Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

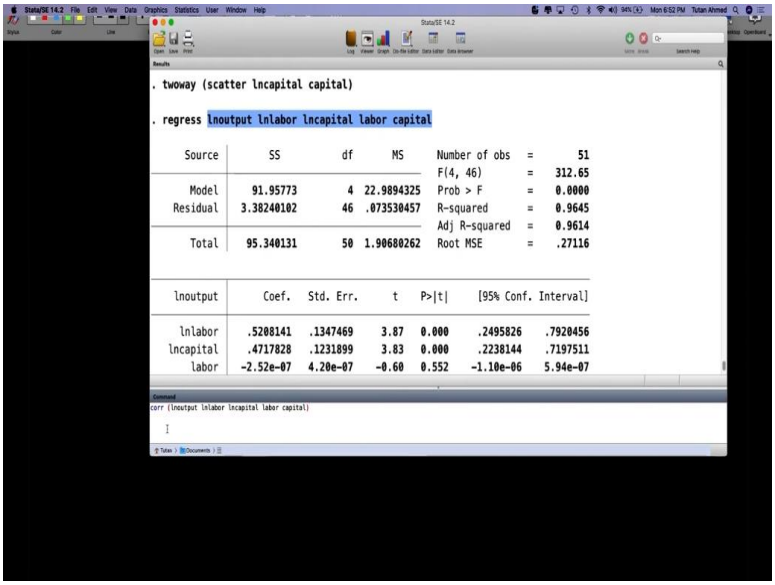
Module - 7
Lecture - 53
Multicollinearity (Contd.)

Hello and welcome back to the lecture on Applied Econometrics, and we are talking about the topic multicollinearity. Now, we have talked about many aspects of multicollinearity. We have talked about what multicollinearity is, how to see if there is a multicollinearity and what are the problems if we have multicollinearity. Now, we will see few more techniques to understand if the multicollinearity is present.

Now, as such, you do not have a test very specific to understand multicollinearity, like we will see in case of, say other problems like heteroscedasticity or autocorrelation, we have specific test to understand if the problems are there. But in multicollinearity, it is a problem of degree, and you try to understand with different ways if there are problems of multicollinearity.

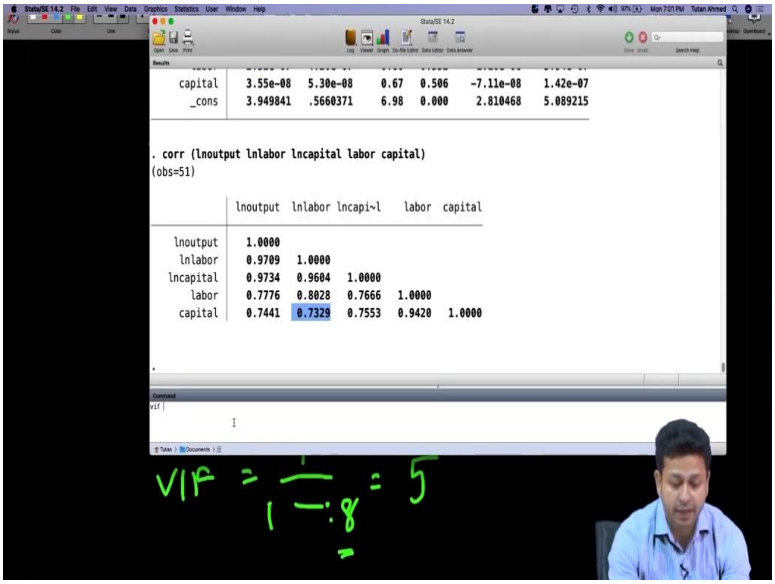
So, one other way to understand multicollinearity is to see; you take simply 2 explanatory variables and you see the pairwise correlation coefficient; as simple as that. And if you see the correlation coefficient is very high, you usually say that you have to look at these 2 variables, because there is a high correlation coefficient. As such, we take any correlation coefficient which is more than 0.5, we consider there might be a problem of multicollinearity, because they are highly correlated and because the correlation coefficient is more than 0.5. So, let me actually do this; let me actually script to show that if there are problems of multicollinearity.

(Refer Slide Time: 01:38)



So, in the way we have taken the variables, there has to be problem multicollinearity, because we have taken the same explanatory variable with different functional form. So, let us see, we can actually, we write the code corr for pairwise correlation coefficient. We can take all the variables. So, you can take only the X variable, you can take all the variables. So, it is okay to see.

(Refer Slide Time: 02:10)

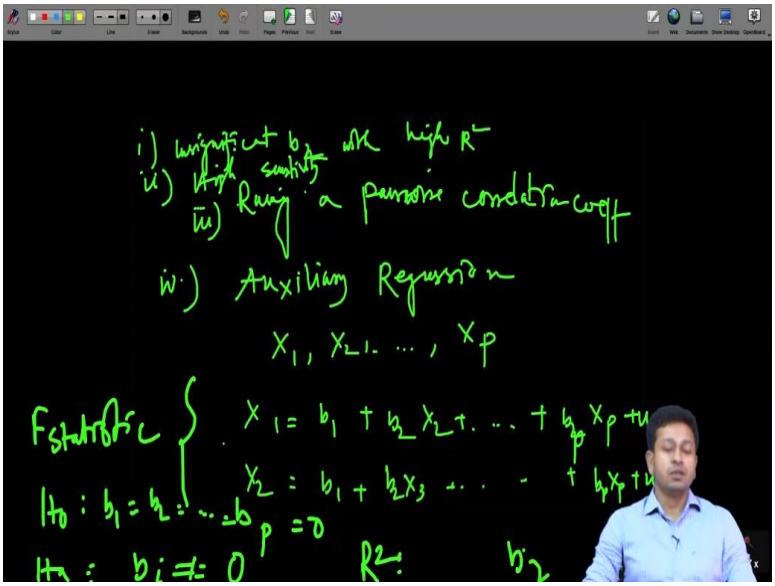


So, it will give the relationship among all the different variables. So, this one is giving the relationship between Y variable and the X variables. And here, we get one-to-one relationship among the X variables. So, here, of course, ln labor, ln labor; so, it will definitely give you 1, correlation coefficient is 1. Here, ln capital, ln labor; it has a high correlation coefficient, 0.96. Labor and ln labor, of course, it will be high, like 0.8.

And capital, ln labor, again is very high, 0.73. So, essentially, as per our criteria, what we said that 0.5 is the cutoff. All the cases you will see that there is a high multicollinearity problem. So, you have to see which one you need to keep, which one you need to drop. So, perhaps, from your knowledge, you know that there might be some relationship between labor and capital, but it is more likely that these 2 functional form, ln labor and labor, and ln capital and capital, they are creating the problem.

So, you might actually want to drop one of those variables. So, maybe the capital and labor you want to drop, and you might want to keep the ln labor and ln capital. So, that is one way of looking at if there are problems of multicollinearity. You simply run a pairwise correlation coefficient. The other way; so, let me write it down.

(Refer Slide Time: 03:29)



So, one way is running a pairwise correlation coefficient. And in the previous video, we have seen, the symptom was, say low. Actually, there are two symptoms we discussed. So, let me write down here; one was insignificant beta or b 2; insignificant b 2 with high R square. In the previous class, we explained this, high, high R square. And the second problem was high sensitivity.

And here, I am saying that we simply take a pairwise correlation coefficient, and we see what happens there. The other way of doing it is, running auxiliary regression. So, what is auxiliary regression? Auxiliary regression is nothing but, you run regression equation for each of these X variables. So, let us say I have my X variable; I have a regression equation

where I have X variable, X 1, X 2 ... X p, and I want to essentially see how each of these explanatory variables are getting determined by other explanatory variables.

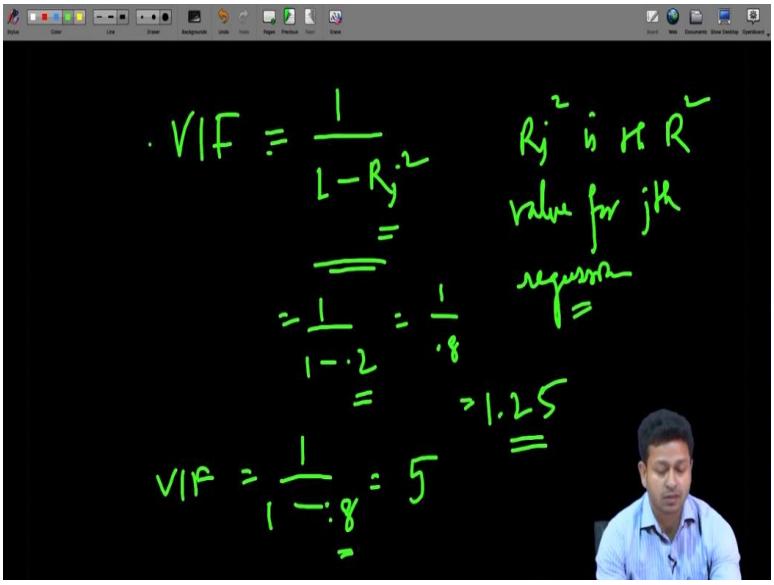
So, let us say beta 1 plus beta 2 X 2 and up to say beta p X p plus some error term. Similarly, you can write X 2 is equal to beta 1 plus beta 2 X 3 and so forth, beta p X p plus error term. And you basically run its regression equation for each of these explanatory variables. So, essentially, if I have say p explanatory variables, so, I will actually run p auxiliary regression equation, because I am trying to explain each of these variable with respect to the rest of the other explanatory variables.

So, that is a little cumbersome process and though it is a good way to understand if there are relationship among these variables; because what we do here, we simply try to see that; first thing we check for each of these regression equation, what we do is, we check if the F statistic is significant, because we know the F statistics is talking about the joint significance of the model, if whether the null hypothesis is beta 1 equal to beta 2 ... to beta p is equal to 0, or my alternative hypothesis is, either of this beta coefficient is not 0.

So, that is when we use the F statistic to understand the joint significance of the model. Now, if the model is actually explaining, so, then we have to be careful that there might be some problem with multicollinearity. And the second thing we need to see is the value of R square. If the R square value is high, so, then, that means, some of the X variables are actually explaining the dependent X variable.

Now, corresponding p value for all these different beta coefficients to understand if the regression equations are actually really making sense, and if actually one of those or different X variables are actually related with the rest of the X variables. So, this is how; let me write down beta coefficient; so, these are the things you actually need to check when you check the auxiliary regression equation; but, the problem as I said is that, it might be a little (cumber).

(Refer Slide Time: 08:08)



Large number of explanatory variables. So, you cannot run so many different auxiliary regressions. For each regression, you have to run it separately. So, that is why we do something called; what you have learnt already; is called VIF. We actually have a command or called VIF that we use to understand the variance inflating factor. So, for auxiliary regression, how it calculates is that, 1 by 1 minus R j square, so, where R j square is the R square value for jth regression.

So, the concept remains same. When we used small r square, the correlation coefficient, we simply took the relationship between 2 explanatory variable, because we had 2 explanatory variable; but when we have multiple explanatory variable, we do not use the row or the small r X 2, X 3, but instead we use the capital R. So, that is what will give me the value of the VIF.

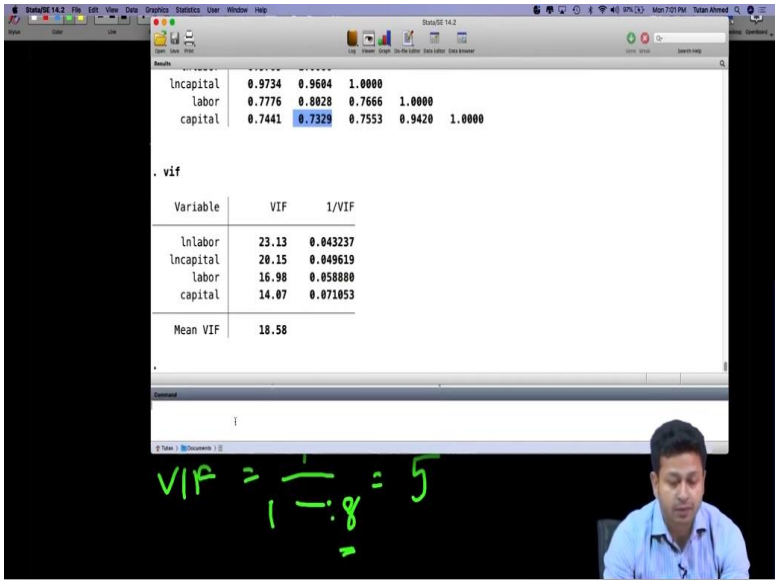
So, in case of the auxiliary regression equation and when we actually calculate VIF, we actually end up calculating 1 by 1 minus R j square. Now, if the R j square value is pretty low, so, let us say 0.2, so, then, my VIF is going to be 1 by 1 - 0.2, 1 by 0.8, which is essentially, I think 1.25. So, which is actually okay, I mean, if the R square value is low, the VIF is also going to be low.

But if the R square value is high, so that, if that means the model, the X variables are actually explained by other explanatory variables, so, then, the VIF value is going to be very high. Let us say, if my R j square is going to be 0.8 instead, it is going to be 1 by 1 - 0.8. So, it is going

to be 5. So, in general, we take a VIF value which is 5 or above. So, that is, the R square value is 0.8 or above.

So, that we consider as a problem, because then, that means the R square value is very high, and we really do not want that. So, when I have the VIF value is equal to 5, we then actually take that into consideration, saying that there might be the multicollinearity problem.

(Refer Slide Time: 10:49)



And in our previous regression, if we simply type VIF, it will provide the VIF value here. And we will see in all the cases, we have a very high VIF value. And it is quite obvious, because that is the reason we have chosen this equation, where we actually show high multicollinearity problem; but we will see in other equations, other examples that the VIF value is varying.

And we have to kind of use the VIF values to understand whether we are going to further dig deep into the multicollinearity problem looking at the VIF values. So, with this, we end this lecture. And in the next lecture, we are going to talk about R square value and correlation coefficient. There are different types of correlation; partial correlation, semi-partial correlation. So, we are going to talk about all these in the next lecture. Thank you.