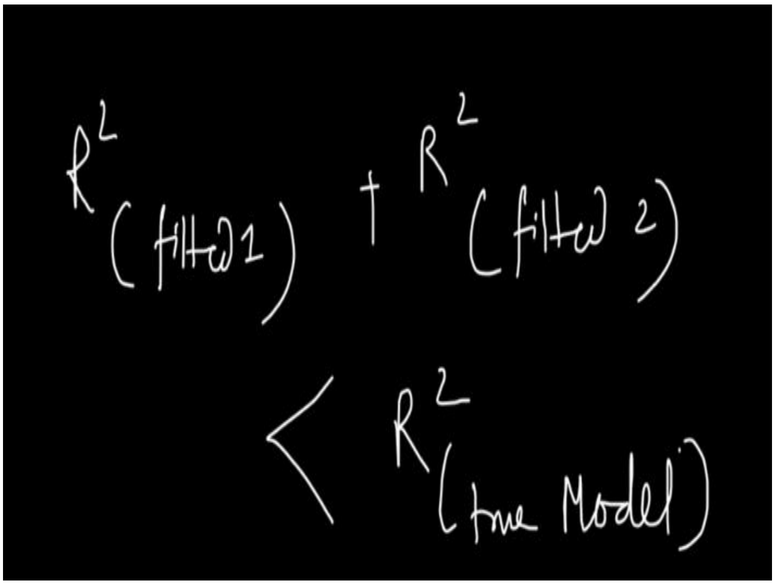
Applied Econometrics

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

Lecture – 88
Model Specification - Continued

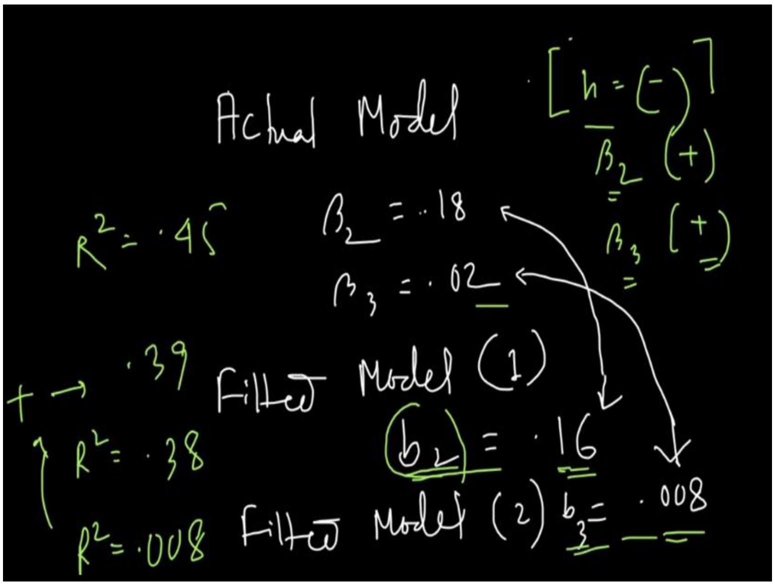
(Refer Slide Time: 00:26)



$$R^2_{(\text{fitted } 1)} + R^2_{(\text{fitted } 2)} < R^2_{(\text{true Model})}$$

So we spoke about two problems. One is the problem related with the coefficients and the other is the problem related to the R squared. Now we can make a sense that it is a problem of overestimation or underestimation and we will try to interpret from the numbers with the problems of overestimation or underestimation and we will try to make sense from these numbers.

(Refer Slide Time: 00:54)



So, let me first write down the coefficient value, let us say in my actual model, so I will write down the values actually. Actual model; so my beta 2 that is or b 2 rather, so actual model I can write beta 2 no problem. So beta 2 is equal to something like 0.18, beta 2 let me just go back there, beta 2 is; this is the real moral I have educational experience 0.18 and experience is 0.02, 0.18 and beta 3 = 0.02 and the fitted model 1 where I only had my b 2.

My b 2 is around 0.16 and fitted model 2 where I only had my b 3, b 3 = something like 0.008 let us say. Now, essentially of course we see that if I compare these, so b 2 is an underestimation of beta 2 and here also B 3 is an underestimation of beta 3, So this is b3. The way exactly this is happening. So, remember the relationship between education and experience that we have defined in our data.

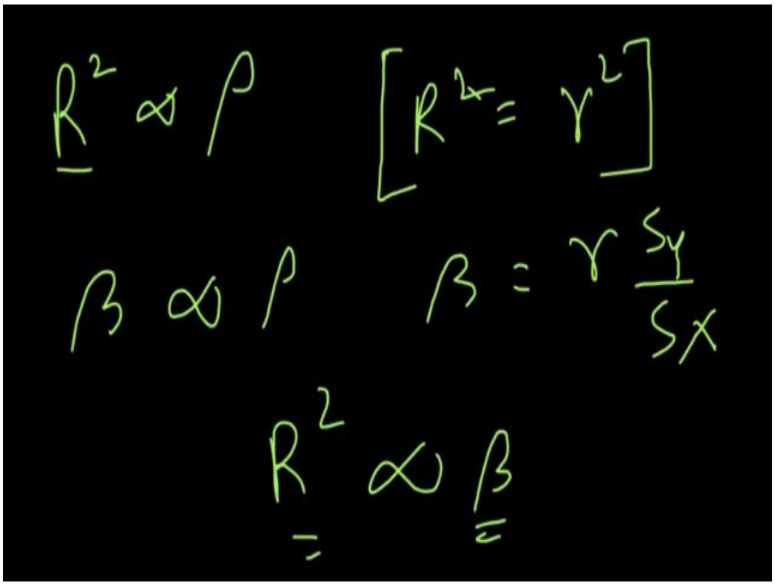
So, if we go back to that do-file you will see experience and education they are actually negatively related. So, education and experience they are negatively related; so with one increasing, the other is decreasing. So what is happening here is that we have this h which is basically the relationship between education and experience that is negative whereas the beta 2 that is the coefficient of education that has a positive contribution towards wage.

And beta 3 that is experience that is a positive contribution towards wage, but because h is negative, so if we console this table beta 3 and h so if one is negative and the other is positive or rather in our case this is the case. So if one is positive and the other is negative, so what we are going to have is a problem of underestimation of beta here. So exactly, that is what we see.

So how would you make sense is that when I am ler us say in this case when I am meeting experience, so that actually undermines the coefficient of education because they are negatively related. So it is cutting down. So when b 2 is actually capturing a part of education because they are negatively related a part of these it is actually cutting away and similarly for **or** experience now because experience and educational are negatively related.

So, the moment I remove education, so what is happening a part of it is actually cutting away, so from 0.02 it is is becoming 0.008. So, that is what we have to keep in mind. Now what about the R squared term? Now in this model R squared was 0.45, in this model R squared was 0.38, and the third model R squared was 0.008. So if I add these two things, I will get, if I sum it up I will get 0.39 which is less than 0.45. And how do I explain that?

(Refer Slide Time: 05:07)



$$\underline{R}^2 \propto \rho \qquad [R^2 = r^2]$$

$$\beta \propto \rho \qquad \beta = r\frac{S_Y}{S_X}$$

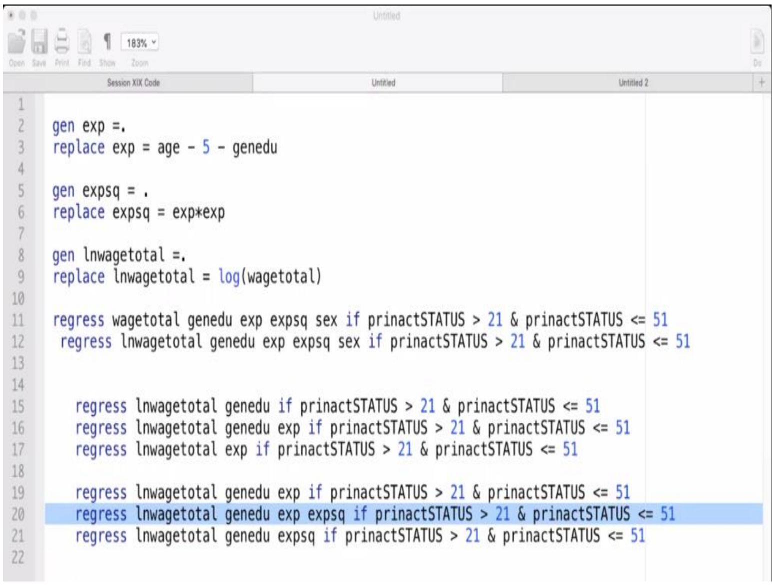
$$\underline{\underline{R}}^2 \propto \underline{\underline{\beta}}$$

How do I remember the fact that R square is proportional to the correlation coefficient between two variables. We have seen that previously. So R square is equal to r squared, we have actually derived that at some point. Now, R square is related to the correlation coefficient and again beta is related to the correlation coefficient because we know that beta is nothing but r S Y by S X, so that is something we get for simple OLS.

And because of these, we can say that R squared is actually proportional to beta. So beta is a coefficient. Now, because of this relationship if I have an increased value of beta it is also going to contribute to the R squared. So now that the the beta values in these two models are actually small, so what we are going to have a corresponding r squared value is also going to be relatively smaller.

But whereas when I have in the first model true model, I have my beta values higher and naturally corresponding R squared value is going to be higher, of course this is the explanation of both the questions. Now, you can actually do this work even further.

(Refer Slide Time: 06:37)

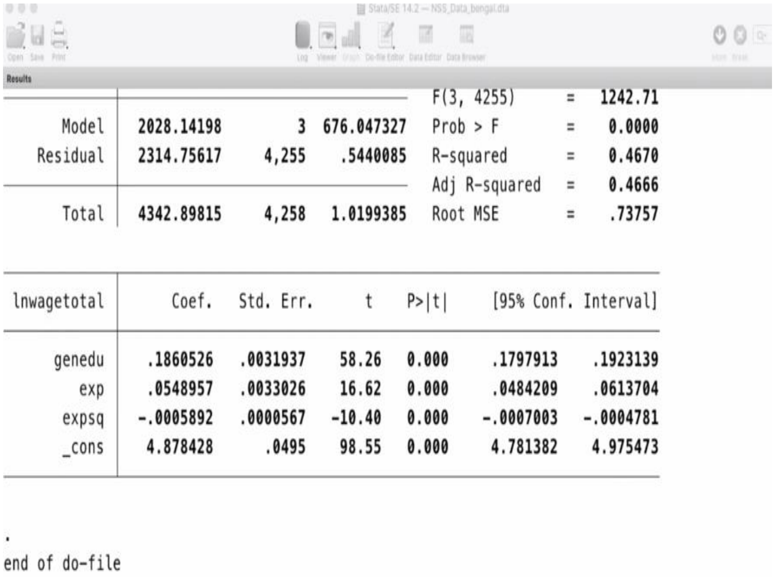


```

gen exp =.
replace exp = age - 5 - genedu

gen expsq = .
replace expsq = exp*exp

gen lnwagetotal =.
replace lnwagetotal = log(wagetotal)

regress wagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51
 regress lnwagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51


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

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

I mean of course, I have kind of claimed that my model is correct only when I include experience and education both, but it can also happen that I can say it is not enough, I have to have experience squared along with the experience. Now, if I want to include experience squared and that is actually what the Mincerian wage regression is. Now then I have to go to the next block of codes and I can actually see that in this case, I have these three regressions. And in this case, this middle regression is going to be the actual the true equation.

And the first and third are actually not going to be the right representation of the equation. So they are like omitted variables. So in the first case, I have omitted the experience squared and in the second case, I have omitted the experience. And we can see actually the consequences of such omission. And let me actually run the regression so that we can actually get a sense. So let me actually include the experience squared term here. And let us see what is happening.

(Refer Slide Time: 07:44)

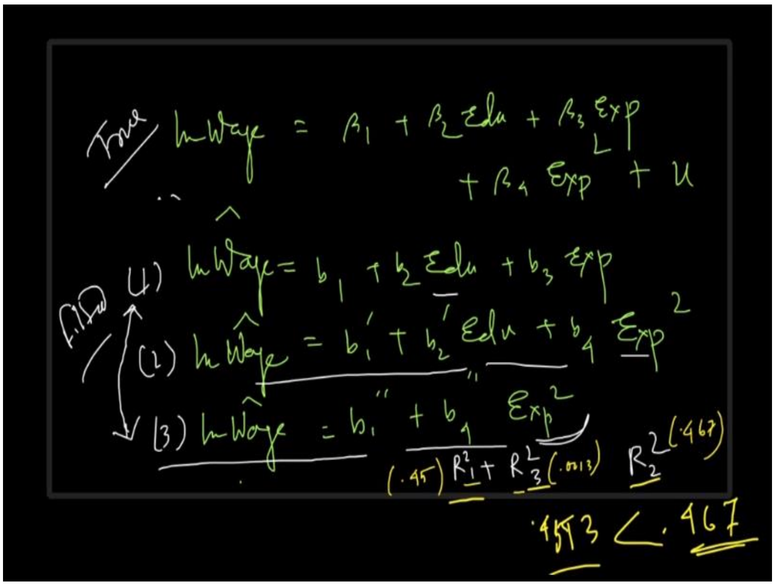

Stata/SE 14.2 — NSS_Data_bengal.dta
Open Save Print
Log Viewer Graph Do-file Editor Data Editor Data Browser
More Break
Results

```
                                                   F(3, 4255)      =   1242.71
       Model    2028.14198         3  676.047327   Prob > F        =    0.0000
    Residual    2314.75617     4,255    .5440085   R-squared       =    0.4670
                                                   Adj R-squared   =    0.4666
       Total    4342.89815     4,258   1.0199385   Root MSE        =    .73757

 lnwagetotal       Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]

      genedu    .1860526   .0031937    58.26    0.000     .1797913    .1923139
         exp    .0548957   .0033026    16.62    0.000     .0484209    .0613704
       expsq   -.0005892   .0000567   -10.40    0.000    -.0007003   -.0004781
       _cons    4.878428      .0495    98.55    0.000     4.781382    4.975473

.
end of do-file
```

So if I run this, I get this regression equation where my adjusted R squared value is pretty high 0.467. And I have all these coefficients, experience squared as I said that it is essentially the representing the tapering off of the wage with experienced and that happens because of the ages with as we increase, as we get older, our experience starts counting less so that is why this negative coefficients.

Normally what happens is experience squared is going to contribute wage in negatively. So essentially, the beta 3 is going to be negative or here in this case beta 4. So let me actually write down the equations.

(Refer Slide Time: 08:42)



True $\ln Wage = \beta_1 + \beta_2 Edu + \beta_3 Exp + \beta_4 Exp^2 + u$

Fitted

(1) $\widehat{\ln Wage} = b_1 + b_2 \underline{Edu} + b_3 Exp + b_4 \underline{Exp}^2$

(2) $\underline{\widehat{\ln Wage} = b_1' + b_2' Edu}$

(3) $\underline{\widehat{\ln Wage}} = b_1'' + \underline{b_4''} \; Exp^2$

$(.45)\ \underline{R_1^2} + \underline{R_3^2}(.0013)$ $\quad \underline{R_2^2}(.467)$

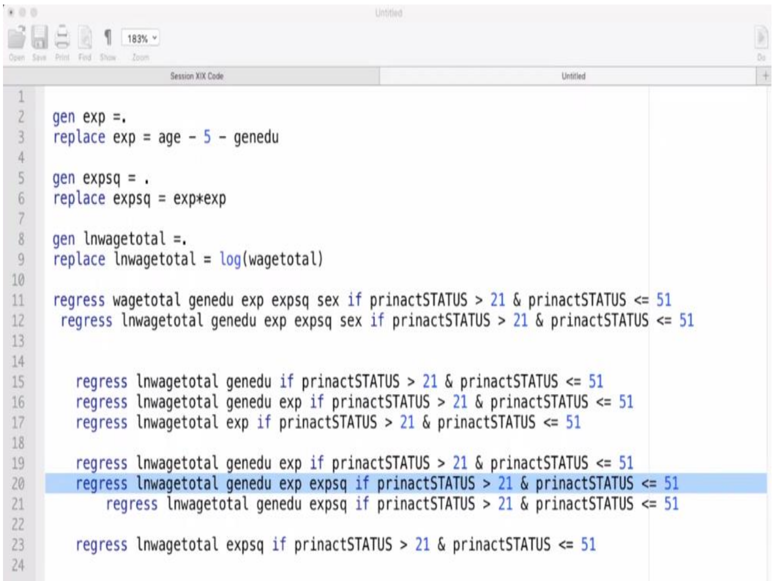
$\underline{.4513} < \underline{.467}$

In this case, the true equation is going to be wage or log of wage let us say is equal to let us say beta 1 + beta 2 into edu beta 3 into experience beta 4 into experience squared and then I have some error term what is is my fitted model ln wage. I will consider actually two fitted models, so let me write it down. So in this case, in the first case, I will actually omit the experience squared. So ln wage is going to be let us say b 1 + b 2 edu and b 3 experience.

I do not have experience squared and I have a fitted model. And in the next fitted model I will have let us say I will omit the experience rather I will have experience squared. So ln wage is going to be let us say b 1 prime + b 2 prime edu, I am keeping edu in both equations. And then I have omitted the experience part, so there is no b 3. And then I have b 4 experience squared. So this is how I have written my equations.

I will also add another equation which will show at the end of it what happens if I have let us say b 1 double prime b 2 double prime actually we will not have any edu and let us say we will only have b 4 double prime experience squared. So I just come come to that to explain the R squared part. Whereas these fit in models let us see what is happening to the coefficients? Fitted 1, fitted 2 and fitted 3. Now, let us look at these equations in my do-file. So let me go to the do-file.

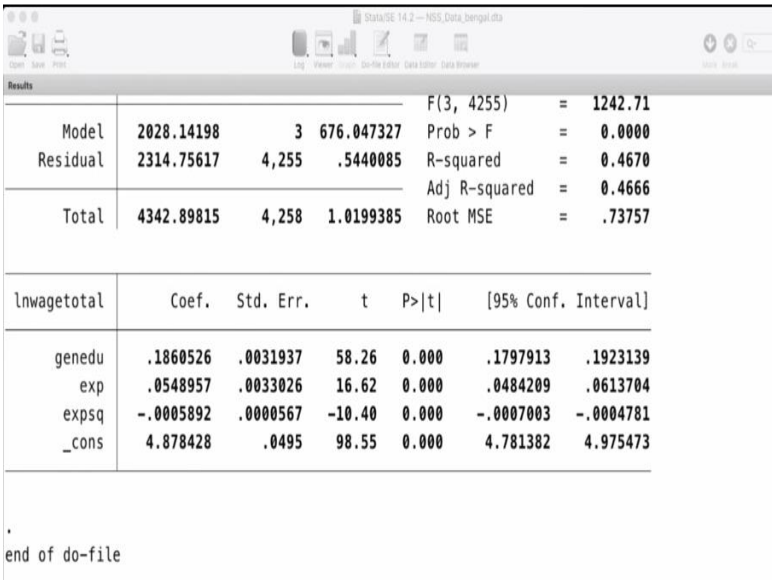
(Refer Slide Time: 10:41)



```

gen exp =.
replace exp = age - 5 - genedu

gen expsq = .
replace expsq = exp*exp

gen lnwagetotal =.
replace lnwagetotal = log(wagetotal)

regress wagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51
 regress lnwagetotal genedu exp expsq sex if prinactSTATUS > 21 & prinactSTATUS <= 51


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

   regress lnwagetotal genedu exp if prinactSTATUS > 21 & prinactSTATUS <= 51
   regress lnwagetotal genedu exp expsq if prinactSTATUS > 21 & prinactSTATUS <= 51
       regress lnwagetotal genedu expsq if prinactSTATUS > 21 & prinactSTATUS <= 51

   regress lnwagetotal expsq if prinactSTATUS > 21 & prinactSTATUS <= 51

```

And in the do-file I can see that in the last block of equations, I have general education experience, general education experience experience squared, and I have only taken experience squared. I just add one more here that is basically I will keep general education and experience square, gen edu and experience squared. So let me run the regression. So this is the main equation, this is the actual equation, true equation.

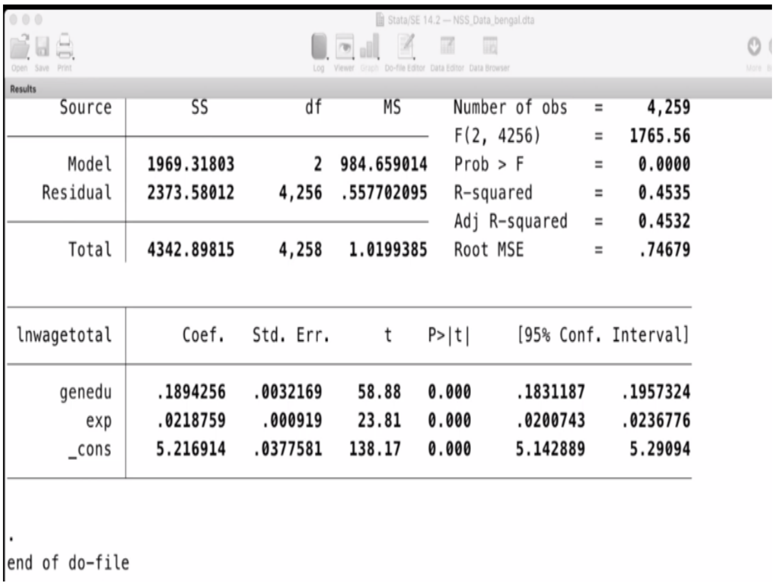
(Refer Slide Time: 11:21)



```
             |                                    F(3, 4255)      =   1242.71
       Model |  2028.14198         3  676.047327  Prob > F        =    0.0000
    Residual |  2314.75617     4,255    .5440085  R-squared       =    0.4670
-------------+----------------------------------  Adj R-squared   =    0.4666
       Total |  4342.89815     4,258   1.0199385  Root MSE        =    .73757

------------------------------------------------------------------------------
 lnwagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   .1860526   .0031937    58.26   0.000     .1797913    .1923139
         exp |   .0548957   .0033026    16.62   0.000     .0484209    .0613704
       expsq |  -.0005892   .0000567   -10.40   0.000    -.0007003   -.0004781
       _cons |   4.878428      .0495    98.55   0.000     4.781382    4.975473
------------------------------------------------------------------------------

.
end of do-file
```

And if I run it, what I will see here is this equation. So I see that experience squared is negative, experience is positive, general education is positive just the way we have expected. Now, interestingly enough, previously my experience coefficient was 0.02, but here I see 0.05, so that is interesting. And le tus say I run the other regression equations. So let us say I take fitted one which is general education experience. And once I run it let us see what is going to the outcome.

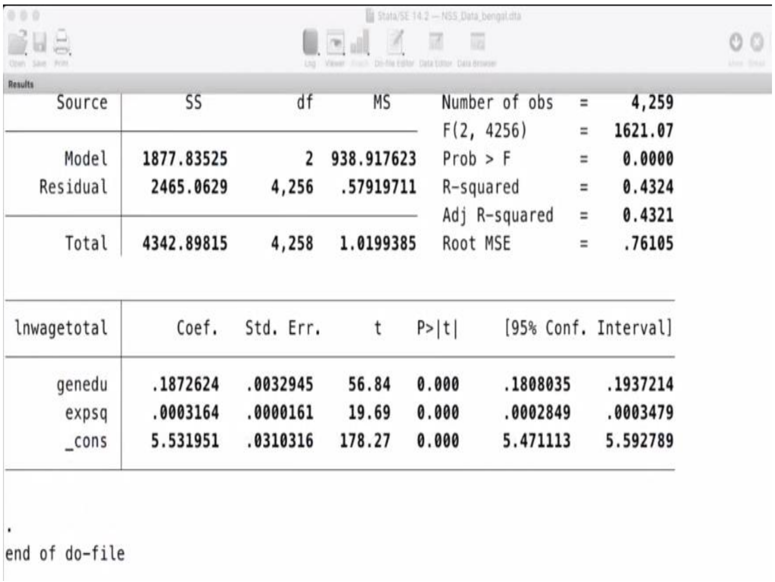
(Refer Slide Time: 11:53)



```
      Source |       SS           df       MS      Number of obs   =     4,259
-------------+----------------------------------   F(2, 4256)      =   1765.56
       Model |  1969.31803         2  984.659014   Prob > F        =    0.0000
    Residual |  2373.58012     4,256  .557702095   R-squared       =    0.4535
-------------+----------------------------------   Adj R-squared   =    0.4532
       Total |  4342.89815     4,258   1.0199385   Root MSE        =    .74679

------------------------------------------------------------------------------
 lnwagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   .1894256   .0032169    58.88   0.000     .1831187    .1957324
         exp |   .0218759    .000919    23.81   0.000     .0200743    .0236776
       _cons |   5.216914   .0377581   138.17   0.000     5.142889     5.29094
------------------------------------------------------------------------------

.
end of do-file
```

So, result is you see, we have we have done it previously. So if I have only experience, I have 0.02, whereas in the previous case I have seen 0.05. So basically, when I include the experience square, the coefficient for my experience is actually increasing just like the previous case.

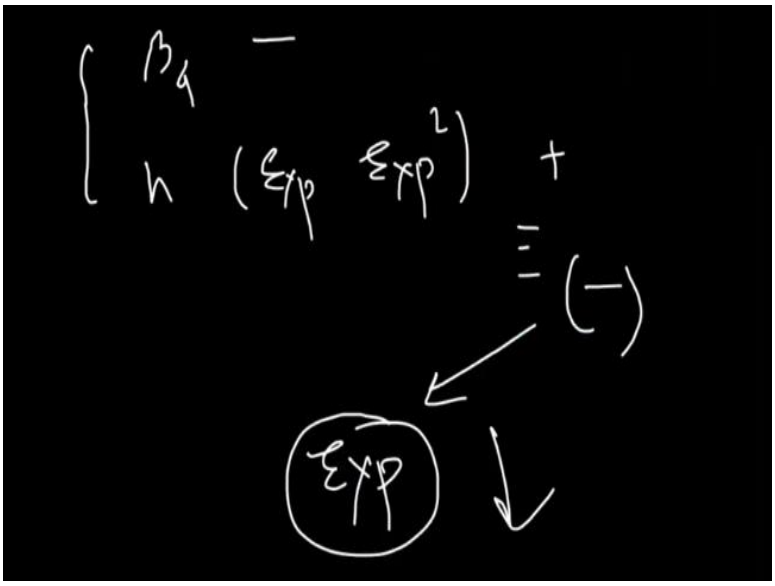
(Refer Slide Time: 12:33)



```
      Source |       SS           df       MS      Number of obs   =     4,259
-------------+----------------------------------   F(2, 4256)      =   1621.07
       Model |  1877.83525         2  938.917623   Prob > F        =    0.0000
    Residual |  2465.0629      4,256  .57919711    R-squared       =    0.4324
-------------+----------------------------------   Adj R-squared   =    0.4321
       Total |  4342.89815     4,258  1.0199385    Root MSE        =    .76105

------------------------------------------------------------------------------
 lnwagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   .1872624   .0032945    56.84   0.000     .1808035    .1937214
       expsq |   .0003164   .0000161    19.69   0.000     .0002849    .0003479
       _cons |   5.531951   .0310316   178.27   0.000     5.471113    5.592789
------------------------------------------------------------------------------

.
end of do-file
```

And if I run the regression where I have only let us general education and experience squared, what what we will see here is that my experience square has become positive instead of having a negative value. And how do I explain in both the cases? So let me explain the first case first. So let us see what happened here? So what happened in the first case is that the moment I included experienced squared, my experience has increased the coefficient for my experience has increased and why is that? Because experience squared is negative, so beta 4 is negative in this case. So let me actually write it down.

(Refer Slide Time: 13:14)



So let me actually go to the whiteboard, beta 4 in this case is negative. Whereas h, h for experience and experience squared is of course positive, it is positive. So the moment I am excluding beta 4, so what I am having is that beta 4 and h that together they are giving me a negative sign, the resultant of this is going to give me a negative sign and that is being captured in the experience.

And what is happening it is undermining the coefficient of experience exactly what you have seen previously. So, this is a case where we can say we are having something like this. So, here my h is positive in this case and my beta 3 is negative in this case. So, essentially it is an underestimation problem and that is exactly what you see in the result and what about the last part? So, if I go to the last part, where I had in my regression I have omitted the experience part, rather I have kept the experience squared part.

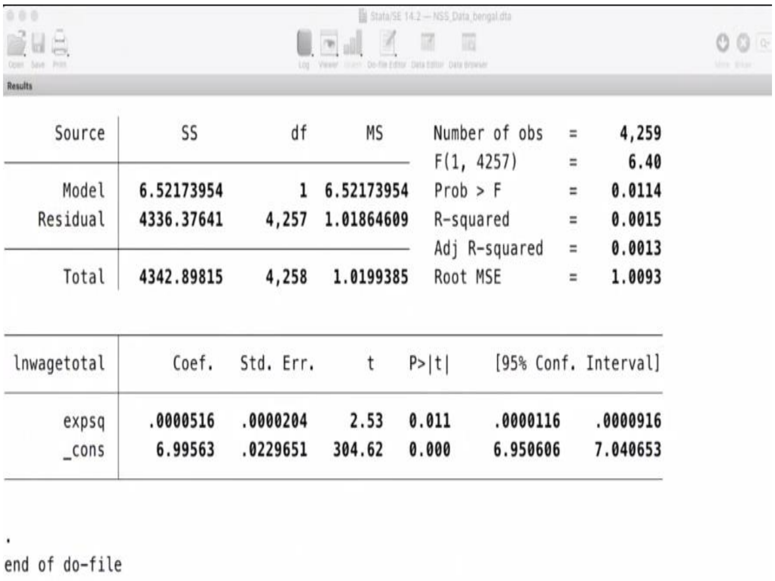
What is happening here is again the experience is being captured by experience square, now experience is related to Y positively whereas experience square is related to Y negatively, but the net effect because the experience has a much higher impact on Y vis-a-vis experience square that only has a role in tapering the whole income experience curve. So, here we will see a net positive and that is again basically an underestimation of the experience square because that is being impacted by experience here.

So that way we should understand how the variables are related within themselves and how they are related with the Y, so that is how we should understand. Now, let us look at the R squared part. The last part here is R squared part. And in the first model we had general education experience. And in the second model we had general education experience squared. So we cannot really add these things up this time because in both cases we have general education, so it is being captured.

But what you can do instead, in one of our models we can actually omit the general education altogether and we can capture it only once and that is in this equation, equation 3 we have basically omitted general education and we have captured only experienced square. So if I add this 3 with 1 and see what happens to the R squared and compare that with 2. So basically, what I want to do is R square in 1 + R square in 3.

And I will see how it is related to the R square in 2 because R square in 2 that is going to be true R squared here. Now, if I look at the R squared value and if I just go back to my data results and if I look at the R squared value, what I will see is, I will see, okay I have to run this at once. Let me run this. Let me go to do-file and let me run this. The last part of the do-file, so that is this one only, this is my equation 3.

(Refer Slide Time: 17:07)



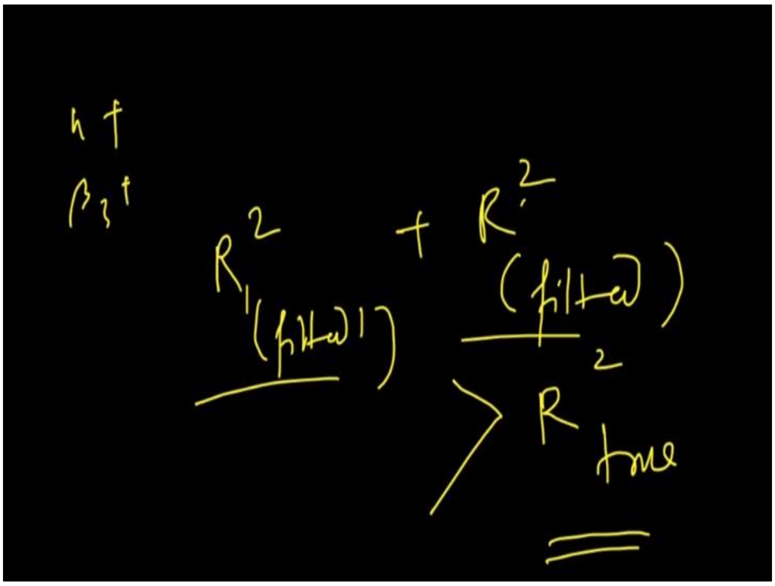
```
      Source |       SS           df       MS      Number of obs   =     4,259
-------------+----------------------------------   F(1, 4257)      =      6.40
       Model |  6.52173954         1  6.52173954   Prob > F        =    0.0114
    Residual |  4336.37641     4,257  1.01864609   R-squared       =    0.0015
-------------+----------------------------------   Adj R-squared   =    0.0013
       Total |  4342.89815     4,258   1.0199385   Root MSE        =    1.0093

------------------------------------------------------------------------------
 lnwagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       expsq |   .0000516   .0000204     2.53   0.011     .0000116    .0000916
       _cons |    6.99563   .0229651   304.62   0.000     6.950606    7.040653
------------------------------------------------------------------------------

.
end of do-file
```

So here I have my R square value is 0.0013. And in the previous case where I had general education and experience, I had 0.45. So if I add these two things up, I will have; back to the whiteboard, so R squared or rather R squared 1 is going to be or R squared 3 is going to be 0.0013 and R squared 1 is going to be 0.45 and my R squared 2 where I have included all these explanatory variables, actually go to the file here this data file, and here I will see I have all the variables that is 0.467.

So let me write it down here 0.467. Now if I add R square 1 and R square 3, I will see three 0.45 + 0.0013 is going to give me something like 0.45; let me actually reduce the size a little bit, it is going to give me 0.4513 which is actually less than 0.467. So, exactly what you have seen previously in the case of underestimation since the beta, actually the R square is actually related to the value of the beta.

So what is going to happen is we are going to see an underestimation of R squared. So even if we add these two R squared values, they are not going to give me or they are actually going to be less than the true model, the R squared corresponding to the true model.

(Refer Slide Time: 19:19)



So this way, I will ask you to also see those cases where my h is positive where my beta 3 is positive and then if I actually run the regressions, I will see R square 1 or for fitted and R square for fitted 2. If I add these things up, we will actually see that is going to be R square true model because here we will see overestimation a beta and therefore correspondingly an inflation of R squared value.

And that is if we add these up, or let me actually show you, they are actually going to be bigger than the R squared value corresponding to the true model. So with this, we end this lecture here where we have explained the impact of omitted variable bias on R square. Thank you.