

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

Module - 7
Lecture - 54
Multicollinearity (Contd.)

Hello and welcome back to the lecture on Applied Econometrics. So, we are talking about multicollinearity, and we have explained pretty much different aspects of multicollinearity. Now, in this lecture, we are actually going to talk about R square and the correlation coefficient in the context of multicollinearity. Now, we have seen previously that our R square value is nothing but my r square value in case of simple OLS.

(Refer Slide Time: 00:49)

$R^2 = r^2$ simple OLS

Y

x_2, x_3, x_4

$R^2 = \sum r^2$

Ideal condition $r_{x_i x_j} = 0$

We have actually proved that this is what we will get.

(Refer Slide Time: 00:57)

StataSE 14.2

	capital	lnoutput	lnlabor	lncapital	labor	capital
capital	3.55e-08	5.30e-08	0.67	0.506	-7.11e-08	1.42e-07
_cons	3.949841	.5660371	6.98	0.000	2.810468	5.089215


```
. corr (lnoutput lnlabor lncapital labor capital)
(obs=51)
```

	lnoutput	lnlabor	lncapital	labor	capital
lnoutput	1.0000				
lnlabor	0.9709	1.0000			
lncapital	0.9734	0.9604	1.0000		
labor	0.7776	0.8028	0.7666	1.0000	
capital	0.7441	0.7329	0.7553	0.9420	1.0000


```
. vif
```



```
Command:
regress lnoutput lnlabor
```

And let me actually do a little bit of demonstration here. So, let us say this is my regression equation, I will just take just a simple one explanatory variable, let us say it is ln labor, and I run a regression.

(Refer Slide Time: 01:14)

StataSE 14.2

```
. regress lnoutput lnlabor
```

Source	SS	df	MS	Number of obs =	51
Model	89.8648125	1	89.8648125	F(1, 49) =	804.22
Residual	5.47531855	49	.111741195	Prob > F =	0.0000
Total	95.340131	50	1.90680262	R-squared =	0.9426
				Adj R-squared =	0.9414
				Root MSE =	.33428

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnlabor	.9795049	.0345397	28.36	0.000	.9100948 1.048915
_cons	4.999017	.42371	11.80	0.000	4.14754 5.850494


```
Command:
corr (lnoutput lnlabor)
```

And what I see is that, the R square value is going to be 0.9426. Now I want to see the correlation coefficient among these two. So, what I will do? I will simply write corr, command, and run it.

(Refer Slide Time: 01:33)

Results

```

. regress lnoutput lnlabor

```

Source	SS	df	MS	Number of obs	=	51
Model	89.8648125	1	89.8648125	F(1, 49)	=	884.22
Residual	5.47531855	49	.111741195	Prob > F	=	0.0000
				R-squared	=	0.9426
				Adj R-squared	=	0.9414
Total	95.340131	50	1.90680262	Root MSE	=	.33428

lnoutput	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnlabor	.9795049	.0345397	28.36	0.000	.9100948 1.048915
_cons	4.999017	.42371	11.80	0.000	4.14754 5.850494


```

. corr (lnoutput lnlabor)
(obs=51)

```

	lnoutput	lnlabor
lnoutput	1.0000	
lnlabor	0.9709	1.0000

Comment

2 Tables 1 Documents 1

And we see, it is going to be 0.97. So, it is 0.97, the small r, the correlation coefficient. And the R square value is 0.94. And why they are not the same? Because here, in the second table, I am only getting the R value, not the R square value.

(Refer Slide Time: 01:54)

Excel - 100%

Formulas Data Chart Layout Styles

File Home Insert References Send To Tell Me

Table 1

X1	X2	X3	X4
10	3	9	3
10	2	7	4
12	1	5	5
18	4	11	6
20	5	13	4

0.9706

0.94206436

Table Style 1

Font: Helvetica, Regular, 10 pt

Character Styles: Character Styles

Alignment: Left, 0 pt

Wing text in cell

Spacing: 10 - Single

Lines: 1

Before Paragraph: 0 pt

After Paragraph: 0 pt

If I actually do 9706; so, if I just simply take a square of that; so, what I am going to get? I am going to get the value of 0.942. I do not know if it is visible; but let me just increase the font. So, you can see, it is 0.942. So, that is basically the R square value. So, this is basically the R square, 0.942. Now, so, that is the case with the simple OLS. We have this equation, capital R square is equal to small r square.

But what happens when we have like a multiple regression? Because, then, the issue is that my R square, the correlation coefficient among the X and Y variables; so, like, if I take

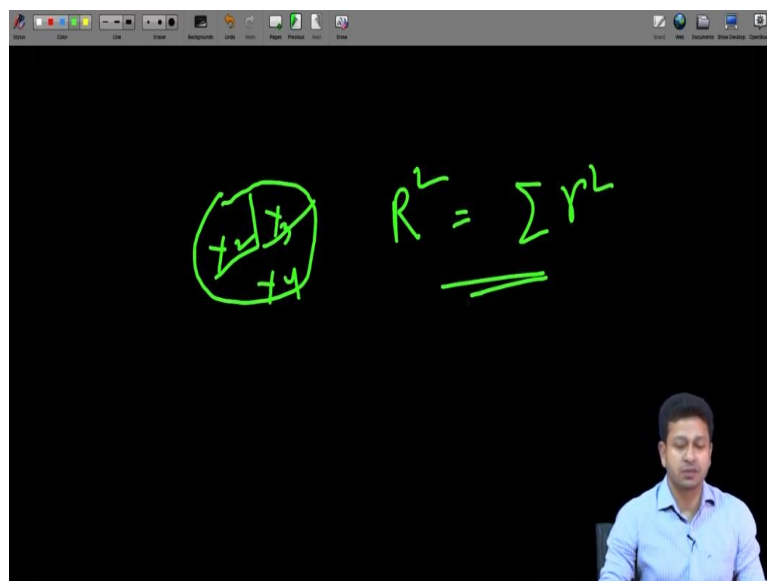
different correlation coefficient for different explanatory variable, they might actually overlap. So, we will see what actually happens in case of this multiple regression. So, let us say the variability of Y is explained by this circle.

Let me actually draw a big circle here; let us say this is the variability of Y. Now, what was happening previously? So, all the variability of Y, whatever variability we are getting, previously, when I was taking correlation coefficient, the R square was exactly the same, because the variability between X and Y was again explained by R square. So, they are explaining basically the same thing.

Now, and your R square is explaining the explained sum of square. So, it was essentially explaining the same thing. Now, the problem with multiple regression is that my different X variables, they have some amount of correlation among themselves. They have some amount of correlation among themselves. And usually what happens is that, if you run, these different X variables will actually explain the different extent of the Y.

And they will also explain each other. So, there will be multiple overlaps. And that is why you really cannot say in case of multiple regression that; ideally I could have said that R square is equal to essentially summation of all these small r squares, but that does not happen because the overlaps. But if for an ideal condition, let us say, in an ideal condition where all the r_{X_i, X_j} is equal to 0; if the correlation coefficient among all these explanatory;

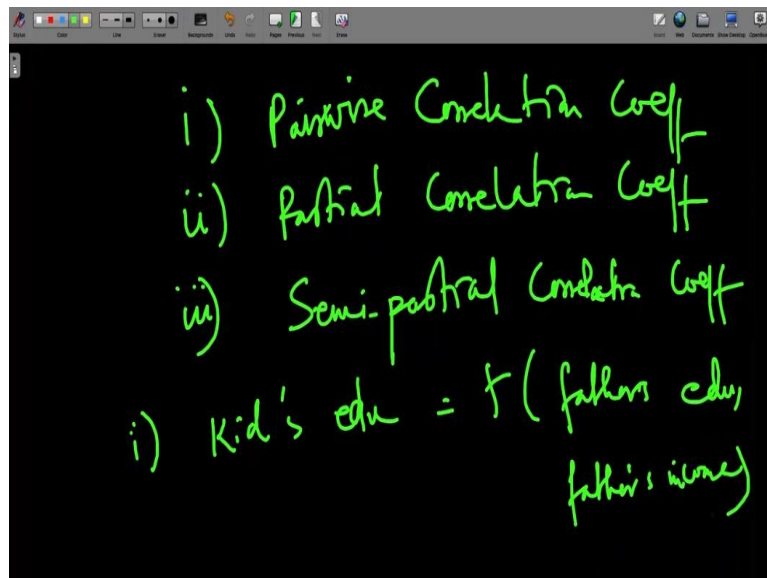
(Refer Slide Time: 04:46)


$$R^2 = \sum r^2$$

Will mean that there is no overlap. So, if X 1 is explaining this part, X 2 is explaining this part, and let us say X 3 is explaining this part; X 2, X 3 and let us say X 4; they are explaining these 3 parts. So, if I get something like this, then, in that ideal condition, my R square is going to be summation of all the small r square. So, that can happen, but only if this ideal condition satisfies, but that is very unlikely because the X variables will always have some amount of relationships among themselves.

Now, because of this overlap problem, we have derived some other types of correlation coefficient. We are just going to see that. So, one we have learnt is that, what we normally know is a pairwise correlation coefficient that we have been doing so far.

(Refer Slide Time: 05:30)



Now, we will learn something called partial correlation coefficient. And there is something called semi-partial correlation coefficient. We will try to understand these 3 terms, and how they are related. Now, we will try to understand that how each of these terms are related. So, before that, let me actually show you in Stata, how this pairwise correlation coefficient matters. So, we already have done pairwise correlation coefficient previously, where we run the code corr. So, let me go back to the table here.

(Refer Slide Time: 06:44)

	lnoutput	lnlabor	lncapital	labor	capital
lnoutput	1.0000				
lnlabor	0.9789	1.0000			
lncapital	0.9734	0.9604	1.0000		
labor	0.7776	0.8028	0.7666	1.0000	
capital	0.7441	0.7329	0.7553	0.9420	1.0000

Variable	VIF	1/VIF
lnlabor	23.13	0.043237
lncapital	20.15	0.049619
labor	16.98	0.058880
capital	14.07	0.071053
Mean VIF	18.58	

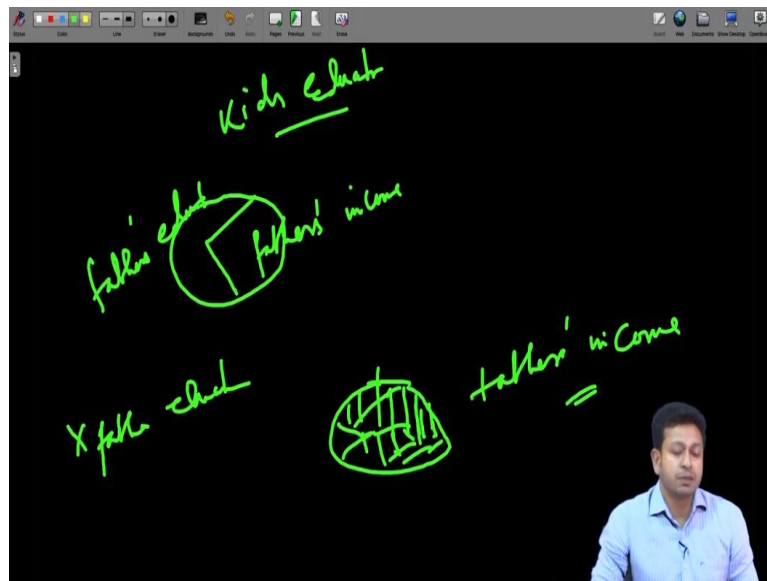
Source	SS	df	MS	Number of obs	=	51
Model	89.8648125	1	89.8648125	F(1, 49)	=	804.22
Residual	5.47531855	49	.111741195	Prob > F	=	0.0000
				R-squared	=	0.9426
				Adj R-squared	=	0.9414

So, here we see that in case of, when we talk about pairwise correlation coefficient, we see that labor, log of labor is actually explaining, like it is 0.97; it is very high correlation coefficient. So, we can say that okay, probably log of labor is actually explaining all the variability in the output. And the same time, if you take log of capital, it is actually explaining 0.97 again. So, it is like explaining everything, all the variability in the model.

Now, why it is happening? It is happening because; when you take one particular variable, so, what will happen? It will actually capture the variations of other variables as well. So, when you take the pairwise correlation coefficient, the case one, it will capture the variations due to other variables. So, we gave an example at the beginning that kid's education is actually some function of father's education and father's income.

So, that is what we said. Now, let us say these are the 2 variables which are very prominently explaining kid's education. Now, if I actually drop, let us say father's income or let us say father's education, whichever you want to drop; so, the other variable will actually; the variations that happen due to the other explanatory variables. The father's income will capture the variations that is happening due to father's education. So, that is where it will actually inflate the influence, the variations in kid's education.

(Refer Slide Time: 08:24)

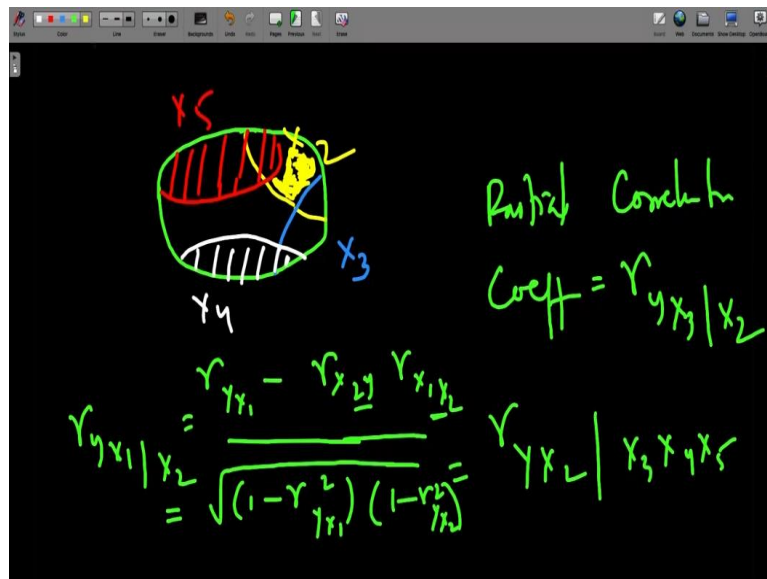


And if, let us say, this is how much father's income. Father's income is actually explaining. And this is, say father's education, this part. And the overall variability is kid's education. The moment you exclude father's education; so, the moment in your model you do not have father's education let us say; so, then, your model will, the variability like the father's income will explain something like this much.

So, almost like everything will be explained by father's income. Father's income will show the impact of other variables, because the other variables are not involved, are not included, so, all the variability of the other variables which are related to father's income, they will be reflected by father's income; because the different other variables which I am omitting, they are also related to father's income.

So, because now I do not include those explanatory variables in the regression equation, so, father's income is going to explain most of the variations. So, that is the problem with pairwise correlation coefficient. And that is why we have derived the term or derived the concept of partial correlation coefficient. And let me actually explain what I mean by partial correlation coefficient.

(Refer Slide Time: 09:58)



Now, I will use different colours here. So, let us say this is again the variability in Y. So, this is how we will try to understand graphically. It makes a lot of sense if we do it graphically. Now, if this is father's, the Y variable, variability of Y variable, and let us say this is the variation that is happening due to X 2, and this is the variation which is happening due to X 3. Now, what pairwise correlation coefficient does is that, it ensures that the variability due to other X variables are actually kept constant.

So, the variability due to other X variables are actually kept constant. Partial correlation coefficient could be expressed in terms of $r_{Y X_3}$, keeping my X_2 constant. So, you can have like many other X variables. So, here you can have X_4 or you can have something like X_5 . Here what you do is, in whatever number of variables you have, so, it is going to be say $r_{Y X_2}$; I just want to; I am interested in X_2 , let us say; $r_{Y X_2}$ keeping X_3, X_4, X_5 constant.

So, this is how we actually express the concept of pairwise correlation coefficient. So, that is how basically we should see how we should understand the pairwise correlation coefficient. Now, mathematical term for pairwise correlation coefficient is basically $r_{Y X_1}$ given X_2 , is going to be $r_{Y X_1} - r_{X_2 Y} r_{X_2 X_1}$ or $r_{Y X_2}$, whichever you want to write, into $r_{X_1 X_2}$ by square root of $1 - r_{Y X_1}^2$ into $1 - r_{Y X_2}^2$ square.

Essentially, you do not have to remember this formula. Essentially, it is just to give you an idea that the partial correlation coefficient, when you want to keep the impact of some X variable constant, so, you have to basically do this, you have to basically incorporate this

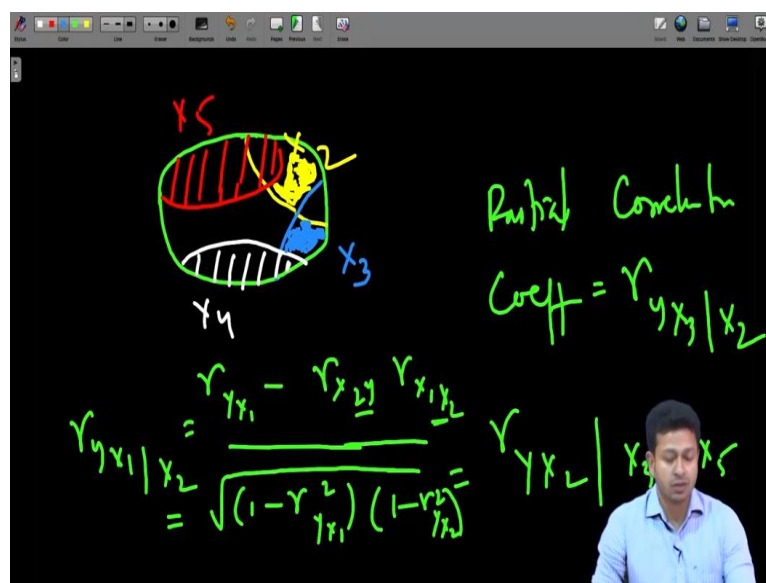
correlation coefficient values here. So, this is about the partial correlation coefficient. And what is the semi-partial correlation coefficient?

The semi-partial correlation coefficient is relatively simple. So, essentially, if I want to look at the partial correlation coefficient here; so, partial correlation coefficient for X 5 is going to be this much. So, it really does not take into account the correlation coefficient or the influence of other variables. So, everything else is constant. So, I will just get the impact of X 5 on Y.

Or if I want to take the partial correlation coefficient between X 4 and Y, so, I will get this much. So, the entire effect by X 4 is actually captured by partial correlation coefficient, because I am keeping the other terms constant. So, it is not inflating the importance of the X X 4 variables here. Now, the semi-partial correlation coefficient on the other hand is basically, shows the independent contribution.

So, what do we mean by that is essentially; let us say, I want to understand the impact of X 2 on Y, and I want to see the semi-partial correlation coefficient. So, what it will do? It will basically show this area where there is no overlap. So, if you want to understand the independent importance of a particular variable of, so far as the determining the dependent variable is concerned, you should look at the semi-partial correlation coefficient. And here, for the semi-partial correlation coefficient, we have, this is the area under the X 2.

(Refer Slide Time: 15:09)



And for X 3, this is going to be the area; so, where they are essentially independent; there is no overlap existing. So, naturally, whenever you get a semi-partial correlation coefficient, what you will get is, it is less than the partial correlation coefficient value. So, let us actually run our small command. And it is like partial correlation coefficient, we write pcorr.

(Refer Slide Time: 15:38)

Partial and semipartial correlations of lnoutput with

Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
lnlabor	0.9388	0.5894	0.8813	0.3474	0.0000
labor	-0.3890	-0.0913	0.1513	0.0083	0.0057
capital	0.4303	0.1031	0.1852	0.0106	0.0020

```
. pcorr(lnoutput lnoutput lnlabor labor capital)
too few variables specified
r(102);

. pcorr(lnoutput lncapital lnlabor labor capital)
(obs=51)
```

Partial and semipartial correlations of lnoutput with

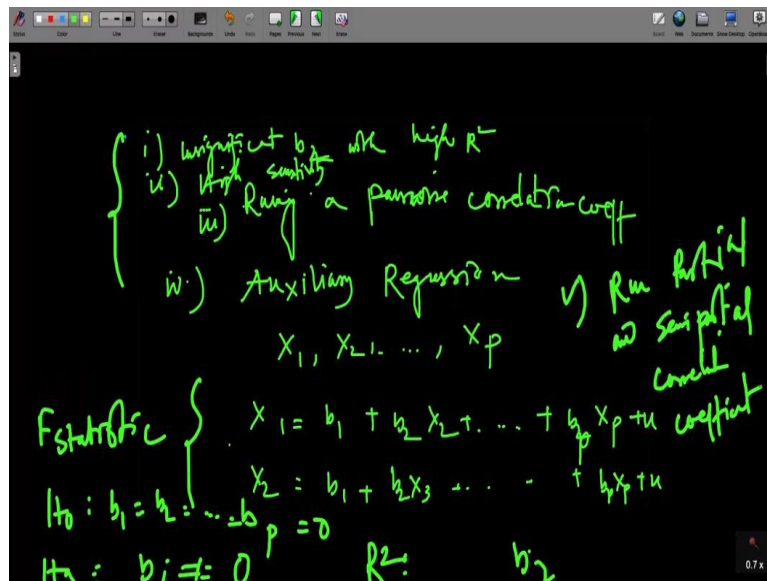
Variable	Partial Corr.	Semipartial Corr.	Partial Corr.^2	Semipartial Corr.^2	Significance Value
lncapital	0.4917	0.1064	0.2418	0.0113	0.0004
lnlabor	0.4951	0.1073	0.2451	0.0115	0.0003
labor	-0.0881	-0.0167	0.0078	0.0003	0.5516
capital	0.0984	0.0186	0.0097	0.0003	0.5060

So, now I get the partial correlation coefficient; I get semi-partial correlation coefficient; I get the square values, and so forth. So, here you see, the ln capital value is becoming 0.49, whereas ln labor is becoming 0.49. So, earlier, if you remember, we had, it is almost; where is that? Previously we have done that; almost 0.97 each, because they are actually explaining the variability due to other factors.

And here, I have like a very clean representation of the variability by, due to ln capital and ln labor; whereas, for only labor and only capital, they are very low. So, we got that. And semi-partial correlation coefficient, as we saw, it is going to be the smaller segment, that part where there is no overlap. So, this part is going to be the semi-partial correlation coefficient for each of this term. So, ln capital, ln labor.

So, usually, if you sum up the square of the partial correlation coefficient, it will be higher and it will be closer to R square, as opposed to if you just square and sum the semi-partial correlation coefficient. So, that is basically the idea of these different correlation coefficients. And when you try to understand the multicollinearity, it is advisable.

(Refer Slide Time: 17:04)



So, previously, we have said that, all these 4 different ways to understand multicollinearity. If there is present or not. So, I will add one more here. I will use the same colour as before. So, I will add one more here. And that fifth is, instead of running just simply pairwise correlation coefficient like we proposed in 3, we can propose a case of 5, where run partial and semi-partial correlation coefficient.

So, now we have seen all the 3 different important terms that we deal with a regression equation, the R square, the standard error and the beta coefficient. So, we have seen how we should look at all these 3 terms in case of multicollinearity problem. Now, with this, we will end this lecture. So far, we have talked about all these different ways to understand multicollinearity, and we have just given small data set to understand how the multicollinearity looks like.

So, finally, what we will do is, we will run a standard regression equation with multiple, different variables, and we will see all these problems, how they look like and how we are going to address them. So, thank you very much. With this, we will end this lecture here.