

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

Lecture - 46
Adjusted R-Squared

Hello and welcome back to the lectures on Applied Econometrics and we are in module 2. And in this lecture we are going to talk about adjusted R square. So we are talking about different terms associated with the regression equation and regression table, the way we understand it.

And it is something that is important for penalizing if we have number of different variables which are irrelevant in regression equation and we are going to just see that how adjusted R square is a better measure for coefficient determination that is represented as R square. So, let us try to understand how adjusted R square works.

(Refer Slide Time: 01:00)

Mincerian Wage Regression Equation									
. regress wagetotal genedu									
Source	SS	df	MS	Number of obs	=	25,488			
Model	2.9466e+09	1	2.9466e+09	F(1, 25486)	=	1932.51			
Residual	3.8859e+10	25,486	1524735.58	Prob > F	=	0.0000			
Total	4.1806e+10	25,487	1640286.32	R-squared	=	0.0705			
				Adj R-squared	=	0.0704			
				Root MSE	=	1234.8			
wagetotal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]				
genedu	100.6434	2.289414	43.96	0.000	96.156 105.1308				
_cons	-250.7105	15.42338	-16.26	0.000	-280.9412 -220.4798				

• X = only human capital variable is education

So let us see, let us first run a regression and I actually have run the regression. So you have to trust me on this. I used national sample survey data, where I took my dependent variable as wage and my independent variable as general education. And what I found is that I have R square value is 0.0705 whereas my adjusted R square value is 0.0704, right?

And I keep on increasing and basically what I, the first thing is slightly, adjusted R square is slightly less than the R square value.

(Refer Slide Time: 01:36)

Mincerian Wage Regression Equation

. regress wagetotal genedu sectornew

Source	SS	df	MS	Number of obs = 25,488
Model	3.0593e+09	2	1.5296e+09	F(2, 25485) = 1006.10
Residual	3.8747e+10	25,485	1520371.94	Prob > F = 0.0000
Total	4.1806e+10	25,487	1640286.32	R-squared = 0.0732
				Adj R-squared = 0.0731
				Root MSE = 1233

wagetotal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
genedu	95.77983	2.354874	40.67	0.000	91.16414 100.3955
sectornew	139.7359	16.22777	8.61	0.000	107.9285 171.5432
_cons	-418.2336	24.81385	-16.86	0.000	-466.8686 -369.5986

• X = only human capital variable is education and other variable is regional characteristics

And again I add another variable here. So this is called sector. Sector means rural, urban. I have included a dummy variable. We will explain what is dummy variable later on. Here again, I see my adjusted R square is slightly less than my R square.

(Refer Slide Time: 01:48)

Mincerian Wage Regression Equation

regress wagetotal genedu sectornew sex

Source	SS	df	MS	Number of obs = 25,488
Model	3.9880e+09	3	1.3287e+09	F(3, 25484) = 895.29
Residual	3.7820e+10	25,484	1484066.51	Prob > F = 0.0000
Total	4.1806e+10	25,487	1640286.32	R-squared = 0.0953
				Adj R-squared = 0.0952
				Root MSE = 1218.2

wagetotal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
genedu	87.95395	2.34757	37.47	0.000	83.35257 92.55532
sectornew	148.6451	16.03681	9.27	0.000	117.212 180.0781
sex	-384.8965	15.40261	-24.99	0.000	-415.0865 -354.7065
_cons	188.4689	34.50257	5.46	0.000	120.8339 256.0879

• X = only human capital variable is education and other variable is regional characteristics and gender variable

And I keep on adding different variables and I see that the adjusted R square remains slightly less than the R square.

(Refer Slide Time: 01:54)

Mincerian Wage Regression Equation

regress wagetotal genedu sectornew sex consump30 consump365

Source	SS	df	MS	Number of obs	=	25,488
Model	4.3633e+09	5	872657452	F(5, 25482)	=	593.90
Residual	3.7443e+10	25,482	1469378	Prob > F	=	0.0000
				R-squared	=	0.1044
				Adj R-squared	=	0.1042
Total	4.1806e+10	25,487	1640286.32	Root MSE	=	1212.2

wagetotal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
genedu	77.57999	2.426704	31.97	0.000	72.82339 82.33659
sectornew	97.87598	16.34893	5.94	0.000	65.84741 129.1045
sex	-397.4577	15.34624	-25.90	0.000	-427.5372 -367.3782
consump30	.0050382	.0007305	6.82	0.000	.0035980 .0064857
consump365	.0007461	.0001709	4.37	0.000	.0004111 .0010811
_cons	203.4703	34.63542	5.87	0.000	135.5829 271.3577

- X = only human capital variable is education and other variable is regional characteristics. I am also including consumption to do an experiment

It is actually even you know, the R square value is increasing, but the adjusted R square value is you know, it is also increasing, but the difference is increasing at the same time. So adjusted R square is lesser than just R square, okay.

(Refer Slide Time: 02:09)

Mincerian Wage Regression Equation

regress wagetotal genedu exp expsq sectornew sex consump30 consump365

Source	SS	df	MS	Number of obs	=	25,488
Model	5.4380e+09	7	775717255	F(7, 25480)	=	543.36
Residual	3.6376e+10	25,480	1427627.81	Prob > F	=	0.0000
				R-squared	=	0.1299
				Adj R-squared	=	0.1296
Total	4.1806e+10	25,487	1640286.32	Root MSE	=	1194.028

wagetotal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
genedu	66.5733	2.536701	26.24	0.000	61.60123 71.54538
exp	26.58767	1.072788	24.78	0.000	24.48495 28.6904
expsq	-.3616782	.0190915	-18.94	0.000	-.3990995 -.3242508
sectornew	78.78556	16.14366	4.88	0.000	47.06306 110.3481
sex	-407.9043	15.13247	-26.96	0.000	-437.5648 -378.2438
consump30	.0055341	.0007299	7.58	0.000	.0041034 .0069648
consump365	.0007654	.0001685	4.54	0.000	.0004352 .0010957
_cons	40.33571	34.85798	1.33	0.184	-21.98788 114.0593

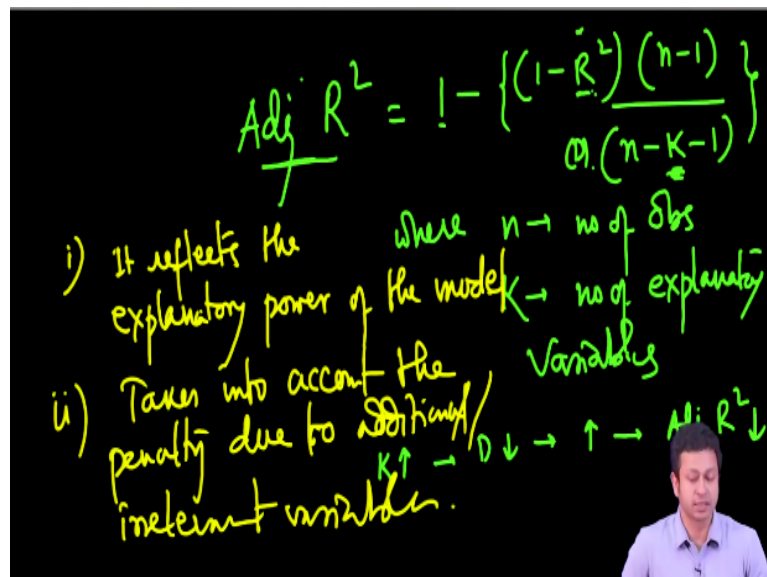
Here what do you see? Now, what exactly is adjusted R square and that is what we are going to explain here in this lecture. So when I talk about the regression equation, and we talk about the different variables we include in regression equation. So what happens is that as I include the number of regression, you know, the independent variables in the regression equation, I would actually see my R square value to increase.

And precisely that is what we have seen. All the different variables I have included, the R square value is actually increasing. So there might be a case that all the variables are important in explaining the R square. But there might also be a case that the variables are actually irrelevant in the model and it is not actually explaining our R square.

Now what happens if the second is true when I have included explanatory variables, which are not really you know relevant to my model? But you know, but still because of their presence, my R square value would increase. And the reason is, sometimes what happens is there are so many, you know, different variations in the data that could somehow be captured by a variable, which is actually irrelevant to the model.

And that is, because of that, as you keep on increasing the explanatory variable your R square value will increase, but you do not want that. So what you want is to kind of penalize if the number of explanatory variable increases in your model.

(Refer Slide Time: 03:38)



Adjusted $R^2 = 1 - \left\{ \frac{(1 - R^2)(n-1)}{n - K - 1} \right\}$

i) It reflects the explanatory power of the model where $n \rightarrow$ no of obs
 $K \rightarrow$ no of explanatory variables

ii) Taken into account the penalty due to adding irrelevant variables.
 $K \uparrow \rightarrow D \downarrow \rightarrow \uparrow \rightarrow \text{Adj } R^2 \downarrow$

So what we do is we actually use this formula is equal to 1 minus 1 minus R square into n minus 1 by n minus K minus 1 where n is the number of observation and K is the number of explanatory variables, K is the number of explanatory variables. So what is happening here? So if I actually increase my K, if I increase my K what will happen is that my denominator here it will reduce, right?

So increase in K means the denominator let us say it is D , the D will reduce and that means the whole term here will increase, right? The whole term here will increase and which would mean if I subtract it from 1, the adjusted R square is actually going to decrease, right?

So you can see the how the change in value of K , which is my number of explanatory variable, or increase in the number of explanatory variable is actually going to influence my adjusted R square, right? So this is how we actually include this penalty term to kind of ensure that adding more variables is also getting penalized.

Now if I add more variables, I can clearly see that it is getting penalized, but suppose my variables are actually helpful in explaining the model. So what will happen? What will happen just what you have seen, the R square value will keep on increasing, right? R square value will keep on increasing because it might be that your variable is actually relevant to your model and R square is increasing.

So if it does, then what will happen is that your, of course it will, there will be a penalty term because of the K , but then your R square value is actually increasing. So what will happen the whole, you know, value adjusted R square value will increase also because of the original R square value.

So, if I keep on adding the relevant variables, because the R square value is increasing, the despite the fact that I am offsetting with this K , my adjusted R square value will also increasing. So that is, what I mean to say here is that the adjusted R square value is accounting for the you know penalty due to additional variables, but at the same time, if my model, explanatory power of my model is increasing, that is also being reflected by my adjusted R square.

So let us write it down. So two things here. It reflects the explanatory power of the model. And at the same time it takes into account the penalty due to additional or irrelevant variables. So this is how we should understand the role of adjusted R square in explaining a regression table. So thank you.