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 – 1

So, welcome once again to our discussion of econometrics and yesterday we completed our discussion on dummy variable models. So, that means, so far we have gained some knowledge about how to estimate an econometric model. We have learned the estimation technique for two variable linear regression model, then we have also discussed about multiple linear regression model, and then lastly we discussed about dummy variable model estimation.

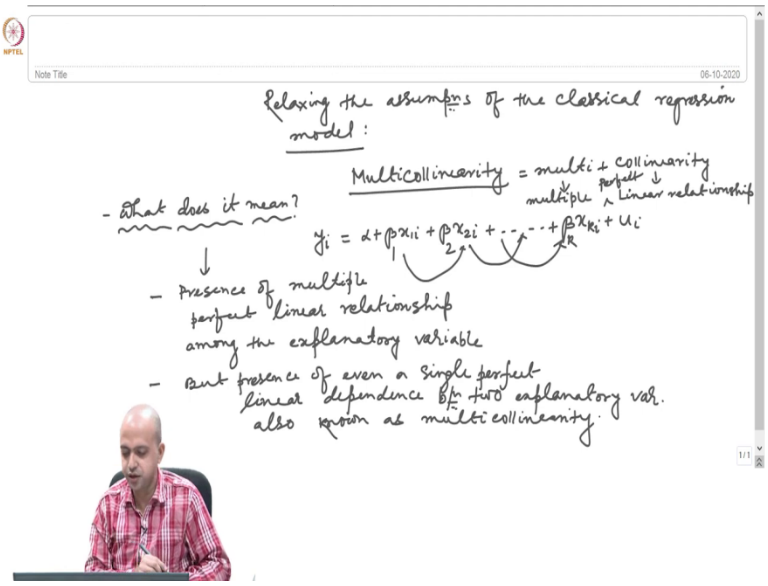
Now once you know how to estimate the model before using the coefficients for your policy purpose or for forecasting, what we need to know or what we need to check actually whether the estimated coefficient exhibit the 3 desirable properties and what are those? They are basically unbiasedness, efficiency and consistency property.

And if you recall we discussed about 10 assumptions. If 10 assumptions are satisfied, then only our estimated coefficients, whatever we are estimating using ordinary least square technique that means using the Euler's technique they will exhibit efficiency, unbiasedness and consistency property. But in reality when you get a dataset for estimation, there is no guarantee that all those 10 assumptions are satisfied by your dataset.

Because this is a data taken from the real life world, taken from the behaviour of particular, certain individuals, certain firms or certain countries whatever and they are behaving based on their own way, we have no control on that, so it is quite unlikely that whenever you get a dataset, all those 10 assumptions are satisfied. But rather it is quite likely that one or more than one such assumptions are violated. And if those assumptions are violated, then your estimates will not give you the desirable properties.

So, that is the reason from today's discussion onward what we need to learn is if those assumptions are relaxed, then what would be the consequences on your estimated coefficients and how to detect whether any of those assumptions are violated, and what are the remedial measures if such assumptions are violated? So, that means, today's discussion would be on relaxing the assumptions of classical regression model.

(Refer Slide Time: 3:23)



Relaxing the assumptions of the classical regression model. To start with we will relax one assumption which we assumed that there should not be any multi collinearity. So, that means, today's topic of discussion is multi collinearity. So, first of all what does it mean when we say multi collinearity?

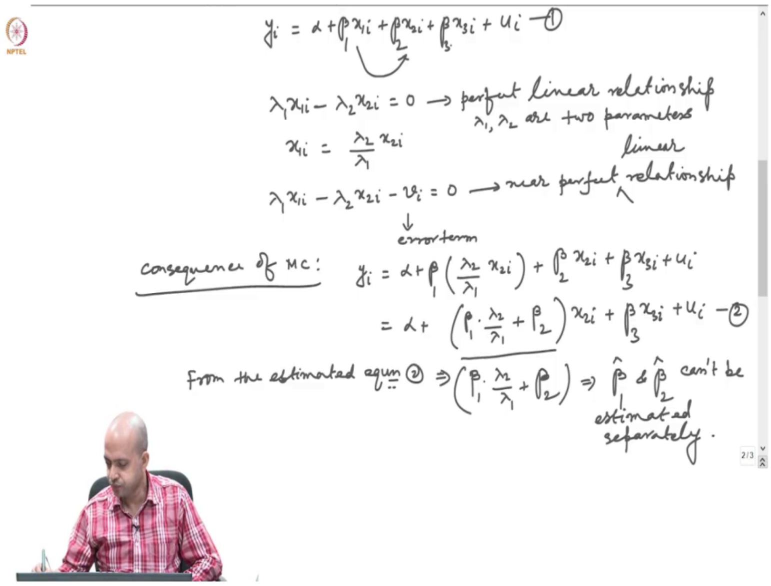
So, let us assume that we are trying to estimate this model yi equals to alpha plus beta1 x1i plus beta2 x2i plus dot dot dot beta k xki plus ui. You have k number of explanatory variables. Then the term multi collinearity actually means, if you decompose this is equal to multi plus collinearity and multi means multiple and collinearity means linear relationship.

So, in your econometric model whatever you have specified it may so happen that this x1 x2 or x2 x3 or x3 x4 they are actually showing liner dependence and you have multiple such relationship between x1 and x2, x3 x5 or x6 x5 x7. So, multiple linear relationship indicates the presence of multi collinearity. That is the actual meaning of multi collinearity, but when you estimate a model, even if you get one such linear relationship and that relationship is called perfect linear relationship, collinearity here I will say perfect.

What is the meaning of perfect? I will explain. So, even if you have one such perfect linear relationship, then also we will say that there is a presence of multi collinearity. So, that means, literally even though multi collinearity means multiple perfect linear relationship, when we estimate a model even one such perfect linear relationship also indicates the presence of multi collinearity. That is all.

Now, what is the mathematical way by which you will understand? So, basically this is the definition: multi collinearity means presence of multiple, perfect linear relationship among the explanatory variables. This is the meaning. But, presence of even a single perfect linear dependence between two explanatory variables also known as multi collinearity. Now I will explain how to write the multi collinearity relationship.

(Refer Slide Time: 9:38)



Let us say that this is our model yi equals to alpha plus beta 1 x1i plus beta 2 x2i plus beta 3 x3i plus ui and we assume that there is perfect linear relationship between x1 and x2. Then the idea here is lambda 1 x1i plus lambda 2 x2i equals to 0. This is called a perfect linear relationship. Because in that case, you can always represent x1 where lambda 1 and lambda 2 they are two parameters. This is the perfect linear dependence.

So, you can easily express or this is you can, if you can put this as minus, it would be easier so that means, I can say that x1i equals to lambda 2 by lambda 1 into x2i. So, if you know the values of lambda 1, lambda 2 and x2i, then you can easily identify what is x1i. This is called a perfect linear relationship. But even though the relationship is not perfect but quite high, then also there would be adverse consequences in your estimates. And how will you represent that? Near perfect relationship.

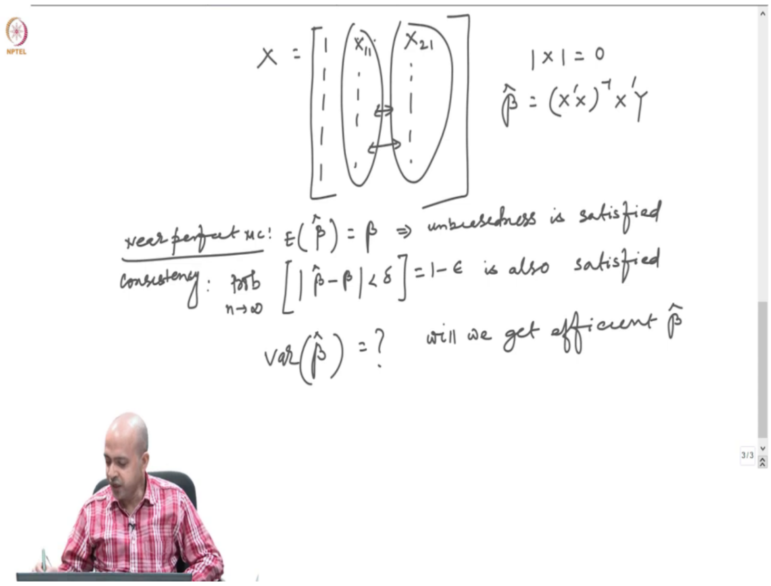
Let us say that this is x1i minus lambda 2 x2i minus some v, v equals to vi equals to 0. Where vi is an error term. So, I can say that there is perfect linear relationship between x1 and x2, because knowing alpha 1, alpha 2, and x2i you cannot perfectly predict the value of x1i because of the presence of vi that is why this is called near perfect relationship. This is the meaning.

Now the question is why should we bother about multi collinearity? That means, if such perfect linear relationship exists between two or more than two explanatory variables, what are the consequences of multi collinearity? So, let us talk about consequence of multi collinearity. In short I will write MC, consequence of multi collinearity. So, when I say that x1i equals to lambda 2 by lambda1 x2i, I can easily substitute the value of x1i in this equation and then what I will get?

Substituting the value of x1i into the original equation, what I can say that, then yi equals to alpha plus beta 1 into x1i. In place of x1i, I will write lambda 2 by lambda1 x2i plus beta 2 x2i plus beta 3 x3i plus ui. So this again I can write alpha plus beta 1 into lambda 2 by lambda 1 plus beta 2, beta 1 I can take this as common plus beta 3 x3i plus ui. Now if you estimate this model, the coefficient of x2i would be this. So, from this model, let us say this is equation 1 and after substituting this is equation 2.

So, from 2, from the estimated equation 2 what you will get? beta 1 into lambda 2 by lambda 1 plus beta 2. so that means I will get only the combined value of beta 1 and beta 2 while my objective was to get beta 1 and beta 2 separately. So, that means, when there is a presence of perfect linear dependence among two variables so that means I cannot get beta 1 hat and beta 2 hat separately. That is the problem. So, I will get a combined value of beta 1 hat and beta 2 hat.

(Refer Slide Time: 17:11)



Moreover, if you have perfect linear dependence, then what will happen? If you recall yesterday's class while discussing about the dummy variable trap, what we say that your X matrix looks like in this you have 1 1 1 1 1 and then if x1 and x2 they are perfectly linearly dependent, then what will happen? Your x1i, let us say this is x1i and this column is let us x1 1 and this is x21. So, this column and this column they are now perfectly dependent.

So, as a result which once again if that is the case, then what will happen? That the determinance actually become 0. Determinance of this matrix is 0 when there is perfect linear dependence. So, that means, beta hat what you are estimating as x prime x inverse x prime y that you cannot estimate when there is perfect linear dependence between x1 and x2 variable. That is the problem because your two columns of this X matrix will be perfectly linearly dependent.

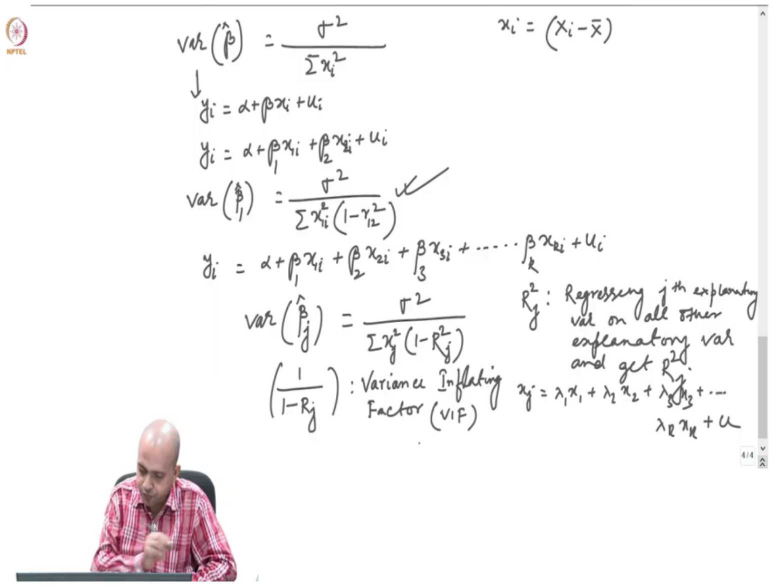
So, that means, from this approach and from this matrix approach, you can easily understand that when there is perfect linear dependence, then your model cannot be estimated. But what will happen, what will happen if you do not have perfect linear dependence? Let us say this is near perfect linear relationship. What will happen? What would be the consequences of that?

Now, let us say we have near perfect MC, what would be the consequence? See, the consequence would be in your estimates. If you take expectation of beta hat, is still beta. That means, unbiasedness property is still satisfied. So, this implies unbiasedness is satisfied. And then beta hat and beta, the difference between these two is and probability of getting such beta when n tends to infinity equals to 1 minus lambda is also satisfied.

So, that means, this is called unbiasedness and what is this called? This is called consistency property. So consistency property is also satisfied. So, even if there is multi collinearity or near perfect multi collinearity, you will still have unbiased estimate, you will still have beta hat which will show the consistency property. But the problem that will happen is the variance of this. What will happen to the variance of beta hat? That means, basically we are talking about the property of efficiency.

Will we get efficient parameter, efficient beta hat? Now what does efficiency mean? Efficiency means basically if you estimate a model using ordinary least square method, then if all other assumptions are satisfied, then Euler's estimate will give you minimum variance in the class of all other unbiased estimates of beta estimated by different estimation technique. But in a given sample when you have endured with particular sample an estimate the model what will happen to the various of beta hat form that particular sample that we need to understand.

(Refer Slide Time: 23:05)



And to know about the variance of beta hat in presence of perfect or near perfect multi collinearity, first of all what is variance of beta hat? Variance of beta hat is defined as sigma square divided by summation xi square where xi equals to capital Xi minus X bar. So, when your model involves only one explanatory variable, that means this is applicable when your model is yi equals to alpha plus beta xi plus ui. So, this is the variance formula when you have only one explanatory variable in the right hand side.

Now, when your model involves more than one explanatory variable alpha plus beta 1 x1i plus beta 2 x2i plus beta 3 let us say x3i plus ui or let us say I have only two for the time being plus ui, then this particular variance formula changes as now variance of beta 1 in this case variance of beta 1 would be sigma square divided by summation x, x1i into square into 1 minus r square 1 minus r square 1 2. You have now two explanatory variable. So, that means the changes will happen like you will have x1i square (Pl) into 1 minus r square 1 2.

So, similarly now this 1 minus when you have multiple linear regression model, let us say you have now your model now changes as alpha plus beta 1 x1i plus beta 2 x2i plus beta 3 x3i plus beta k xki plus ui. And now suppose you want to get the variance of beta let us say j, beta j. So, what you will do? You have multiple linear relationship and now when you have multiple variables how this formula, this particular formula will change? When you have two, you can easily calculate the correlation between 1 and 2 and you can square it off.

But when you have k number of explanatory variable and you want to get the variance of a particular variable, let us say this is beta j. So, how will you modify this formula? This modify, modification of this formula will happen like summation xj square into 1 minus capital R square j. And what is R square j? R square j is basically how will you get? By regressing jth explanatory variable on all other explanatory variable and get R square and that is known as R square j. Is this clear?

So, when you have multiple regression model, then this R square j is nothing but the R square from the R square what you get by regressing the jth explanatory variable on all other explanatory variable. And why I am regressing this? Because this regression basically shows if other explanatory variable can explain a significant portion in the variation in xj that basically shows that all other variables are highly correlated with the jth explanatory variable, quite simple.

So, that means, they are basically asking you to run this type of regression. So, this is your xj equals to some lambda 1 let us say x1 plus lambda 2 x2 plus lambda 3 x3 plus dot dot dot lambda k xk plus some u. So, that means I am regressing the jth explanatory variable on all other explanatory variable and collecting the R square and that R square is basically this R square. That is how you can change the variance formula.

Now, this 1 minus 1 min, 1 by R squared j, 1 by minus R square j, this is given a specific name in the literature which is called variance inflating factor. What I am saying? Variance inflating factor or in short, VIF.