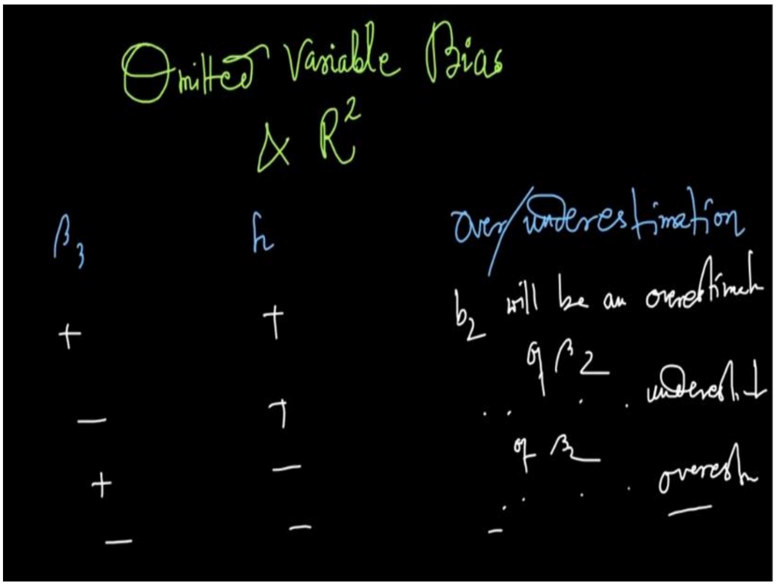
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Lecture – 87
Model Specification - Continued

(Refer Slide Time: 00:26)



Hello and welcome back to the lecture on applied econometrics. We have been talking about model specification and within model specification we have been detailing the problem of omitted variable bias. Now, in this lecture we are going to see with some examples what happens if we have omitted variable bias? How do we see the underestimation and overestimation of the coefficient? And what happens to R square because of that?

So, let me first actually write down the relationship between h, beta 3 and our beta 2 that we have derived in the previous lecture. So, let me write down. So, we have our beta 3, we have seen there is h which is nothing but the regression coefficient when we ran a regression between two explanatory variable X 2 and X 3 and then we have the problem of overestimation or underestimation.

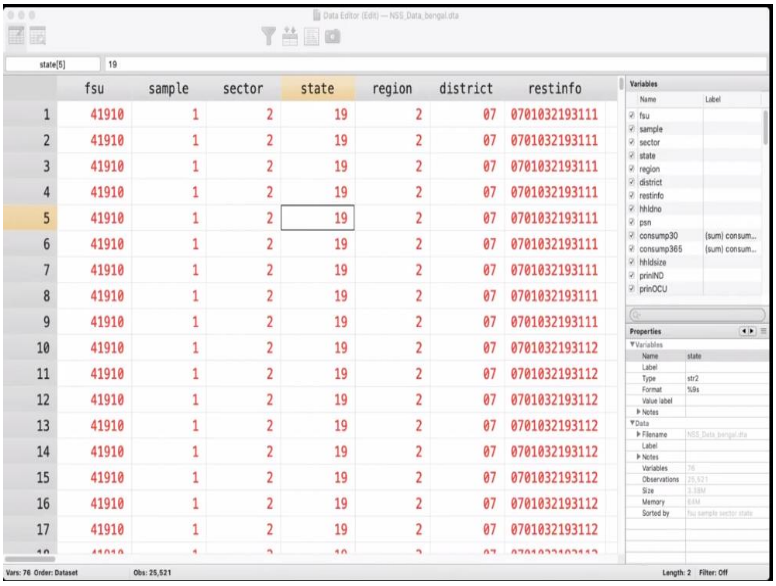
Now, we had 4 different cases. In first case we had both positive. So, when the relationship between the omitted variable and the Y variable is positive that is beta 3 is positive and when the relationship between these two explanatory variable that is X 2 and Yor X 3 is positive, so then h is positive. So, then what we will have is a problem of overidentification. So, beta 2 will be an overestimation of beta 2 because here we have omitted beta 3, so beta 2 is kind of is going to capture both beta 2 as well as some effect of beta 3.

So, that is something that you understood already. So, the other possible cases if let us say the beta 3 is negatively or X 3 is negatively related to the Y, so then the beta 3 is going to be negative whereas I can still have my X 3 and X 2 positively related in which case h is going to be positive, and the other cases just the opposite. And in both the cases we have seen that b 2 is going to be an underestimation of beta 2.

So, we have explained that with examples. Whereas in the last case, we have both beta 3 and h negative that is the X 3 variable is negatively related to Y and h that is a regression coefficient between X 2 and X 3 that is also negative that is X 2 and X 3 are negatively correlated. So, then again we will have the problem of overestimation. So, in this case b 2 is going to an overestimation of beta 2.

So, these 4 cases we have already seen, but now what are you going to do is that instead of understanding this theoretically, we will try to see some examples and we will try to see what is happening when we look into these all possible cases and what is happening to R square because of that most importantly.

(Refer Slide Time: 03:34)

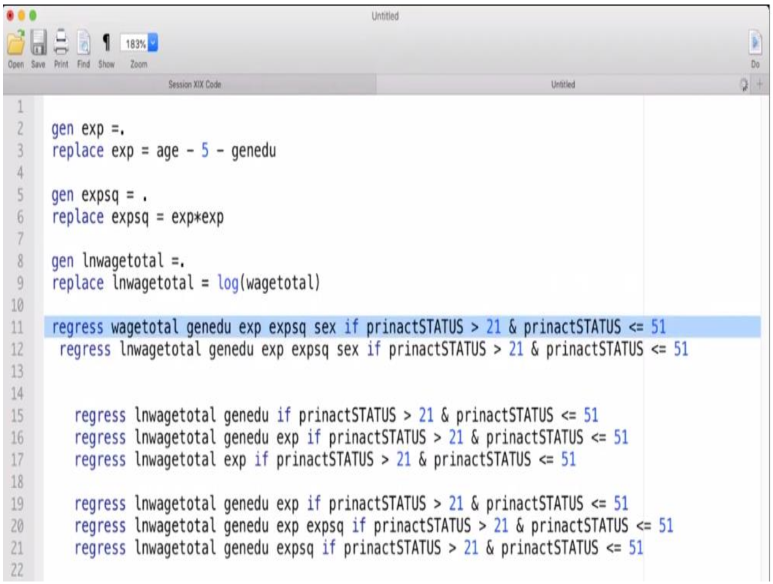


	fsu	sample	sector	state	region	district	restinfo
1	41910	1	2	19	2	07	0701032193111
2	41910	1	2	19	2	07	0701032193111
3	41910	1	2	19	2	07	0701032193111
4	41910	1	2	19	2	07	0701032193111
5	41910	1	2	19	2	07	0701032193111
6	41910	1	2	19	2	07	0701032193111
7	41910	1	2	19	2	07	0701032193111
8	41910	1	2	19	2	07	0701032193111
9	41910	1	2	19	2	07	0701032193111
10	41910	1	2	19	2	07	0701032193112
11	41910	1	2	19	2	07	0701032193112
12	41910	1	2	19	2	07	0701032193112
13	41910	1	2	19	2	07	0701032193112
14	41910	1	2	19	2	07	0701032193112
15	41910	1	2	19	2	07	0701032193112
16	41910	1	2	19	2	07	0701032193112
17	41910	1	2	19	2	07	0701032193112

So, let me first actually take you to data file, so let me actually take you to this data editor part. So, here we actually have the national sample survey data and that is only for West Bengal, so state is 19. And so what do you want to do here? We want to do a regression of wage, which is my Y variable whereas I will have all the independent variables like education, experience, experience square, and so forth.

So, what I am going to do in this lecture is that I am going to actually show you the changes or basically if we omit a variable and then again include a variable, so how the coefficients are going to change their value and how the r square is going to change this value.

(Refer Slide Time: 04:18)



So, first let me actually take you to; so there are 2 files, so I actually have done this little bit of work. So we have defined experience and as I mentioned experience is not provided in the national sample survey data. So how we do it? We usually take age minus 5 minus general education assuming that someone starts the class 1 when they are 5 years of age and then the general education is the years of general education.

So if you subtract these two, you will get experience approximately assuming that people start working after their schools without any break or anything. And experience square is basically the square of the experience. You basically just multiply experience with experience. You can also have log of wage instead of just wage, so wage total is basically the numerical value of the wage, all the wages one received.

But log of wage total is something that theoretically we have seen or empirically we have seen that is actually a better fit; of better wide than wage total because I get a better model. So, this is a problem of functional specification. Now, we have run, we have basically written down several regression equation. In the first equation we have written general, so wage total is just the wage total, not the log, and my X variable is general education, experience, experience square and sex.

And here this principle occupational status is above 21 or less than 51. That means people, so this essentially tell basically creates a group of people who are not self employed, who are not unemployed, and so forth. So, these are the people who are basically wage earners. So, either they are casual or they are salaried, but all of them actually earn some wage, so that is basically these numbers mean.

So if you look at national sample survey data and if you look at the principal activity status, you will see all these codes, so I have just taken the code from there. Great. So this next regression instead of wage total I have actually done a log of wage total because I know it is a better fit for this particular type of regression. Going forward, what we have done? We have just taken general education. So we are now entering into the territory of omitted variable bias, which you want to explain.

So I have omitted all these variables experience, experience squared and sex. Let say in the next regression equation, I have included general education and experience and in this block the last regression equation, I have only have the log of wage total and experience instead of general education and experience. So let us we will see the assumptions here and we will see how the coefficients behave and how the R squared behave.

And then in next block, what we have done is we have taken basically the log of wage total, general education, experience. In the next regression equation log of wage total, general education, experience and experience squared. And then we have taken log of wage total, general education and experience squared, I have omitted experience here. So we will see what are the different assumptions and results.

Let me actually, run this regression in the first place. So let me actually go there to the data editor and I will just run the regression again. So, I am not running this part, this already knows what I am talking about so how the variables are defined. So if I run it against, it will give me an error.

(Refer Slide Time: 07:58)

Model	2076.35749	4	519.089374	Prob > F	=	0.0000
Residual	2266.54065	4,254	.532802222	R-squared	=	0.4781
				Adj R-squared	=	0.4776
Total	4342.89815	4,258	1.0199385	Root MSE	=	.72993

lnwagetotal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
genedu	.1838988	.0031687	58.04	0.000	.1776864	.1901112
exp	.0555027	.003269	16.98	0.000	.0490937	.0619116
expsq	-.000595	.0000561	-10.61	0.000	-.000705	-.0004851
sex	-.2830172	.0297511	-9.51	0.000	-.3413448	-.2246896
_cons	5.213611	.0603429	86.40	0.000	5.095307	5.331914

.
end of do-file

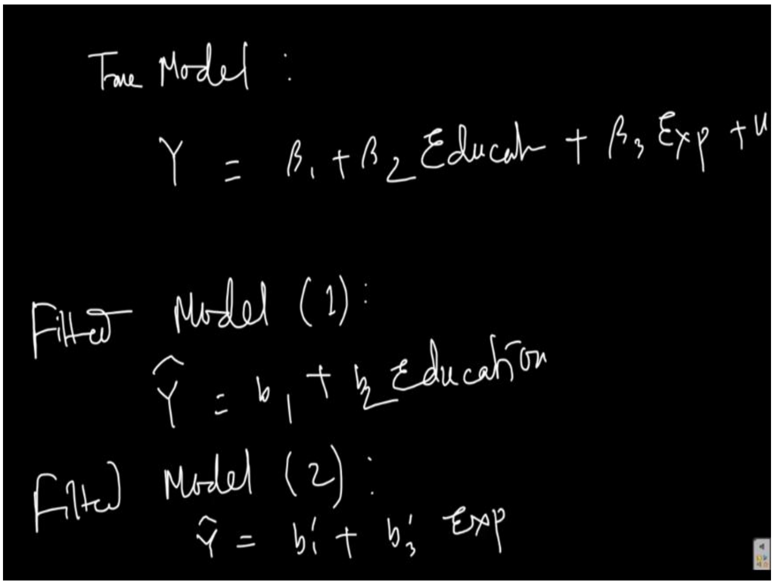
So let me first run this one, this regression and if I run this regression, what I will see. I need to actually go to the this file, and this is the regression equation, the results. So we see R squared is actually 0.38. I have general education, experience, experience squared and sex. Not too bad, not too good. But I know for a fact that I actually need to regress this log of wage total, it is a functional specification problem.

So if I run it, what I get is this, if I go there I see that, let me actually go back again, somewhere click this. So I do not need to define it again, let me see, it should not be any problem anymore. So if I go there, so now I have my regression equation. So now r squared has improved quite a lot, it is 0.47, so pretty good model and it has happened because I have taken log of wage total as a Y variable, my X variables are constant, so just to show you the functional specification problem here.

Now all these variables you see are basically significant, so the P value is pretty small. Now, what we want to do here is we actually want to go to omitted variable bias problem. So let us say my actual model is for the timing, I do not get into sex and experience squared, I will just talk about let say my real model is this one, let me show you my real model is this one, the general education and experience let us say.

So, if I have my regression equation as X variable as only general education that is wrong basically that will have an omitted variable bias problem. And in this model also if I only have experience, so that will also have omitted variable bias problem. So in both the cases, I have to basically address that and we will see what happens So let me actually write down these equations here.

(Refer Slide Time: 10:12)

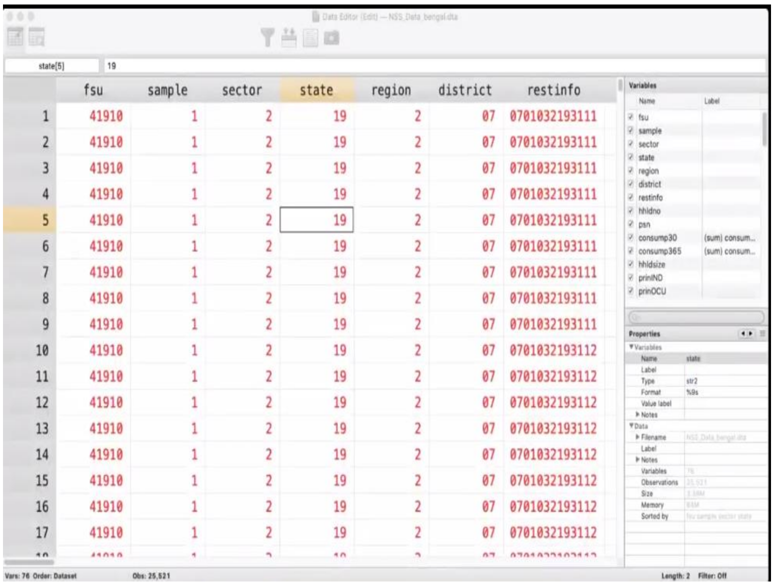


So essentially, in this case my true model is Y = beta 1 + beta 2 into let us say education, and let us say beta 3 into experience and then there is some term. Let us say my fitted model 1 is Y = beta 1, I have fitted model so b 1 + b 2 into education, we must be careful about the notations here. And because it is a fitted model I do this and the fitted model 2 is Y hat is equal to let us say some b 1 prime and b 2 or let us say b 3 prime let us say experience.

So now basically this is exactly what we have done where we have done the regression part. So if we go back, so this is do-file, this is this exactly what you have done in these equations or variable, taken only general education, I have only considered the b 2. When I have only taken experience, I considered the b 3 and I have actually omitted the general education component. So let us see what happens if we run this.

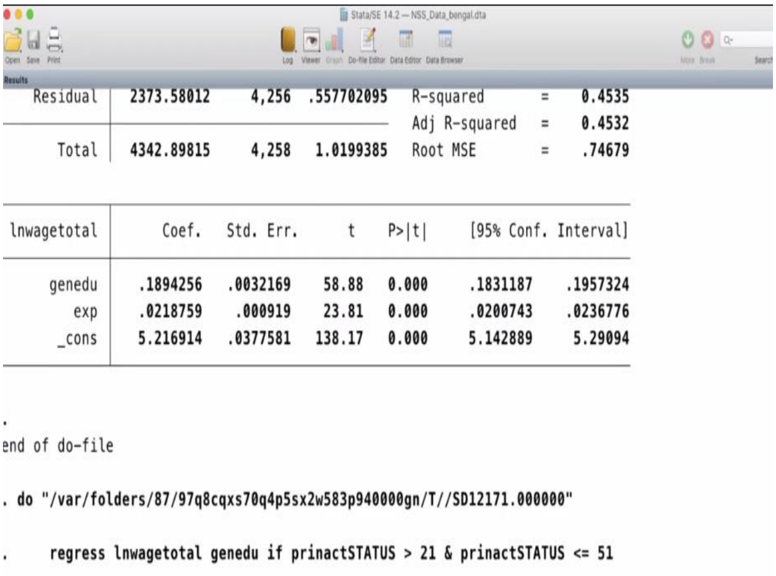
So let us again go back to the do-file and now I am going to run this. So this is my true regression. Let us this is the true model and we will see this results one by one. This is the regression where I have omitted the experience and this is the regression where I have omitted the general education and I will see the results.

(Refer Slide Time: 12:23)



And I got the results. Now in this lecture, I will just let you see this results and I want you to get an explanation of this results. What do you see here?

(Refer Slide Time: 13:00)



Let me explain you see that if I have general education so if I have the true model so in true model my general education is 0.189, experience is 0.02, adjusted R squared is 0.45.

(Refer Slide Time: 13:09)

```
    Residual |  2689.62176     4,257  .631811548   R-squared       =    0.3807
-------------+----------------------------------   Adj R-squared   =    0.3805
       Total |  4342.89815     4,258   1.0199385   Root MSE        =    .79487

------------------------------------------------------------------------------
 lnwagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   .1684559   .0032931    51.15   0.000     .1619997    .1749122
       _cons |   5.920188   .0250271   236.55   0.000     5.871122    5.969254
------------------------------------------------------------------------------

.
end of do-file

. do "/var/folders/87/97q8cqxs70q4p5sx2w583p940000gn/T//SD12171.000000"

.         regress lnwagetotal exp if prinactSTATUS > 21 & prinactSTATUS <= 51
```

Now, in the model with omitted variable when I only have general education, I have my R squared as 0.38, general education value is 0.16, so it has reduced. So when I remove the experience, the general education, the coefficient of general education reveals that is pretty interesting.

(Refer Slide Time: 13:32)

```
. do "/var/folders/87/97q8cqxs70q4p5sx2w583p940000gn/T//SD12171.000000"

.         regress lnwagetotal exp if prinactSTATUS > 21 & prinactSTATUS <= 51

      Source |       SS           df       MS      Number of obs   =     4,259
-------------+----------------------------------   F(1, 4257)      =     35.15
       Model |  35.5696043         1  35.5696043   Prob > F        =    0.0000
    Residual |  4307.32854     4,257  1.01182254   R-squared       =    0.0082
-------------+----------------------------------   Adj R-squared   =    0.0080
       Total |  4342.89815     4,258   1.0199385   Root MSE        =    1.0059

------------------------------------------------------------------------------
 lnwagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0070584   .0011905     5.93   0.000     .0047245    .0093924
       _cons |    6.85659   .0343481   199.62   0.000      6.78925     6.92393
------------------------------------------------------------------------------
```

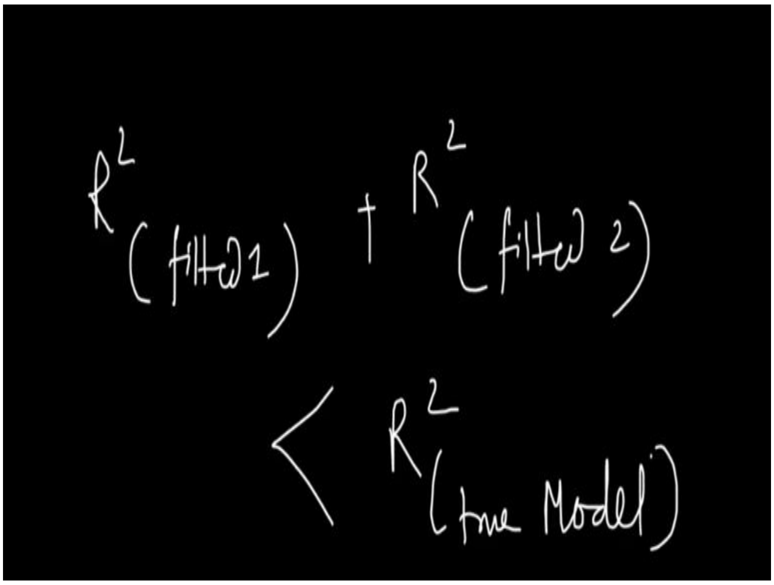
Now, in the next part if I can see that if I have included only experience R squared value quite low that is fine, but the coefficient of experience squared has also reduced, previously it was 0.02. So, what is happening here? Let us try to understand that. So, essentially in the last part when I have general education, my coefficient is actually decreasing and similarly R square values decreasing 0.38.

In the previous case I have both the case, both the variables genral education and experience and my R squared value is 0.45, it is higher than the last one with only general education and whereas if I have the independent variable only experience I see that this is again less than the coefficient of experience I had in my actual model. My actual model is is this one here. I have both general education as well as the experience, alright.

So this is something we have to interpret that why in this true model while both my coefficients are actually higher and in the fitted model where I have omitted variable bias problem, why both the coefficients for the independent variable is actually lower? So how exactly I can explain that? So this is something we have to see. So this is my problem number 1.

The second thing that I want to explain is that here I have the R squared value 0.3807 where I have taken only general education and here what I have taken if I have both of the explanatory variable I have 0.453 and the last part where I have taken only experience my R squared value is 0.008. So, if I add these two let us say, 0.38 + 008, it will be point 0.388 or something, let us say 0.39, whereas my R squared if I have both explanatory variables 0.45 so it is pretty high.

(Refer Slide Time: 16:08)



So essentially it means if I can write it down, essentially it means that here what is happening is that the R squared values for the two equations for the fitted; let us say if I can write it down R square 1 fitted and R square 2 fitted. So what I am going to have is; so if I can write it down R square 1 fitted or R square fitted 1 because that is how I have named my model R square fitted 1 + R squared fitted 2.

If I sum them up, this is going to be less than R square true model. How can I really explain that? So I want you to think through this problem and I will explain both the problems that I just talk about in a while.