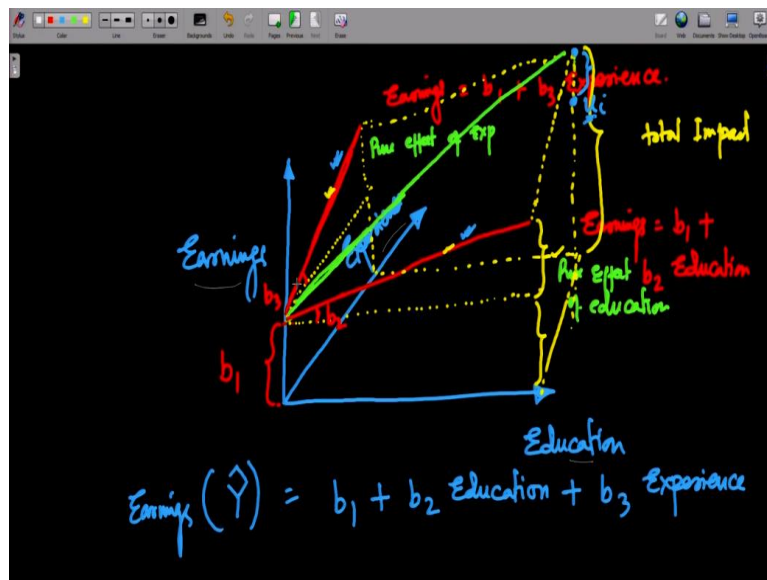


Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 6
Lecture - 49
Multiple Regression (Contd.)

Hello and welcome back to the lecture on Applied Econometrics. So, we have been talking about multiple regression. In the last lecture, we have actually shown you how a multiple regression would, we can visualise with the different explanatory variable.

(Refer Slide Time: 00:32)



We had education and experience as our explanatory variables. And we had our dependent variable earnings. So, we had 2 independent variable, 1 dependent variable. And I kind of showed you how can I represent that regression equation with this plane where I had drawn line corresponding to education and corresponding to earning. And then, we kind of got the result in plane. Now, that was a visual depiction of this whole idea. Now, how will it look when we actually work with the data set. So, let me actually show you a data set.

(Refer Slide Time: 01:12)

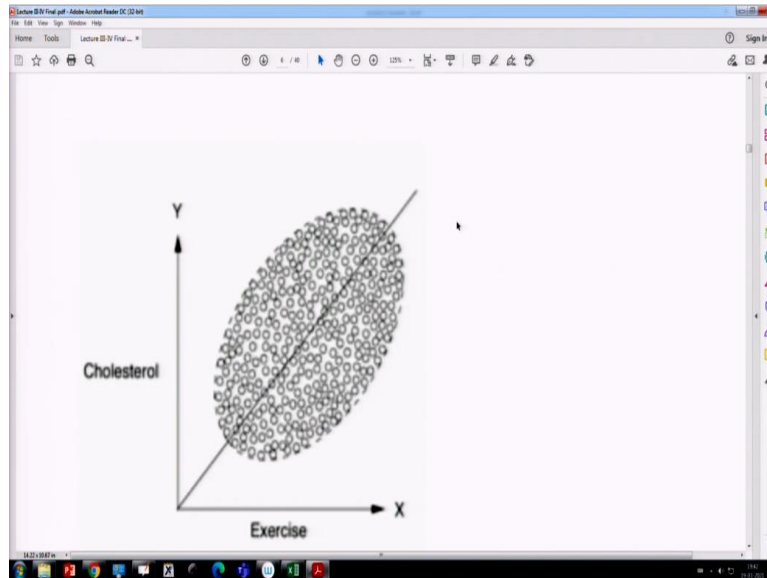
Cholesterol	exercise	age
100	35	61
98	36	61
95	37	60
90	38	60
81	25	51
80	26	51
77	27	50
72	27	50
62	20	43
60	21	42
58	22	41
56	23	40

Let us say I want to explain the cholesterol level of people with their level of exercise and with their age. So, intuitively we can understand that cholesterol level might have some relationship with exercise, and cholesterol level might also have some relationship with age. So, what should be the intuitive understanding behind it? So, if I do exercise, usually, I can say that my cholesterol level probably would go down.

And with age, usually, we know that, as people grow old, the cholesterol level actually increases a little bit. And let us say this is bad cholesterol, we are not talking about good cholesterol. So, we will see that these numbers we have, and we have like observation for 12 individuals, let us say. We are keeping the dataset short, so that we can just get the concept here. And we also have the exercise.

There are some units, let us say this is the amount of time, like everybody spends every day in exercise. And age is age in years. Now, we need to understand how the concept of multiple regression would play a role here to explain the cholesterol level as an outcome of exercise and age. Now, let me actually show you the plot.

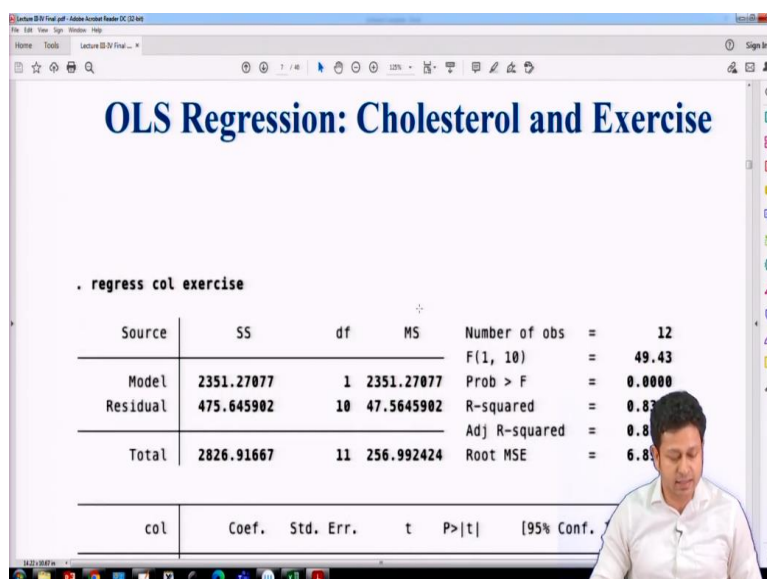
(Refer Slide Time: 02:33)



So, actually, what I have plotted here is the cholesterol and exercise. So, my cholesterol is a Y variable, the dependent variable, and my exercise is the explanatory variable, the X variable. Now, if we look, if we go back; and you will see that there is something interesting about this diagram. So, the interesting part of this diagram is; so, I see that the, my cholesterol level is actually increasing with my level of exercise.

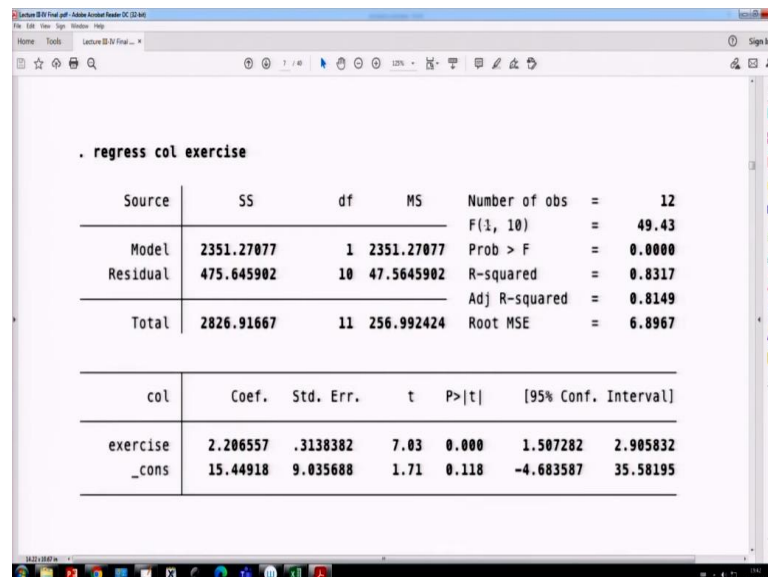
So, that is kind of counterintuitive; because, if I am doing exercise, usually, I can think that my cholesterol level should go down; but it is happening on the contrary. Now, what I am doing here? So, let us say, then, now that we know a little bit of regression, we actually run a regression. And let us say I have run the regression. So, the first regression here, I will just show you here is this.

(Refer Slide Time: 03:24)



So, because I am interested to know the impact of exercise and cholesterol, what I have taken here is, I simply have taken cholesterol as the dependent variable and my exercise as the independent variable. Now, if I actually see the regression result here; let us actually explain the regression results. We will actually talk about this whole table later on.

(Refer Slide Time: 03:46)



The screenshot shows a software window with a command bar containing the text `. regress col exercise`. Below the command bar, there are two tables. The first table is a summary of the regression results, and the second table is a table of coefficients.

Source	SS	df	MS	Number of obs	=	12
Model	2351.27077	1	2351.27077	F(1, 10)	=	49.43
Residual	475.645902	10	47.5645902	Prob > F	=	0.0000
Total	2826.91667	11	256.992424	R-squared	=	0.8317
				Adj R-squared	=	0.8149
				Root MSE	=	6.8967

col	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exercise	2.206557	.3138382	7.03	0.000	1.507282 2.905832
_cons	15.44918	9.035688	1.71	0.118	-4.683587 35.58195

But for now, we will just focus on the coefficient for these different explanatory variables and the P-value; just these 2 items we will focus on for now. Now, if I run the regression between cholesterol and exercise, what I see is that, the exercise has a coefficient which is 2.206. And it has a corresponding P-value which is approximately closely equal to 0. So, we know that for a very low P-value, we reject the null hypothesis.

And our alternative hypothesis, that is the exercise is actually impacting the cholesterol level; we kind of accept that. Now, if that is the case, then the exercise is actually, what it shows? The result shows that exercise has a positive impact on cholesterol levels. So, that means, if you do exercise, your cholesterol level is actually increasing, right? And that too, it is significant, because the P-value you get here is actually close to 0.

So, that is actually showing the significance. Now, that does not make sense really, right? I mean, how come the exercise is actually impacting, like increasing the cholesterol level? Now, being curious about this whole scenario, what I did is actually, I thought, let me actually run a regression where I will have the cholesterol level in my Y axis. So, here, this cholesterol level, I am representing at col as a variable, and the independent variable as age, in the X axis.

(Refer Slide Time: 05:19)

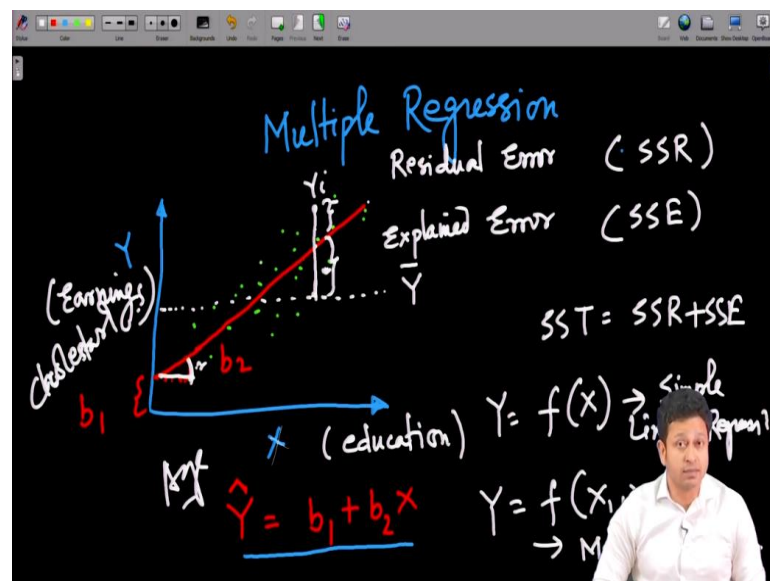
`. regress col exercise age`

Source	SS	df	MS	Number of obs =	12
Model	2783.58591	2	1391.79296	F(2, 9) =	289.08
Residual	43.3307532	9	4.81452813	Prob > F =	0.0000
Total	2826.91667	11	256.992424	R-squared =	0.9847
				Adj R-squared =	0.9813
				Root MSE =	2.1942

col	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
exercise	-.7390924	.3264973	-2.26	0.050	-1.477681 - .0005042
age	2.516963	.2656155	9.48	0.000	1.916099 3.117827
_cons	-29.77275	5.571237	-5.34	0.000	-42.37577 -17.16974

So, let us see what happened here. Now, when I run this regression, what I see is that, well, I am again focusing on the coefficient age and its corresponding P-value. And what I see is, age is again positive 1.9; so, which means, if I go back to the previous diagram, which would mean like, if I take one variable, let us say this one X variable, and let us say this is age.

(Refer Slide Time: 05:45)



And if I actually plot it here, so, what I will have is that, the coefficient value which I am getting as 1.94, so, it will be something like that. So, that is basically represented by this beta coefficient, only for age, when I do. So, it will somewhat show, if I kind of use this diagram for representing the cholesterol and age, I can actually write cholesterol and age in this axis. And if I plot it, it will show a positive trend.

And the same thing is true when we have done cholesterol and exercise. Now, what is going on here. So, it means that this table actually makes sense, because, with age, cholesterol level is increasing. That kind of makes sense; that is intuitive. But the previous one, the cholesterol level is also increasing with exercise; that really does not make sense. Now, how do I this result? And actually, to make sense of this result, well as, with 1 explanatory variable is not sufficient.

This diagram, this visualisation, this concept of 1 variable regression is not able to actually sufficiently explain the phenomena that is happening. So, what we have to do is, we actually have to do a multiple regression or table to see how the multiple regression is playing a role here. So, let me actually run a regression where I have plotted both cholesterol as a dependent variable and the age and exercise both as independent.

Now, when I run this regression with both the variables, exercise and age as dependent variable, we see something different. And what we see here? We see that exercise is now, it has become negative. You see it is a negative sign to it, right? Whereas age, it has remained positive. And if we know how to read the P-values, so, we will know that the P-value here for exercise is significant, it is 0.05 and P-value for age is of course significant, I mean, it is kind of showing the significance; so, it is 0.000 which is basically very close to 0.

So, that means, the area basically for the corresponding t is very low. So, that means, you accept the alternative hypothesis or you reject the null hypothesis. So, that means, you reject the null hypothesis that age does not have any impact on cholesterol level; similarly, exercise does not have any impact on the cholesterol level. So, these are the null hypothesis, and we basically reject them.

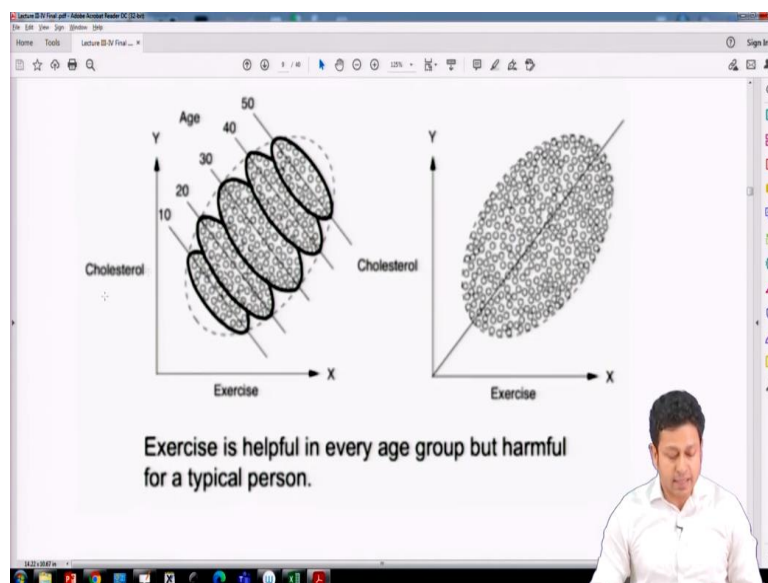
So, both the variables are actually significant, but the interesting part is, exercise is here showing a negative coefficient and it is significant; whereas, age continued to remain as a positive coefficient and is significant. So, now it is making sense. So, it shows that, the moment you do multiple regression, some part of the variability is actually explained by age, and some part of the variability is actually explained by our exercise.

The moment age is actually explaining its own part and whatever is left out is explained by your exercise, that is actually making a lot of sense, that yes, so, with age, my cholesterol

level is increasing, but with exercise, my cholesterol level is actually decreasing. And you see that in all the cases, this table R-squared value is very high; it is like 0.98. And the previous tables also, we had R-squared value very high.

It was something like, for exercise is 0.81; for age, it was 0.97 and so forth. But to understand what is happening here, actually you have to run the multiple regression where you will have both the explanatory variables present. Now, let us try to make sense of it.

(Refer Slide Time: 09:51)



And if we actually try to draw a diagram, so, how it is different? In the first diagram, we have seen this one, where we had like this, we plotted cholesterol and exercise, and we saw something like this. Now, what actually is happening is this. So, the moment we actually bring in age, it makes a lot of sense. So, what is happening here is that, once we draw these different lines, different segments for this different age group, so, 10, 20, 30, 40 and so forth, what we get is, for each of this age group, the cholesterol level is definitely increasing 10 to 20, 20 to 30, 30 to 40 and 40 to 50, so forth.

Now, for each of this age group, what is happening is that the people of that 10 age group keep, when they are doing the exercise, those people are having low cholesterol. So, for a given age, if you do exercise, then your cholesterol level is actually declining. So, it is true for all the different age groups. So, if you consider only age group of 10, they are having the, with exercise, their cholesterol level is decreasing.

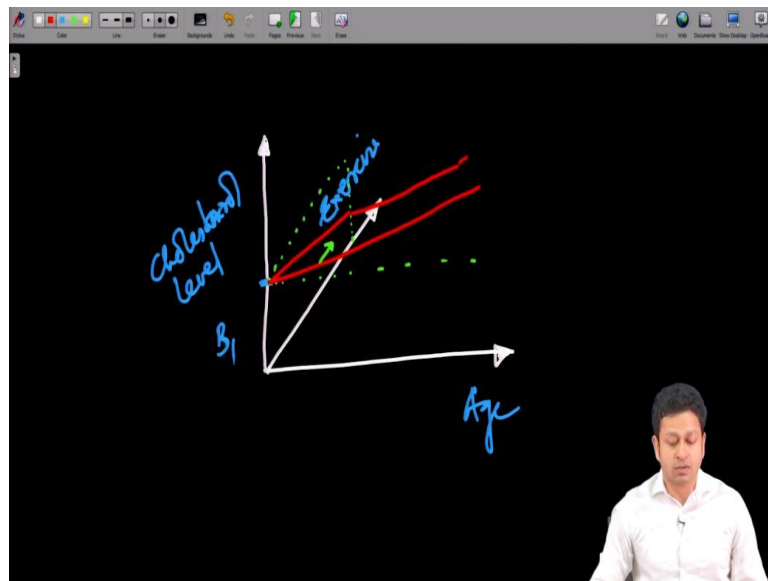
If you are taking the age group of 20, if you would exercise, your cholesterol level is declining. If your age group is 30, if you would exercise, your cholesterol is declining, and so forth. So, for each of the age group, we see the same phenomenon. So, essentially, it means that, the moment you keep the age constant, the moment you take the variability away due to age, you can actually see how exercise is playing role.

And then only it makes sense that; that actually tells you that how exercise is actually helpful to get rid of the cholesterol problem. So, that is what the table shows. So, essentially, if we do not consider the age, it will show that exercise is actually harming us, exercise is actually increasing the cholesterol level. So, for a typical person, if I do not consider the age, like, if you take the whole sample, the whole population, it will show that exercise is actually harming you, because you have not controlled the age; but the moment you control for the age, you will see that exercise is actually helpful.

So, that is very radically different interpretation. But only if you are able to use the concept of multiple regression, it is likely that you will come to explanation which is close to reality. So, that is the idea of multiple regression. And I took this example is a very famous example; it is called Simpson's paradox. So, you can actually search for it. So, essentially, we will see going forward.

This is a very, like, it is used for illustration purpose, it is prominent, but it could be much more tricky when you do multiple regression in terms of how these variables are related and so forth. We will see that. But before I proceed to the next topic, let me actually try to explain how it is happening. So, this is what we have seen for education on earning.

(Refer Slide Time: 12:50)



But if I take this example where we explained how exercise and our age is actually explaining the cholesterol level, so, we have this cholesterol level; explain it correctly. And here I have my age, and here I have my exercise. So, what is happening? So, let us say this is the, our constant term. So, let us say beta 1. And then what I will do? I will use a different colour first. So, this is for age.

And I know that what is happening for age is, actually it is increasing the cholesterol level; but whereas for exercise, what is happening is, it is actually decreasing. So, if this is the case, I can actually do a resultant; I can actually draw something resultant of it. Let us say I try to project it. Let us say I will project this line here. This line will move to here. And I will project it here.

So, essentially, if I draw parallel line here, so, that will kind of give me the plane where I will have the resultant value. So, it is actually, the resultant is actually going to be little less than the overall effect of age. So, exercise is kind of taking it down. So, if you see the resultant plane, it is going to be somewhat lower than if you just draw with the age. So, that is basically the idea of multiple regression.

And in the next lecture; so, we have kind of given you these tables, but we have not really explained the different terms which are there in the table. And in the next lecture, I am going to explain these different terms of regression table. Thank you.