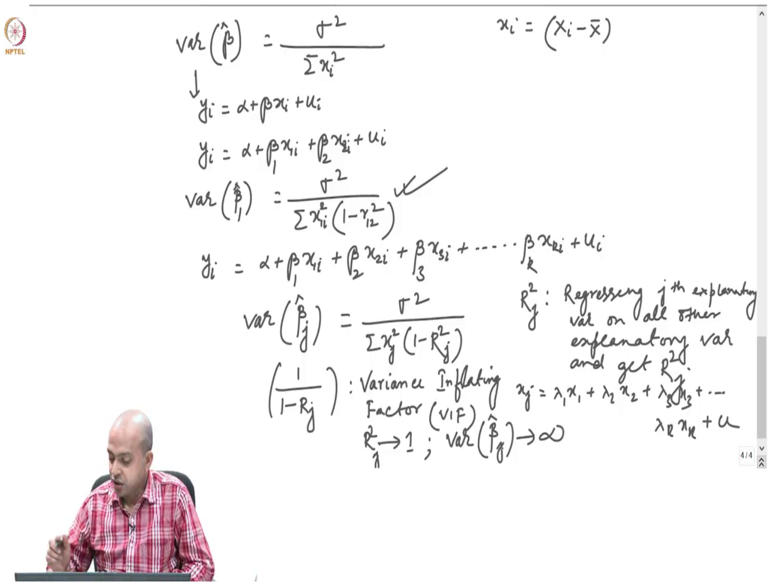
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Relaxing the assumptions of CLRM-Multicollinearity and Autocorrelation Part – 2

(Refer Slide Time: 0:15)



$$var(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2} \qquad x_i = (X_i - \bar{X})$$

$$\downarrow$$

$$y_i = \alpha + \beta x_i + u_i$$

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_{2i} + u_i$$

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum x_i^2 (1 - r_{12}^2)} \checkmark$$

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots \beta_k x_{ki} + u_i$$

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2 (1 - R_j^2)}$$

R_j^2 : Regressing j^{th} explanatory var on all other explanatory var and get R_j^2.

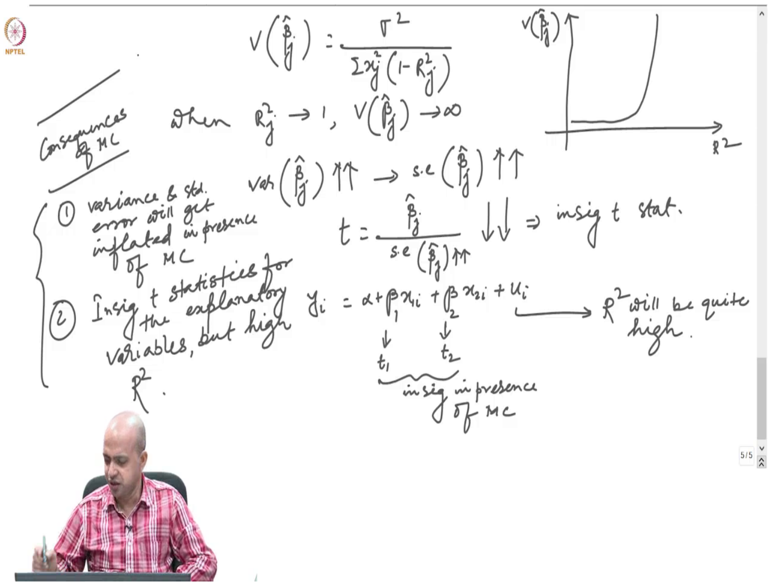
$\left(\frac{1}{1 - R_j}\right)$: Variance Inflating Factor (VIF)

$x_j = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \cdots \lambda_k x_k + u$

$R_j^2 \to 1$; $var(\hat{\beta}_j) \to \infty$

Now why this is called variance inflating factor because from here you can understand if the R square tends to be, if the R square j tends to 1, then what will happen, v, variance of beta j will tend to infinity, is not it? Variance of beta j will tend to infinity. That is why this 1 by 1 minus r square j is called the variance inflating factor. There is high degree of linear dependence among several explanatory variable that means you can expect R square j would be quite high and if R squared j tends to infinity, then, tends to 1, you will get variance of beta j tends to be infinite. So, that means, this is a serious consequence.

(Refer Slide Time: 1:27)



$$V(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2 (1-R_j^2)}$$

Consequences of MC

when $R_j^2 \to 1$, $V(\hat{\beta}_j) \to \infty$

① variance & std. error will get inflated in presence of MC

$var(\hat{\beta}_j) \uparrow\uparrow \to s.e(\hat{\beta}_j) \uparrow\uparrow$

$t = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)\uparrow\uparrow} \downarrow\downarrow \Rightarrow$ insig t stat.

② Insig t statistics for the explanatory variables, but high R^2.

$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \longrightarrow R^2$ will be quite high.

$\underbrace{t_1 \quad t_2}_{\text{insig in presence of MC}}$

So, what will happen, that means variance of beta j beta hat j as I told you this is sigma square divided by summation xj square into 1 minus R square j, and when R square j tends to 1, variance of beta j tends to infinite and if you plot R square in this axis and variance of beta hat in the y axis, then you will see the relationship will look like this.

So, as R square increases, u will get this type of relationship, it will increase like anything, this type of relationship. Now when variance increases, variance of beta j increases, that will lead to the standard error of beta j also will increase. And if standard error increases of beta hat, what will happen to the t statistic which is nothing but beta hat j divided by standard error of beta j?

What will happen if this increases? Then t statistic will fall drastically. So, t statistic will become insignificant. So, that means, in presence of multi collinearity what you observe as the t value that is actually not the actual t value because the variance got inflated in presence of MC, standard error got inflated in presence of MC.

That means, this t statistic appears to be artificially low. And if the t statistic is low, then what will happen? You will see that your individual t statistic is insignificant, so that means variables will appear to be insignificant. But, even though this t statistic, individual t statistics are insignificant, you will still get a high r square because that is coming from here. So, that means, what would be the consequence of multi collinearity?

Suppose this is your model yi equals to alpha plus beta1 x1i plus beta2 x2i plus ui and you see by individual t statistics, let us say this is t1, and this is t2 both are insignificant in presence of MC. But the R square what you get from this model, the R square will be quite high. So, this is the classic symptom as well as consequence of multi collinearity. So, that means, many of your explanatory variables will appear to be insignificant but still you see the high R square value from the model.

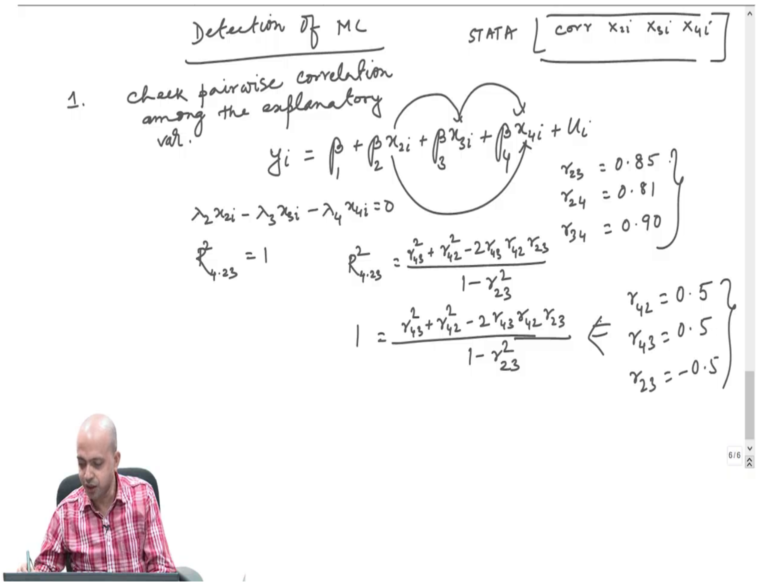
So, whenever you get this type of result, you estimate a model and immediately after estimation you see that your R square is quite high, but many of your explanatory variables are insignificant individually then immediately it should click in your mind that means my data is suffering from multi collinearity problem. This is the classic symptom as well as consequences of multi collinearity problem.

This will happen. So, that means, your variance will get inflated, I will write shortly consequence variance and standard error will get inflated in presence of MC, multi collinearity and as a result of which you will get insignificant t statistics for the explanatory variable, but high R square.

So, these are the two consequences of multi collinearity. So, that means, out of the 3 desirable properties- consistency and unbiasedness properties are still maintained, but efficiency property get disturbed and when I say efficiency, it means for a given sample the variance of particular beta hat let us say beta j if you assume beta j is correlated with it variable may get disturbed and that will result in this type of situation.

This is the classic consequence as well as symptom of multi collinearity problem. These are the consequences, then the next question is how will you detect multi collinearity, detection of multi collinearity?

(Refer Slide Time: 8:54)



Detection of MC

STATA [corr x_{2i} x_{3i} x_{4i}]

1. check pairwise correlation among the explanatory var.

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$$

$r_{23} = 0.85$

$r_{24} = 0.81$

$r_{34} = 0.90$

$$\lambda_2 x_{2i} - \lambda_3 x_{3i} - \lambda_4 x_{4i} = 0$$

$$R^2_{4.23} = 1$$

$$R^2_{4.23} = \frac{r^2_{43} + r^2_{42} - 2 r_{43} r_{42} r_{23}}{1 - r^2_{23}}$$

$$1 = \frac{r^2_{43} + r^2_{42} - 2 r_{43} r_{42} r_{23}}{1 - r^2_{23}} \Leftarrow$$

$r_{42} = 0.5$

$r_{43} = 0.5$

$r_{23} = -0.5$

Detection of multi collinearity - First, this is a simple measure, you just do what you do, check pair wise correlation among the explanatory variable. For example let us say my model is yi equals to let us say beta 1 plus beta 2 x2i plus beta 3 x3i plus beta 4 x4i plus ui, then the simple measures that they say when you write a paper also before presenting your regression result, what you show actually in your paper what is the pair wise correlation between x2 x3, x3 x4 and x1 x2 x4.

So, this is called pair wise correlation and the statistical software what you use, if you simply after importing the data if you simply put the STATA comma which is corr, then x2, then x3 and x4, if you put this command, this is the STATA command, then that will give you this pair wise correlation.

And any pair wise correlation which is 0.8 or more than 0.8 that means when r, pair wise correlation let us say between 2 3 equals to 0.85 or let us say r 2 4 is 0.81 or let us say r 3 4 equals to 0.90. So, these are basically high pair wise correlation. So, if you get this type of result, then you will understand that there is some kind of multi collinearity problem in our dataset and we must rectify that.

But this is a simple measure, as I said that pair wise correlation is not a very rigorous statistical measure of detecting multi collinearity because this is only simple pair wise correlation measure and it has some limitation also. For example, let us say I am assuming that this x4 is highly correlated with x2 and x3 that means let us lambda 1, lambda 1 x, there is perfect linear dependence, lambda 2, lambda 2 x2i minus lambda 3 x3i minus lambda 4 x4i equals to 0. There is perfect linear dependent.

So, we can expect that R square 4 equals to 1. So, that means, when I am regressing the fourth explanatory variable on the second and third explanatory variable, the R square from this model is coming out to be 1. But if you write this, this formula, what is R square 2 3 actually?

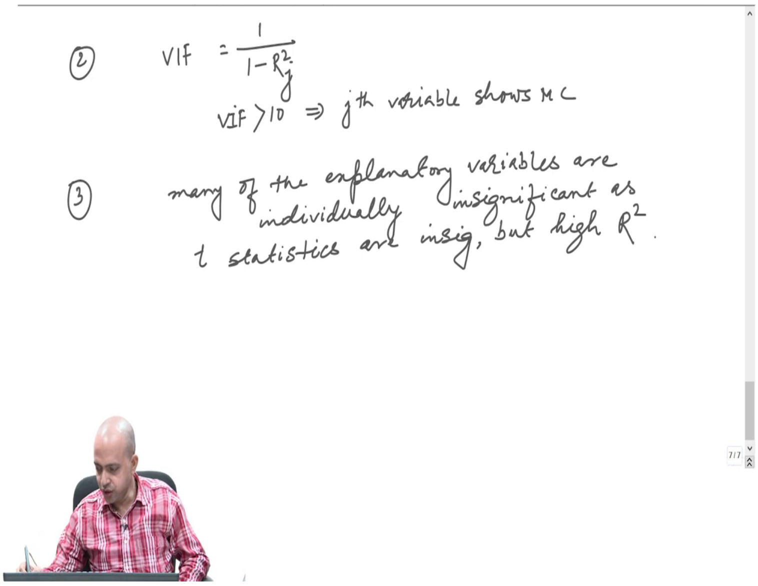
R square 4 2 3 is basically r square 4 3 plus r square 4 2 minus 2 into 2 into r square, sorry this is not r square, this is r 4 3, then r 4 2, then r 2 3, 1 minus r square 2 3. So, that means, if you substitute 1 here, so for 1 this is r 4 3 plus r square 4 2 minus 2 into r 4 3 r 4 2 r 2 3 this is 1 minus r square 2 3. Where r square indicates the squared correlation or you can say that by regressing the fourth variable on this and this is partial correlation coefficient.

Now, if this relation to be satisfied so that means if this relations to be satisfied, then what you can understand is that from there you just look at this relationship carefully that this left hand side is 1 and you need to have specific value for this. Now this relation if you look at even for r square r42 if you assume 0.5 and r43 is also 0.5 and r 2 3 is minus 0.5 you will get these condition to be satisfied.

That means pair wise correlation if you see it is not very high, but still you have high degree of multi collinearity. So, that is why I said that pair wise correlation may not always guarantee that means low pair wise correlation does not always rule out the possibility of multi collinearity problem. That is why this is not a very rigorous statistical measure because from these relationship what I see that even though r square 4 2 3 is quite high, that means, when r square 4 2 3 is 1, from you have serious multi collinearity problem.

Fourth explanatory variable is perfectly linearly dependent on 2 and 3, second and third variable perfectly explain the fourth explanatory variable. But from the partial pair wise correlation we see they are only 0.5. So that is why it may not, if the pair wise correlation is quite high, then you will understand there is multi collinearity problem, but if they are low, they do not rule out the possibility of multi collinearity problem, we have to check for some other measure.

(Refer Slide Time: 17:37)



② VIF $= \frac{1}{1-R_j^2}$

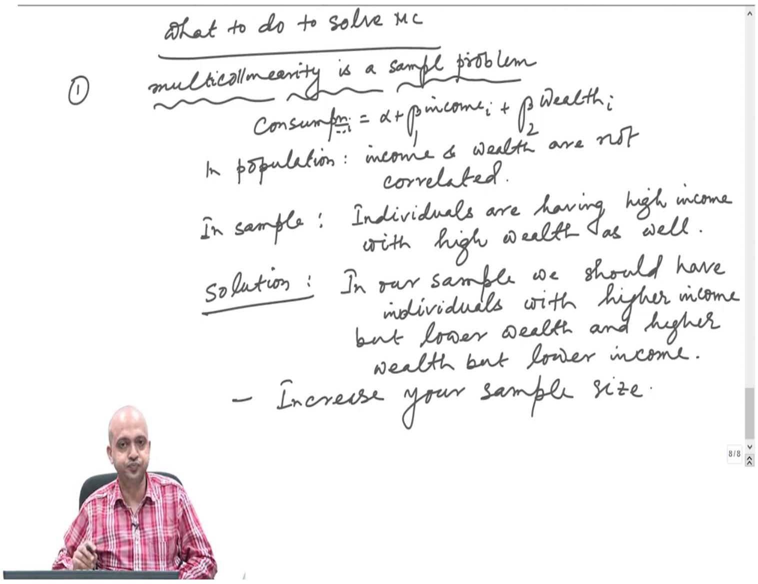
VIF $> 10 \Rightarrow j^{th}$ variable shows MC

③ many of the explanatory variables are individually insignificant as t statistics are insig, but high R^2.

And what is the other measure that is simple to apply also, that is VIF that means 1 minus 1 by 1 minus R square j. This is the other way of detecting this, so if the VIF so defined is greater than 10 for this particular jth variable that will imply that the jth variable shows MC. That means, jth variable is perfectly linearly dependent with others. So, this is a simple formula to apply and it is very easy to estimate from the statistical software also.

And third one is as I said how will you detect that many of the explanatory variables are individually insignificant as these statistics are insignificant, but high R square. So, this will also help you detecting multi collinearity in your dataset. So, that means, we have discussed the consequences of multi collinearity and this is how you can detect multi collinearity and next thing what we need to do is the solution for multi collinearity.

(Refer Slide Time: 20:25)



What to do to solve MC

① multicollinearity is a sample problem

$\text{Consumptn}_i = \alpha + \beta_1 \text{income}_i + \beta_2 \text{Wealth}_i$

In population: income & wealth are not correlated.

In sample: Individuals are having high income with high wealth as well.

Solution: In our sample we should have individuals with higher income but lower wealth and higher wealth but lower income.

– Increase your sample size.

What to do to solve multi collinearity. Now in this context econometrician generally say that multi collinearity is a sample problem. What I am saying, this is a very very important statement I am making, they say that multi collinearity is a sample problem. So that means we need to modify our sample appropriately to solve the multi collinearity problem.

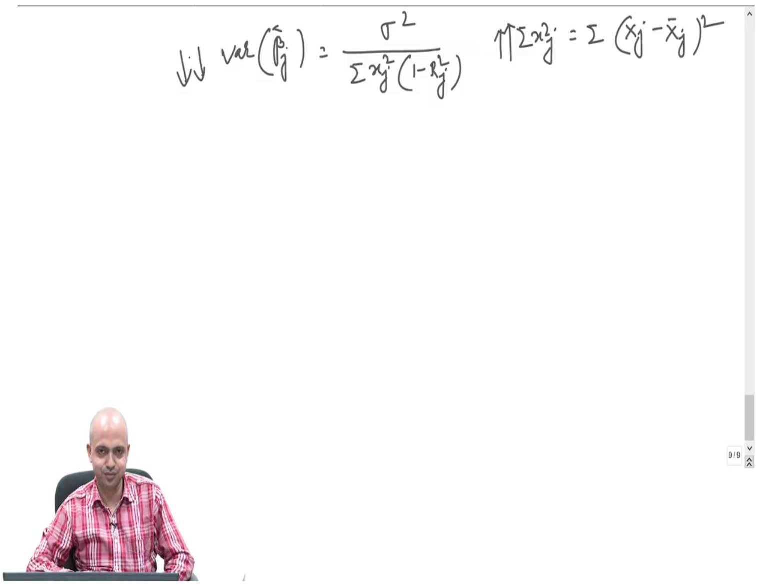
I will explain this. Let us say that your model is consumption equals to alpha plus beta 1 income, this is i income plus beta 2 wealth. So, when I include both income and wealth in my model to estimate the consumption function that means we assume that in population income and wealth are actually are not correlated. But in sample what is happening here it may so happen in sample the sample what you have collected, then you got information from those individuals who are having.

So, in sample individuals are having high income with high wealth as well. So, that means, even though in population, we assume income and wealth to be are not correlated because that is how consumption function theory says that consumption function, the consumption depends not only individual's income but also wealth, we assume in the population they are not correlated. But when we collect a sample, it may so happen that individuals are having high income with high amount of wealth as well.

So, how will you solve this problem? To solve this problem you need to collect a sample wherein, there are individuals. So, what is the solution? In our sample we should have individuals with higher income but lower wealth and higher wealth but lower income. So, we need to design our sample in such a way that this particular is satisfied. We have individuals which are having high income but low wealth and we have other individuals who are having low income but high wealth.

So, we need to solve this problem by designing the sample strategy appropriately that is why they say that multi collinearity is basically a sample problem. But it is not always possible to get that type of sample. And how will you get this sample? Just increase your sample size. If you increase your sample size, then you have higher probability of getting that type of individuals who will have higher wealth, low income and higher income low wealth.

(Refer Slide Time: 26:30)



$$\downarrow\downarrow \; var(\hat{\beta}_j) = \frac{\sigma^2}{\Sigma x_j^2 (1-R_j^2)} \qquad \uparrow\uparrow \Sigma x_j^2 = \Sigma (X_j - \bar{X}_j)^2$$

And when you increase the sample size, one more thing happens. Let us say beta hat j what I say this is equals to sigma square divided by summation xj square and 1 minus R square j and when you increase, what is x square j? This is x square j is basically this is summation xj minus xj, this is capital. So, this is Xj minus Xj bar whole square.

So, if you increase the sample size, obviously these will also increase. That is why when this increases, then variance of j will come down that is why increasing the sample size will solve two problem, we will have individuals with higher income, lower wealth and lower wealth high income at the same time you are reducing the severity of multi collinearity problem as well. Now there are other ways by which we can solve the multi collinearity problem and that those other measures we will discuss in our next class.

Thank you very much.