gives some counter impact on the VIF and as a result of which severity of the multi collinearity goes down as you increase the sample size.

The other thing also that happens when you increase the sample size, as I told you earlier that suppose you are working on a consumption function which is basically a function of income and wealth, so what basically you need in your sample; the two variables income and wealth not to be correlated. Because while specifying the model, the theory suggest in the population these two variables are not correlated.

So, when you get a sample, as you increase the sample size, there is a higher probability that you will collect information from those individuals in your sample, there are individuals with high income but low wealth, at the same time high wealth but low income. So, this two things you can achieve by increasing the sample size.

Now, today we will talk about another solution, solution number 2 but please keep in mind all the solutions what we are talking about they have some benefit, some cost as well. So, we do not know which particular solution will work in which particular context that is why no particular measure can give you a universally true kind of solution. So, that is the reason we are discussing so many alternatives. Otherwise, we would have stopped our discussion only by here by increasing the sample size.

Because increasing the sample size also has its own cost, it is not always possible to collect a bigger sample without involving your so much of time, energy even in terms of monetary costs. If it is possible, you can increase the sample size and your severity of the multi collinearity problem will come down. If it is not possible to always increase the sample size, we need to think about the other ways by which you can actually reduce the severity of multi collinearity.

And the second solution is the transformation of variable and here we will first discuss two things, within the transformation of variable, this is 2a which says that you can reduce the severity by taking the first difference. How, we will explain. Let us say that your model is yt, you have a time series data on alpha plus beta1 x1t plus beta2 x2t plus beta3 x3t plus ut. Where let us say yt is basically your consumption expenditure at the macro level, then x1t is let us say GDP of the country.

Let us say that I have only two variable here. I have income which is given by GDP and also population, these are the two factors that determines the consumption expenditure of the country. And we assume that GDP and population they are highly correlated, they have high pair wise correlation. Now, how will you apply the first difference? If the relationship holds true for the t-th period, obviously the relationship holds true for the t minus oneth period.

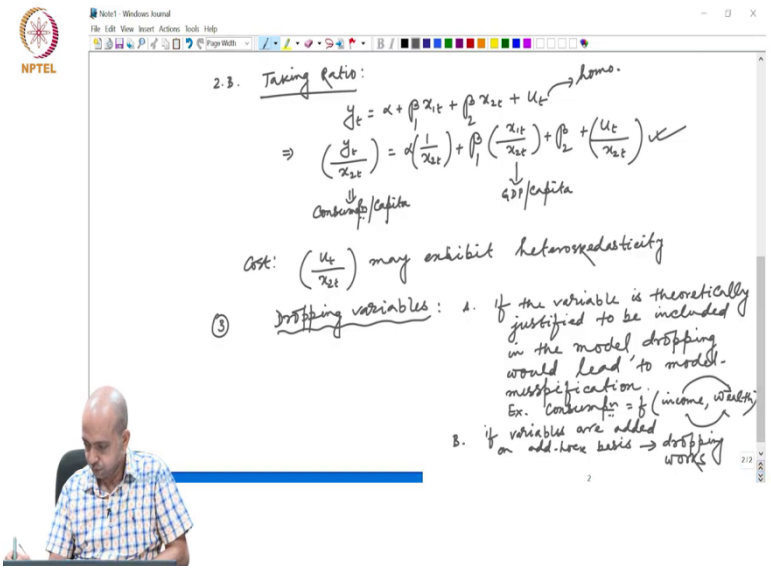
That means the same equation we can write for the t minus oneth period. So, what will happen if I take the equation like this? This should become alpha plus beta 1 x1t minus 1 plus beta 2 x2t minus 1 plus ut minus 1 and then what you do, you take this yt minus yt minus 1 equals to beta 1 x1t minus x1 t minus 1 plus beta 2 x2t minus x2 t minus 1 plus ut minus 1. So, if you do so, then what happens?

If x1t and x2t are correlated, there is no reason to believe that x1t minus x1 t minus 1, so let us say that this is delta yt equals to beta 1 delta xt x1t plus beta 2 delta x2t plus plus, this should become ut minus ut minus 1 plus delta ut. So, that means, what we are saying that just because x1t is correlated with x2t we cannot say that delta x1t would be correlated with delta x2t and even if they do so, the severity of the correlation would be much lower when you take the difference.

So, that means, by doing the transformation if you have a time series data, when you take the first difference, then severity of the multi collinearity problem goes down. But as I said, every measure has some cost also. What is the cost? The cost here if this ut and ut minus 1 actually they may show you some kind of correlation. So, if your original variable does not show any kind of serial correlation, the moment you take the difference, then that may show you some kind of auto correlation problem.

So, auto correlation problem may arise in this first difference when you assume this ut does not show any kind of auto correlation, that is the problem with the first difference. And also if you see our original interest in this model was to see the impact of x1t on yt. But ultimately after transformation what we are getting is actually impact of delta x1t on delta yt. That is the model we have derived. So, this is some, this is the cost involved with the first difference method.

(Refer Slide Time: 10:56)



Then, under transformation this is 2b I would say and this transformation is let us say taking ratio. So, your model is yt equals to alpha plus beta1 x1t plus beta2 x2t plus ut where x1t is GDP and x2t is population. And over a period of time GDP and population show some kind of linear trends, so there is a chance that GDP and population would be highly correlated. So, what you do to overcome that? You divide both sides of the equation by x2t. So, what you will get? You will get yt by x2t equals to alpha into 1 by x2t plus beta 1 x1t by x2t plus beta 2 plus ut divided by x2t. So, that means, after this transformation what I am saying that Consumption per capita is actually a function of GDP per capita.

So, instead of planning the regression in its absolute form, you are basically running the regression on per capita form, which make sense. So, you can also transform some of your variables in ratio form so that I am not losing the any information because my information on population is also there in the form of consumption per capita and here GDP per capita. But what is the cost here?

The transformed error term which is ut by x2t may exhibit heteroskedasticity. We have not yet discussed heteroskedasticity but for the time being just remember this means that error variance is not constant. So, to solve one problem you have end up with the another problem. You were trying to solve multi collinearity problem, but after transformation, you may end up with heteroskedasticity problem.

Hetero, because here ut is also divided by x2t, the population. So, if your original variable ut is homoscedastic, then the transformed variable may show you heteroskedasticity. That is

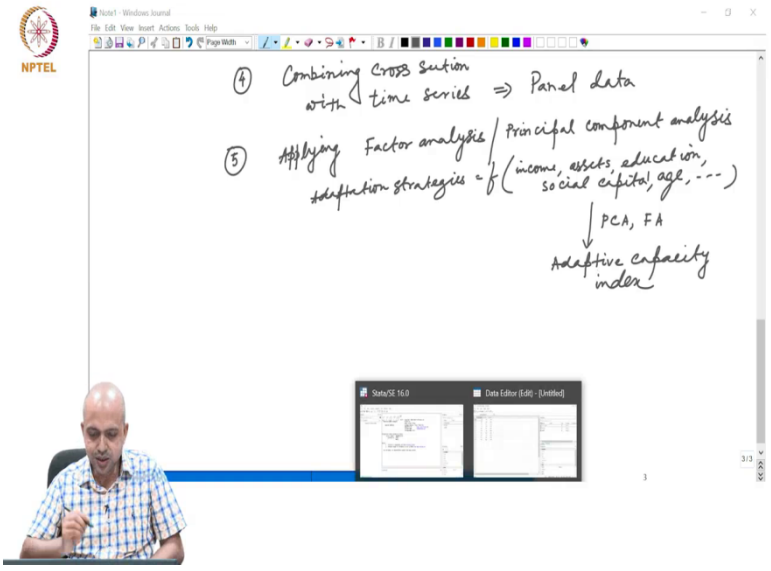
actually the cost of transforming this variable in this format, cost of taking the ratios. So, we have to again carefully think that whether to apply this method or not and then third solution is dropping variable.

So, generally, we think that whenever some variable is correlated with the other one, we will drop one variable and keep the other one. But here, what is the suggestion? The suggestion is if the variable is theoretically justified to be included in the model, dropping would lead to model misspecification. So, you cannot drop a variable when the variable is theoretically justified to be included in the model.

For an example, we can say that consumption is actually a function of both income and wealth. So, even though they are highly correlated with each other, income and wealth both are theoretically justified to be included in the model. So, I cannot drop neither income nor wealth. What is the solution? You might have to increase your sample size. So, it is not possible, so dropping is not possible when it is theoretically justified.

If variables are added on adhoc basis, then dropping is okay. So, you do not have a theoretical justification, you have simply included many variables and now you see some of the variables are inter correlated, so you drop one or those variables and keep the other one. So, it all depends on whether there is any theory behind this or not. So, dropping the variable will work in this way.

(Refer Slide Time: 18:59)



And then the fourth solution is combining cross section with time series. So, if you can do so, then you may combine cross section with time series and get a panel data. So, in panel data

the severity of the multi collinearity problem goes down. That is also another solution, combining cross section with time series data. And fifth is applying factor analysis or principal component analysis.

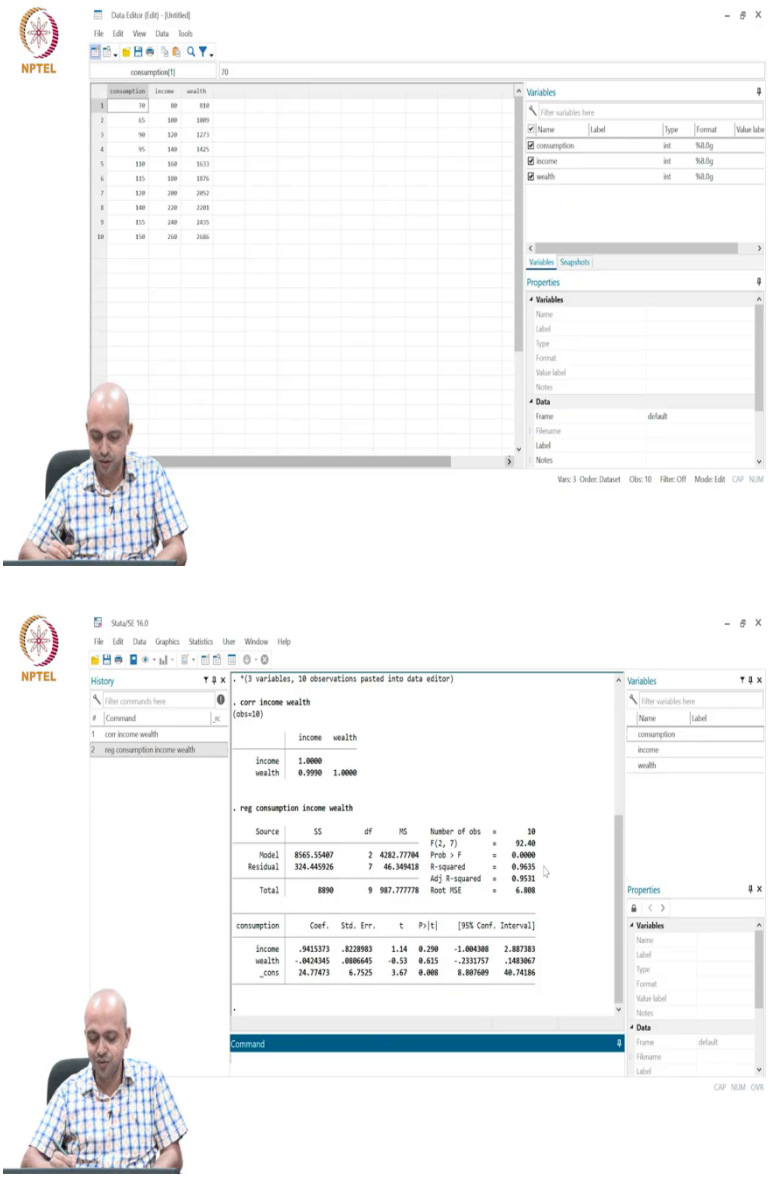
Here the idea is let us say that you are trying to estimate what are the factors that determine the ways people adopt to climate change. So, that means, your dependent variable is adaptation strategies which is a function of let us say your income, then assets and then your education, then your social capital, then your age, male, female, so on and so forth.

There will be many factors that determines the ways people adapt to climate change adaptation strategies. Now, these independent variables, income, asset, education, social capital, so many of these variables they show inter correlation. Then what you do? You apply principal component analysis here to identify some of the principal component out of your n number of such independent variables.

And then you construct some kind of index, if you apply here PCA or factor analysis, you will get some kind of adaptive capacity index. So, instead of using all the variables where you will apply principal component analysis or factor analysis, here we will identify the important factors out of those and then we will construct some kind of index, this is also another method of solving the multi collinearity problem.

So, depending on the situation, depending on your dataset we have to think about which particular method to apply and to solve the multi collinearity. Now, we will take one hypothetical dataset and we will see what type of consequences may arise due to the multi collinearity problem.

(Refer Slide Time: 23:44)



And this is a data, look at the dataset, this is a hypothetical data on consumption, income and wealth. And what I said if you remember that when you import the data before estimating the model applying regression, what you actually do immediately after incorporating the data, you just look at the pair wise correlation. Notice the command, this is corr, then income and then wealth.

Now look at the inter correlation between income and wealth is actually 0.9990. So, that means, almost 1. So, there is high degree of multi collinearity. So, if you estimate the model in presence of multi collinearity what will happen? Consumption, income and wealth. Now, look at what is the cons, what are the consequences of multi collinearity as we discussed yesterday. Your income variable is insignificant.

The p value is 0.290. If you multiply that by 100, that would become 29 so that means I cannot say that the variable is insignificant. What about wealth? Wealth is also insignificant because p value is 0.615. So, that means, you have estimated a model including income and wealth but ultimately both the variables are insignificant.

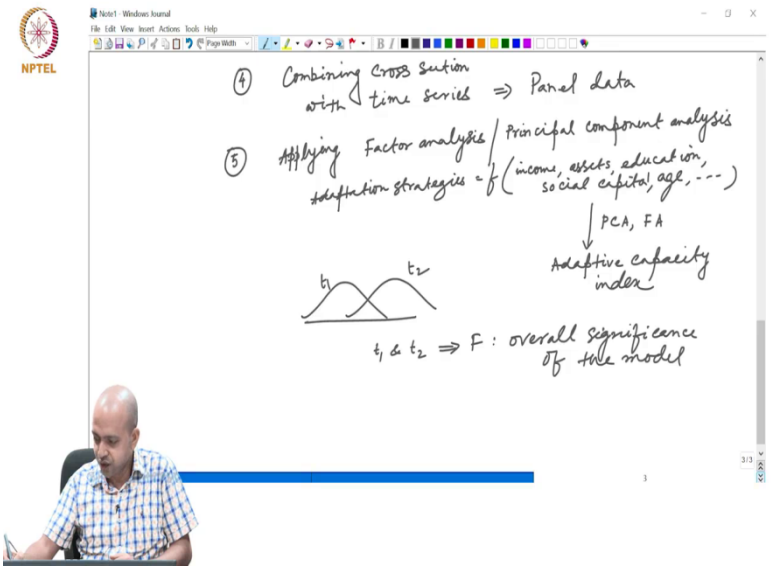
Not only that you see the sign of wealth variable, it is not only insignificant, but the sign what you are getting is negative which is very difficult to justify. How come wealth could have a negative impact on the consumption which is a very difficult thing to justify because our theory says that as wealth increases, that also gives you some positive impact on your consumption, here it is coming totally different.

So, you can now understand the severity of the multi collinearity problem, variables are insignificant and they are coming up with unexpected sign also. And then when you look at the R square, the R square is 0.9635 that means your model explains 96.35 percentage variation in consumption, but none of your variable is significant. How is it possible? Who is explaining them? And there comes the significance of our app test. If you remember that even though variables are insignificant individually, the model becomes overall significant.

So, individual insignificance but overall significance of the model as I told you earlier is a classic symptom of multi collinearity. Here, both the variables are insignificant by their individual t statistic, but the variable is insignificant overall. So, that means, f statistic should only be applied in this context to see whether the model is significant or not.

Individual t statistic will mislead you in the presence of multi collinearity, because these t statistic they are not the actual one. They are the artificial t statistics as a result of the variance inflating factor that means, the standard error got inflated in presence of multi collinearity. And as a result of which this t1 and t2 they become insignificant.

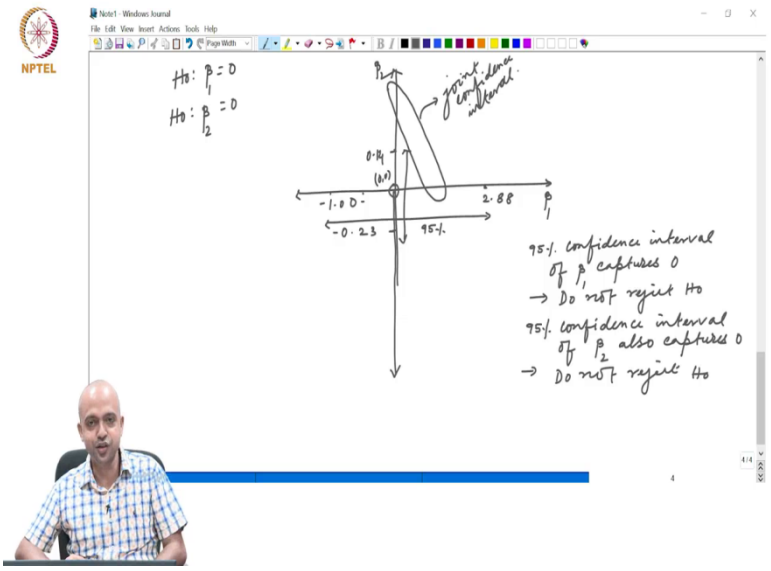
(Refer Slide Time: 28:20)



So, that means, here what is happening once again I will apply the diagram, let us say this is t1 distribution, this is t2 and both are insignificant, but t1 and t2 just because both are insignificant, we cannot say that model is also overall significant. That means, t1 and t2 should not map with F. Why? Because you are drawing the sample, you are drawing the distribution from the same sample.

And that is why this F gains extra significance in this context. So, F gives you overall significance of the model. And the value of f value is 92.40 and probability p that means p value associated with that f is 0.0000 which is highly significant. That means, what is happening here in the context of multi collinearity.

(Refer Slide Time: 29:51)



In the context of multi collinearity if you look at your let us say I will draw you an interesting diagram, so let us say this is the distribution of beta 1 and this is beta 2. Now if you look at the confidence interval of beta 1 and beta 2 look at first the income. The confidence interval 90 percent is minus 1.00 to 2.88. Let us say this is minus 1.00, this is 2.88.

This is the interval 95 percent. Now this 95 percent confidence interval of beta 1, 95 percent confidence interval of beta 1 captures 0. And what was our null hypothesis? Null hypothesis beta 1 equals to 0. Similarly, the null for beta 2 is 0. So, 95 percent confidence interval captures 0, so that means we do not reject the null, do not reject null.

Similarly, if you look at the 95 percent confidence interval for beta 2, what is happening here? Look at the wealth; this is minus 0.23 to 0.14. So, again if you take this 95 percent confidence interval for beta 2, that also captures 0. So, 5 percent confidence interval of beta 2 also captures 0. And what is the solution then? Do not reject H naught. So, as a result of which, both income and wealth, they become insignificant, but what about the joint significance. Joint confidence interval if you apply they have basically an elliptical shape. This is the joint confidence interval that does not capture 0.