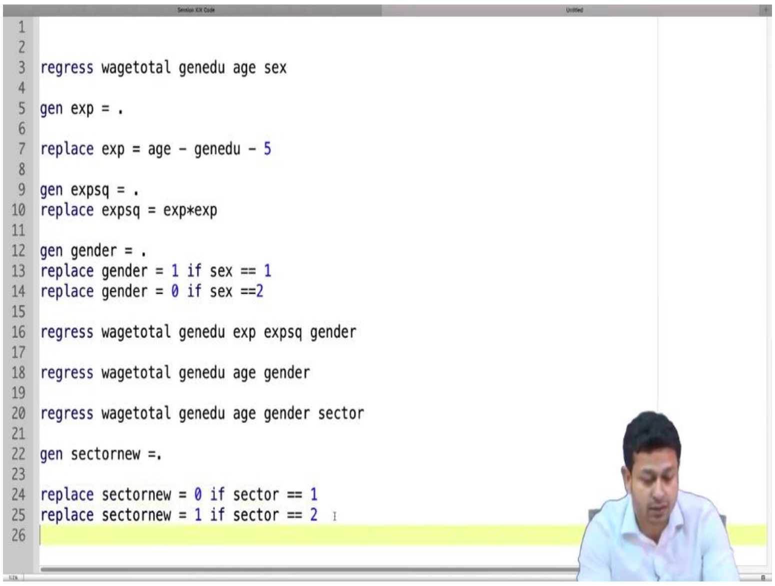
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 7
Lecture - 58
Dummy Variable (Contd.)

Hello and welcome back to the lecture on Applied Econometrics, and we are talking about dummy variable. And we have been doing an example where we use national sample survey data for West Bengal, and we want to understand the impact of different variables including different types of dummy variables on my dependent variable which is wage.

(Refer Slide Time: 00:44)



```
regress wagetotal genedu age sex

gen exp = .

replace exp = age - genedu - 5

gen expsq = .
replace expsq = exp*exp

gen gender = .
replace gender = 1 if sex == 1
replace gender = 0 if sex ==2

regress wagetotal genedu exp expsq gender

regress wagetotal genedu age gender

regress wagetotal genedu age gender sector

gen sectornew =.

replace sectornew = 0 if sector == 1
replace sectornew = 1 if sector == 2
```

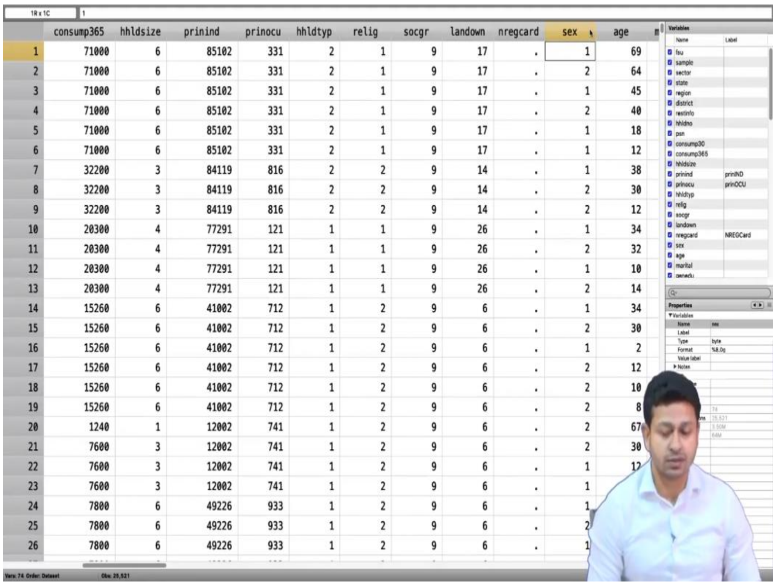
And as I said previously, we used; I will just refine the variables a little bit. So, normally, we can use age as a proxy to experience, but we can also create a variable called experience and we can use that variable experience in my regression equation. So, let us say I have created a variable called experience. And how we code it? So, I will replace experience as; so, I have already run this gen exp equal to dot.

So, basically, that is how you create a new variable in your Stata. And this is going to be your, let us say, a person; I will consider a person age and let us say general education. I will just explain this; minus 5 is the experience. And how would we do that? So, we assume that someone joins school at the age of 5. And then, that person has some years of general education.

And then, if you subtract that from age, so, that is basically the experience of that person. So, we assume that after the completion of the education, that person has joined the labour market; and that becomes their experience. So, that is one. And we are actually going to run a very standard wage regression where we use Mincerian wage regression. And we actually also use experience square with experience.

And that is because of how the experience variable behaves, which we have seen from lot of empirical data that over a period of time, it actually plateaus. And that is why we need to use another explanatory variable that is experience square. And let us say that is, I replace it with the value of expsq is equal to exp into exp. So, basically, that is experience square. So, this is one variable I am going to include in my regression equation. And now, we have seen something interesting that we have to explain here. We are including 2 dummy variable here.

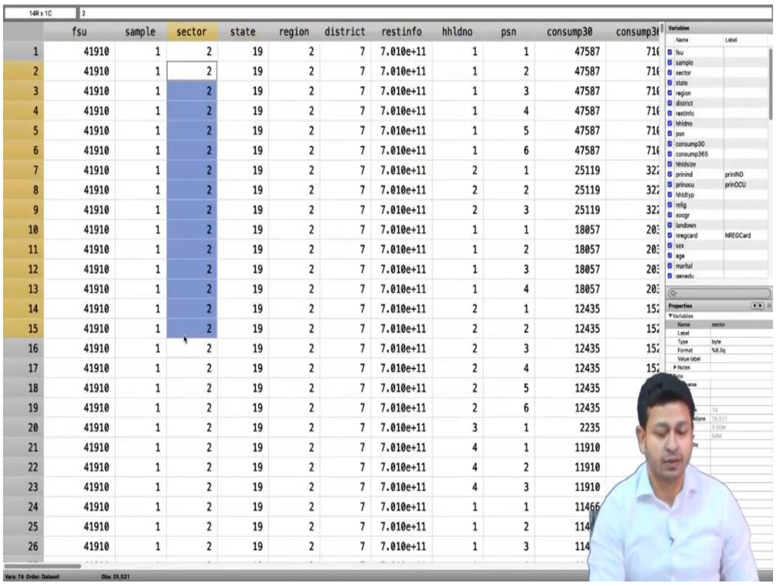
(Refer Slide Time: 02:45)



	consump365	hhldsize	prinind	prinocu	hhldtyp	relig	socgr	landown	nregcard	sex	age
1	71000	6	85102	331	2	1	9	17	.	1	69
2	71000	6	85102	331	2	1	9	17	.	2	64
3	71000	6	85102	331	2	1	9	17	.	1	45
4	71000	6	85102	331	2	1	9	17	.	2	40
5	71000	6	85102	331	2	1	9	17	.	1	18
6	71000	6	85102	331	2	1	9	17	.	1	12
7	32200	3	84119	816	2	2	9	14	.	1	38
8	32200	3	84119	816	2	2	9	14	.	2	30
9	32200	3	84119	816	2	2	9	14	.	2	12
10	20300	4	77291	121	1	1	9	26	.	1	34
11	20300	4	77291	121	1	1	9	26	.	2	32
12	20300	4	77291	121	1	1	9	26	.	1	10
13	20300	4	77291	121	1	1	9	26	.	2	14
14	15260	6	41002	712	1	2	9	6	.	1	34
15	15260	6	41002	712	1	2	9	6	.	2	30
16	15260	6	41002	712	1	2	9	6	.	1	2
17	15260	6	41002	712	1	2	9	6	.	2	12
18	15260	6	41002	712	1	2	9	6	.	2	10
19	15260	6	41002	712	1	2	9	6	.	2	8
20	1240	1	12002	741	1	2	9	6	.	2	67
21	7600	3	12002	741	1	2	9	6	.	2	30
22	7600	3	12002	741	1	2	9	6	.	1	12
23	7600	3	12002	741	1	2	9	6	.	1	[illegible]
24	7800	6	49226	933	1	2	9	6	.	1	[illegible]
25	7800	6	49226	933	1	2	9	6	.	2	[illegible]
26	7800	6	49226	933	1	2	9	6	.	1	[illegible]

In the first case, I am going to include just one dummy variable, and that is, let us say, the gender or sex. Now see, the way people in the field who collected data or people who actually processed the data, that the way we got this raw data, we had the dummy; the values were 1 and 2. And as an econometrician, as a person who is running the regression, we have to define how we are going to use the different values of this dummy variable. Similarly, for sector which is rural and urban, we will have 1 and 2; we have seen that.

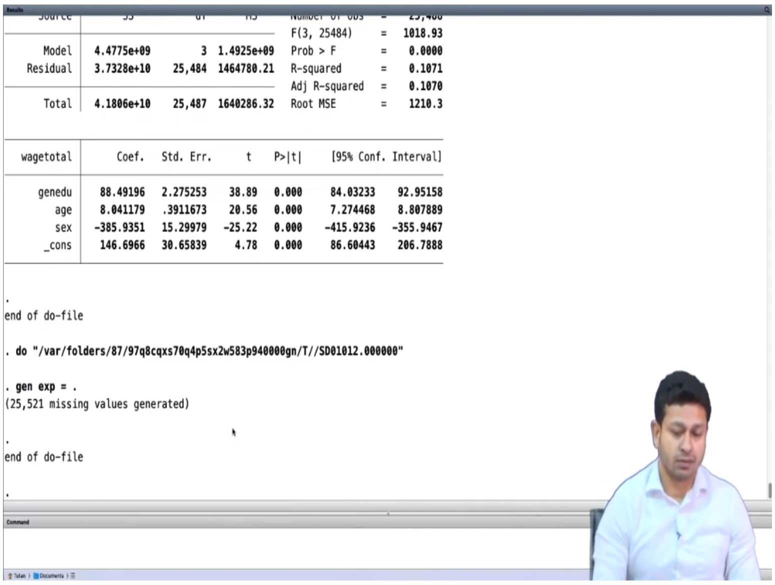
(Refer Slide Time: 03:18)



	fsu	sample	sector	state	region	district	restinfo	hhldno	psn	consump30	consump3
1	41910	1	2	19	2	7	7.010e+11	1	1	47587	716
2	41910	1	2	19	2	7	7.010e+11	1	2	47587	716
3	41910	1	2	19	2	7	7.010e+11	1	3	47587	716
4	41910	1	2	19	2	7	7.010e+11	1	4	47587	716
5	41910	1	2	19	2	7	7.010e+11	1	5	47587	716
6	41910	1	2	19	2	7	7.010e+11	1	6	47587	716
7	41910	1	2	19	2	7	7.010e+11	2	1	25119	322
8	41910	1	2	19	2	7	7.010e+11	2	2	25119	322
9	41910	1	2	19	2	7	7.010e+11	2	3	25119	322
10	41910	1	2	19	2	7	7.010e+11	1	1	18057	203
11	41910	1	2	19	2	7	7.010e+11	1	2	18057	203
12	41910	1	2	19	2	7	7.010e+11	1	3	18057	203
13	41910	1	2	19	2	7	7.010e+11	1	4	18057	203
14	41910	1	2	19	2	7	7.010e+11	2	1	12435	152
15	41910	1	2	19	2	7	7.010e+11	2	2	12435	152
16	41910	1	2	19	2	7	7.010e+11	2	3	12435	152
17	41910	1	2	19	2	7	7.010e+11	2	4	12435	152
18	41910	1	2	19	2	7	7.010e+11	2	5	12435	
19	41910	1	2	19	2	7	7.010e+11	2	6	12435	
20	41910	1	2	19	2	7	7.010e+11	3	1	2235	
21	41910	1	2	19	2	7	7.010e+11	4	1	11910	
22	41910	1	2	19	2	7	7.010e+11	4	2	11910	
23	41910	1	2	19	2	7	7.010e+11	4	3	11910	
24	41910	1	2	19	2	7	7.010e+11	1	1	11466	
25	41910	1	2	19	2	7	7.010e+11	1	2	114	
26	41910	1	2	19	2	7	7.010e+11	1	3	114	

Now, how do I really want to use it? So, as a convention, what I do is, I actually convert these values into 0 and 1.

(Refer Slide Time: 03:30)

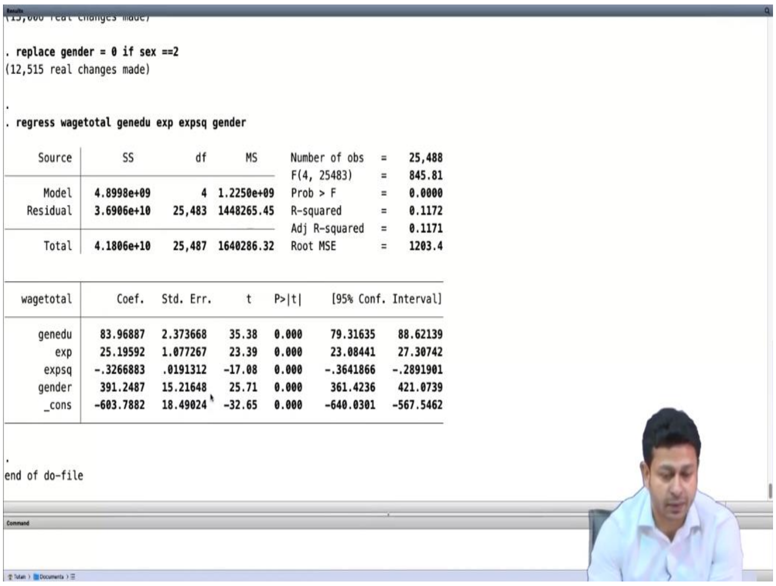


```
      Source |       SS           df       MS      Number of obs   =    25,488
-------------+----------------------------------   F(3, 25484)     =   1018.93
       Model |  4.4775e+09         3  1.4925e+09   Prob > F        =    0.0000
    Residual |  3.7328e+10    25,484  1464780.21   R-squared       =    0.1071
-------------+----------------------------------   Adj R-squared   =    0.1070
       Total |  4.1806e+10    25,487  1640286.32   Root MSE        =    1210.3

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   88.49196   2.275253    38.89   0.000     84.03233    92.95158
         age |   8.041179   .3911673    20.56   0.000     7.274468    8.807889
         sex |  -385.9351   15.29979   -25.22   0.000    -415.9236   -355.9467
       _cons |   146.6966   30.65839     4.78   0.000     86.60443    206.7888
------------------------------------------------------------------------------

.
end of do-file

. do "/var/folders/87/97q8cqxs70q4p5sx2w583p940000gn/T//SD01012.000000"

. gen exp = .
(25,521 missing values generated)

.
end of do-file

.
```

And I will tell you the different implications of these different values that we use. So, I have this variable sex already. Let us say I create a new variable called gender, and gender is equal to dot. And I replace the gender is equal to, let us say, 1. So, this is for, let us say, for females or let us say it is for males. If let us say my sex, the other variable was 1. So, sex had 2 values. And I am just going to explain what I am doing here.

Replace gender is equal to 0 if sex is equal to 2. So, it means that, what I have simply done is, someone having a sex is equal to 2, means female; I have assigned a value is equal to 0 and I have called it a new variable gender. So, basically, what all I have done, I have incorporated a new variable gender and I have changed the values from 1 and 2 to 0 and 1. Now, if I run a regression, let us say, instead of; so, I have created this variable called experience, experience squared and gender.

(Refer Slide Time: 04:58)



```
(15,000 real changes made)

. replace gender = 0 if sex ==2
(12,515 real changes made)

.
. regress wagetotal genedu exp expsq gender

      Source |       SS           df       MS      Number of obs   =    25,488
-------------+----------------------------------   F(4, 25483)     =    845.81
       Model |  4.8998e+09         4  1.2250e+09   Prob > F        =    0.0000
    Residual |  3.6906e+10    25,483  1448265.45   R-squared       =    0.1172
-------------+----------------------------------   Adj R-squared   =    0.1171
       Total |  4.1806e+10    25,487  1640286.32   Root MSE        =    1203.4

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   83.96887   2.373668    35.38   0.000     79.31635    88.62139
         exp |   25.19592   1.077267    23.39   0.000     23.08441    27.30742
       expsq |  -.3266883   .0191312   -17.08   0.000    -.3641866   -.2891901
      gender |   391.2487   15.21648    25.71   0.000     361.4236    421.0739
       _cons |  -603.7882   18.49024   -32.65   0.000    -640.0301   -567.5462
------------------------------------------------------------------------------

.
end of do-file
```

We will see that, of course, all these different variables are significant, because I know from my experience, it is a standard regression equation. So, I have included these variables because these variables are going to be important. And this gender, of course it has taken a value, and it has shown something positive here. And what do we actually interpret? So, previously, it showed something negative in the previous example, -385.

And in this example, it is showing something positive which is +391. So, how do we explain that? So, that is our first task here. Now, look at the gender dummy. Think about what we really assign the values as. So, we assign the value of gender is equal to 1 in case it is male; and value of gender is equal to 0 in case it is female. Now, I said that when we interpret dummy variable and the way the Stata or other; when I do a regression equation and we use dummy variables, so, one category always becomes the reference category.

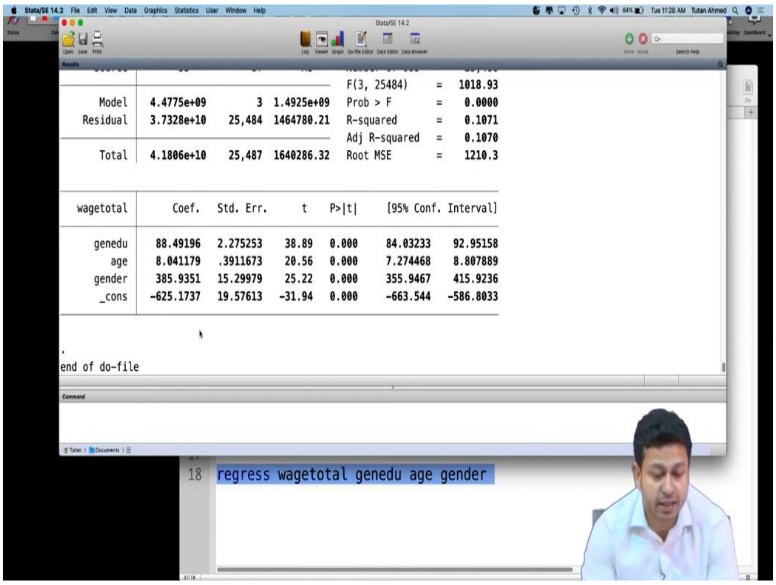
So, here, the reference category is equal to 0. And my reference category means here the females. And the way I have assigned value and I have given a value of male higher than female, so, it is how the basically convention is. So, basically, what Stata is doing, it is trying to see the impact of this 1 unit change from 0 to 1. So, the moment you move from 0 to 1, how much your wage is changing. So, it is 391 rupees a weak for males.

So, now, I will interpret it as; because I know that my female is a reference category, so here, a male actually earns 391 rupees more than a female in a given week, let us say. And you see the values are little different. It is 391; previously it was 385. It is simply because I have changed the other explanatory variable; because, instead of age, here I have used experience and experience squared.

So, if I actually use the same explanatory variables, we will see; let us actually see if I have the same explanatory variable, just age. So, it is just like this equation. All I have done, I have changed sex to gender, which is nothing but converting the values 1, 2 into 0, 1. And previously, in this case, my reference dummy was male, because male was valued as 1, female was valued as 2, and by convention, Stata will pick the lower value as the reference category which is the male.

So, we saw the impact of being female with respect to male. And here, what you are seeing is impact of being male with respect to female. So, let us see what happens if I run the regression. I have run it. So, now you see what we got.

(Refer Slide Time: 08:00)



```
                                                  F(3, 25484)     =   1018.93
       Model |  4.4775e+09         3  1.4925e+09  Prob > F        =    0.0000
    Residual |  3.7328e+10    25,484  1464780.21  R-squared       =    0.1071
-------------+----------------------------------  Adj R-squared   =    0.1070
       Total |  4.1806e+10    25,487  1640286.32  Root MSE        =    1210.3

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   88.49196   2.275253    38.89   0.000     84.03233    92.95158
         age |   8.041179   .3911673    20.56   0.000     7.274468    8.807889
      gender |   385.9351   15.29979    25.22   0.000     355.9467    415.9236
       _cons |  -625.1737   19.57613   -31.94   0.000     -663.544   -586.8033
------------------------------------------------------------------------------

.
end of do-file
```

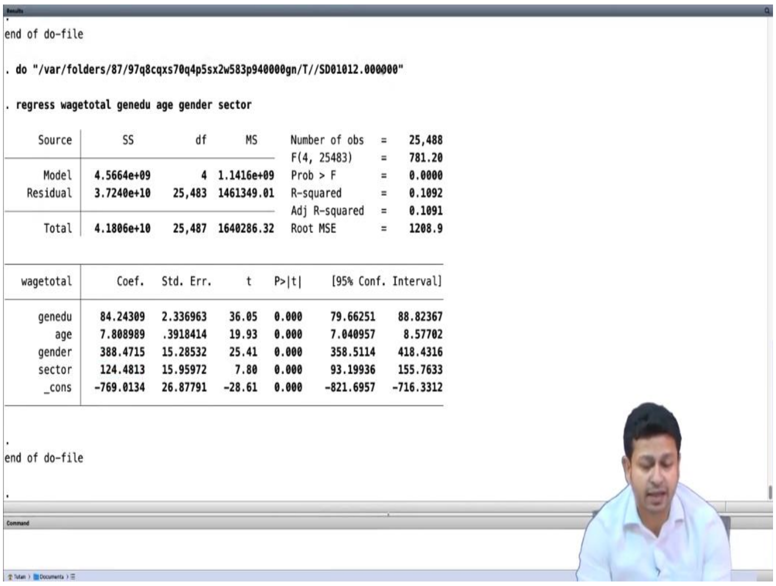
```
regress wagetotal genedu age gender
```

What we got is the same 385.9, but with a positive sign. So, that means, the impact is same; impact is 385.9 rupees. Here also we had 385.9 rupees, but it has a minus sign, because there I was comparing females wage with respect to males. And here, I am comparing males wage with respect to female. So, the quantity, the magnitude remains exactly the same. It only changes the sign. And the impact of other variables are going to be same.

So, it does not matter which one you choose the reference category. So, I could have very well scored in my; the way I have created the variables, I could have very well scored female as 1 and male as 0, instead of male as 1 and female as 0. So, all that would have happened, I would have gotten a minus sign here. So, it does not matter how we assign the values of the dummy variable, it will give you the result, the interpretation of the result remains the same.

You will just see the change in sign. So, now, let us see, we include another dummy variable in the same category, binary category, and we have seen that is we want to include sector. And here also, we see the values of sector is 1 and 2. And let us first simply run a regression equation where we will use sector. So, it is a dummy variable. It is like Stata will analyse the dummy variable assuming that it has two values, 1 and 2. And let us see what happens here. I have run it and what I see?

(Refer Slide Time: 09:58)



```
.
end of do-file

. do "/var/folders/87/97q8cqxs70q4p5sx2w583p940000gn/T//SD01012.000000"

. regress wagetotal genedu age gender sector

      Source |       SS           df       MS      Number of obs   =    25,488
-------------+----------------------------------   F(4, 25483)     =    781.20
       Model |  4.5664e+09         4  1.1416e+09   Prob > F        =    0.0000
    Residual |  3.7240e+10    25,483  1461349.01   R-squared       =    0.1092
-------------+----------------------------------   Adj R-squared   =    0.1091
       Total |  4.1806e+10    25,487  1640286.32   Root MSE        =    1208.9

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   84.24309   2.336963    36.05   0.000     79.66251    88.82367
         age |   7.808989   .3918414    19.93   0.000     7.040957     8.57702
      gender |   388.4715   15.28532    25.41   0.000     358.5114    418.4316
      sector |   124.4813   15.95972     7.80   0.000     93.19936    155.7633
       _cons |  -769.0134   26.87791   -28.61   0.000    -821.6957   -716.3312
------------------------------------------------------------------------------

.
end of do-file

.
```

I have seen the sector; if sector is 2, it has; so, since I have 2 values, sector is equal to 1 and sector is equal to 2, and by convention, Stata will choose sector is equal to 1, the lower value as the reference category. So, the category, sector is equal to 2, that has 124.4 rupees higher earning per week. So, here, the way the NSS data is coded, the sector is equal to 2 means it is urban; it is a city. So, city people actually earns rupees 124.4 higher than a rural person.

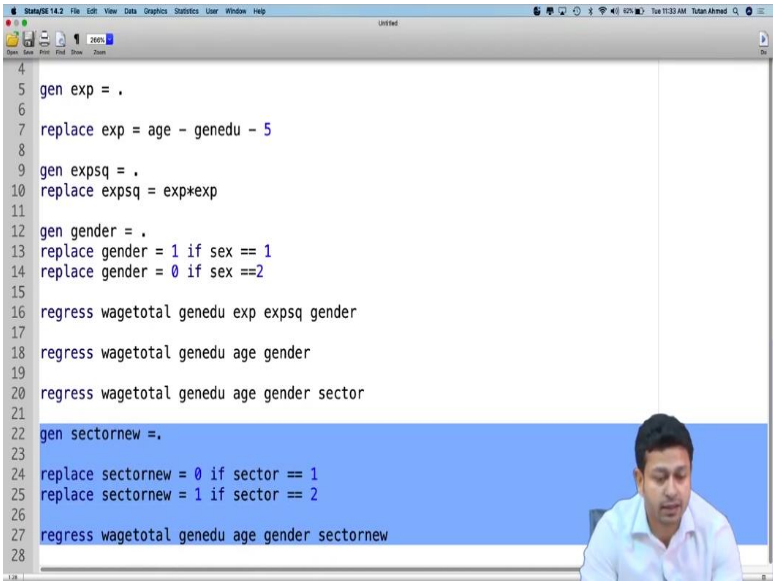
So, that is what the interpretation is. And the P value is very low; so, it means it is a significant result. The difference is significant. Now, that is, I have not changed anything. I have used the variable just as it is. Now, if I actually; like in the previous case, if I want to create a new variable, let us say sector new, is dot. And what I am doing, I am simply changing the value of the dummy variable; that is all. And why you do that?

Usually, when we have the dummy category, we keep it as 0 and 1, binary dummy category; 0 as the reference category and 1 as the other category. So, you see, the results are not changing. In the previous example, we have seen, the results are not changing. If it is 1 and 2 and 0 and 1, there is no difference, because it is always considering the relative difference. So, 1 and 2, the relative difference is 1; the 0 and 1, the relative difference is 1.

So, there is no change because of that. Only thing that here we are ensuring that we basically can assign which one we want as a reference category. So, the moment we put a value is equal to 0, so, that category becomes a reference category. So, let us do that. So, replace sector new is equal to; let us say I want to have my, let us say rural as 0, if sector; so, previously, my rural was I think 1. Replace sector new is equal to 1 if sector is equal to 2.

So, if it is a city, in a NSS Stata city means 2. So, I have basically created the variable 1. So, in this case, if I run a regression, my base category is going to be the rural, and the urban is going to be the other category. So, it will measure the impact of being in the city vis-a-vis the impact of being in a rural area. So, the regression equation will not change, because previously also, the ordering was that, like, the way we have seen the values, the 1 was for the rural category and 2 for the urban category. So, it will always by default measure the impact of being in 2 vis-a-vis the impact of being in 1. So, here, the regression equation, we will get the same regression equation.

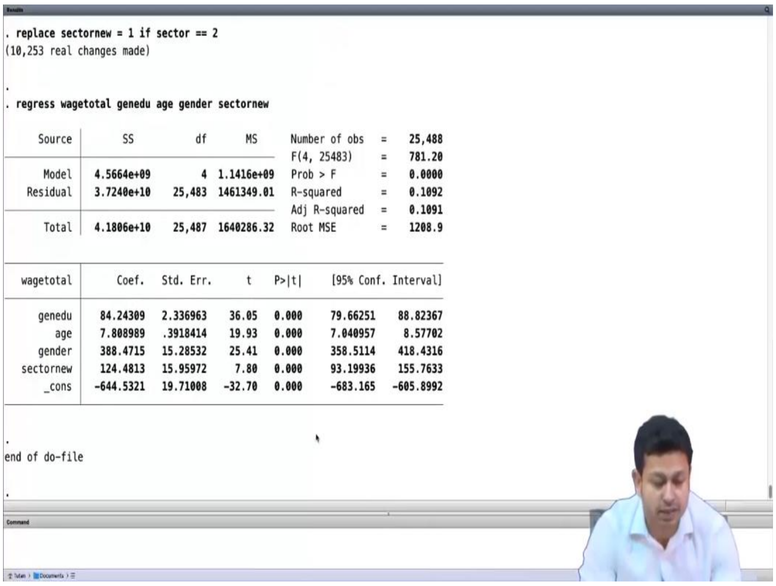
(Refer Slide Time: 13:17)



```
gen exp = .

replace exp = age - genedu - 5

gen expsq = .
replace expsq = exp*exp

gen gender = .
replace gender = 1 if sex == 1
replace gender = 0 if sex ==2

regress wagetotal genedu exp expsq gender

regress wagetotal genedu age gender

regress wagetotal genedu age gender sector

gen sectornew =.

replace sectornew = 0 if sector == 1
replace sectornew = 1 if sector == 2

regress wagetotal genedu age gender sectornew
```

All I have done, I have created a new variable, sector new. So, let me run the full code here. And we will see, we got the same regression equation here.

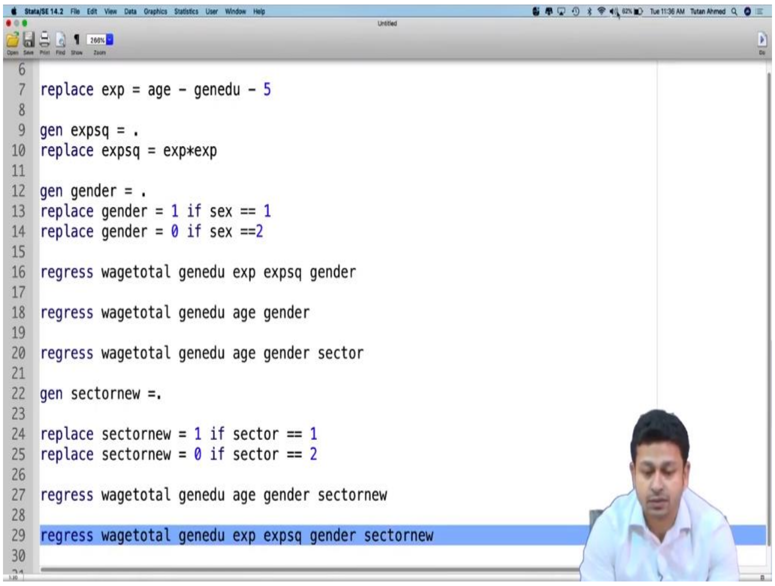
(Refer Slide Time: 13:32)



```
. replace sectornew = 1 if sector == 2
(10,253 real changes made)

.
. regress wagetotal genedu age gender sectornew

      Source |       SS           df       MS      Number of obs   =    25,488
-------------+----------------------------------   F(4, 25483)     =    781.20
       Model |  4.5664e+09         4  1.1416e+09   Prob > F        =    0.0000
    Residual |  3.7240e+10    25,483  1461349.01   R-squared       =    0.1092
-------------+----------------------------------   Adj R-squared   =    0.1091
       Total |  4.1806e+10    25,487  1640286.32   Root MSE        =    1208.9

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   84.24309   2.336963    36.05   0.000     79.66251    88.82367
         age |   7.808989   .3918414    19.93   0.000     7.040957     8.57702
      gender |   388.4715   15.28532    25.41   0.000     358.5114    418.4316
   sectornew |   124.4813   15.95972     7.80   0.000     93.19936    155.7633
       _cons |  -644.5321   19.71008   -32.70   0.000     -683.165   -605.8992
------------------------------------------------------------------------------

.
end of do-file

.
```

Just the same regression, 124.48, right? And of course, the other variables are remaining constant; I am not doing anything with other variables. Only thing that we are concerned here is the sector new value. And we will see, the sector new value is going to give the same value for the coefficient, 124.48, because we have only changed the value of the dummy variable. And since the measurement is happening on a relative scale, there is no difference in the regression equation.

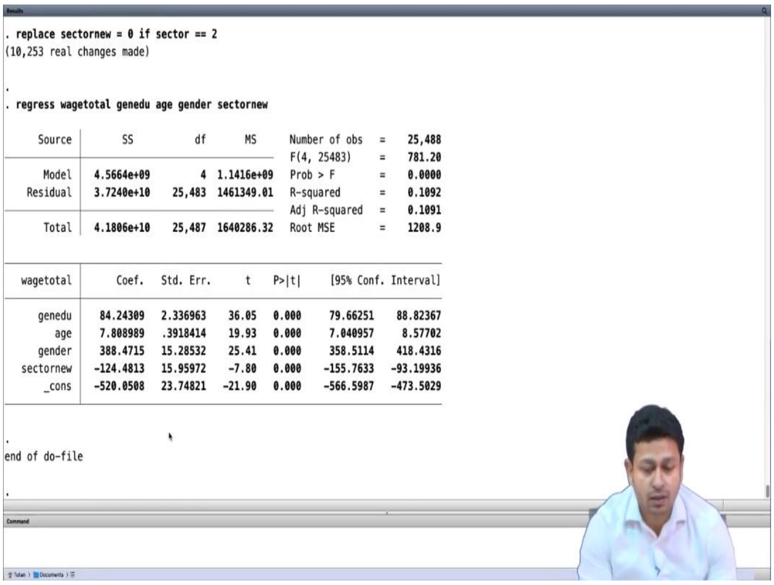
(Refer Slide Time: 14:03)



```
replace exp = age - genedu - 5

gen expsq = .
replace expsq = exp*exp

gen gender = .
replace gender = 1 if sex == 1
replace gender = 0 if sex ==2

regress wagetotal genedu exp expsq gender

regress wagetotal genedu age gender

regress wagetotal genedu age gender sector

gen sectornew =.

replace sectornew = 1 if sector == 1
replace sectornew = 0 if sector == 2

regress wagetotal genedu age gender sectornew

regress wagetotal genedu exp expsq gender sectornew
```

But if you suppose, instead of having rural as the base, let us say I want to have urban as the base. So, I create sector new is equal to 1 if the sector is equal to 1, so, which is rural. I still have rural is equal to 1 in sector new. And here, I create urban is the reference category. So, it is 0. So, all I have done, I have changed the reference category. So, now my urban is my reference category. So, now, if I run this, all that I am going to see is that, the sign has changed. Only thing that has happened is that the sign has changed; the value remains the same.

(Refer Slide Time: 14:44)



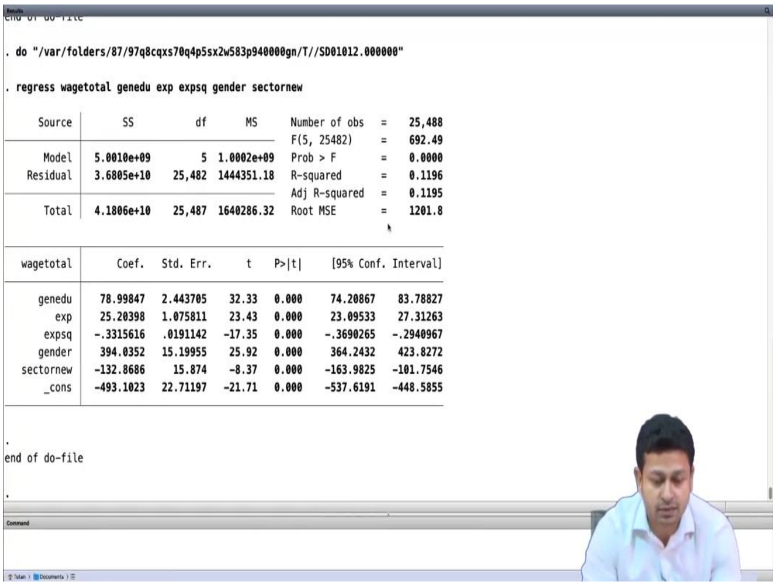
```
. replace sectornew = 0 if sector == 2
(10,253 real changes made)

.
. regress wagetotal genedu age gender sectornew

      Source |       SS           df       MS      Number of obs   =    25,488
-------------+----------------------------------   F(4, 25483)     =    781.20
       Model |  4.5664e+09         4  1.1416e+09   Prob > F        =    0.0000
    Residual |  3.7240e+10    25,483  1461349.01   R-squared       =    0.1092
-------------+----------------------------------   Adj R-squared   =    0.1091
       Total |  4.1806e+10    25,487  1640286.32   Root MSE        =    1208.9

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   84.24309   2.336963    36.05   0.000     79.66251    88.82367
         age |   7.808989   .3918414    19.93   0.000     7.040957     8.57702
      gender |   388.4715   15.28532    25.41   0.000     358.5114    418.4316
   sectornew |  -124.4813   15.95972    -7.80   0.000    -155.7633   -93.19936
       _cons |  -520.0508   23.74821   -21.90   0.000    -566.5987   -473.5029
------------------------------------------------------------------------------

.
end of do-file

.
```

So, -124.48 with a P value very small, which is basically, the result is significant. So, if you compare with the previous one, it is a positive one; and here it is a negative one. And it is only because I have changed the reference category. Now, I am comparing the wage of a rural person with respect to a wage of an urban person. So, all that I am saying, the wage of a rural person is rupees 124.48 less vis-a-vis the wage of a urban person every week.

So, essentially, the interpretation remains same. Now, we will just do a little bit of modification in the regression equation to fit into a standard regression equation. It is just for cosmetics purpose; it is not anything to do with the dummy variable, but I am just bringing a little change. So, because I have seen the R square value very small, I want to improve the R square value. And I can actually improve it by having right set of variables.

I can have experience, the experience we have created; experience square. If we run it, we will get somewhat better regression. So, we have already seen these in our functional specification.

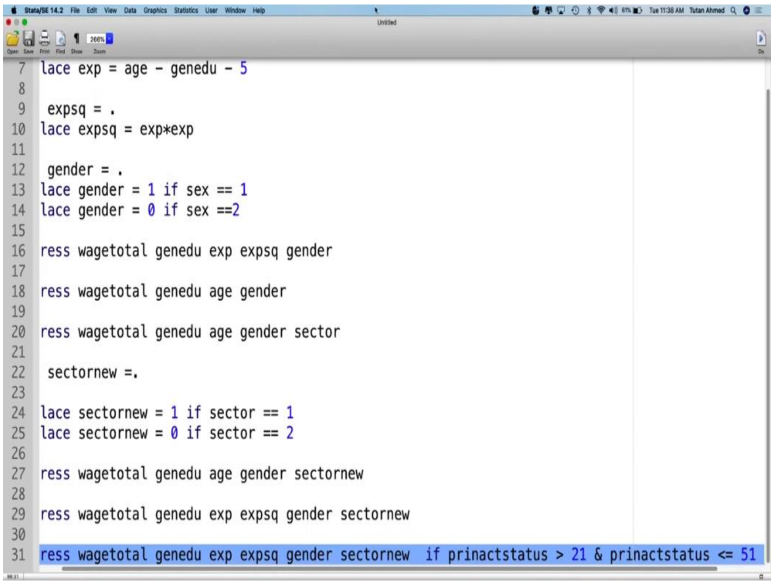
(Refer Slide Time: 16:16)



```
. do "/var/folders/87/97q8cqxs70q4p5sx2w583p940000gn/T//SD01012.000000"

. regress wagetotal genedu exp expsq gender sectornew

      Source |       SS           df       MS      Number of obs   =    25,488
-------------+----------------------------------   F(5, 25482)     =    692.49
       Model |  5.0010e+09         5  1.0002e+09   Prob > F        =    0.0000
    Residual |  3.6805e+10    25,482  1444351.18   R-squared       =    0.1196
-------------+----------------------------------   Adj R-squared   =    0.1195
       Total |  4.1806e+10    25,487  1640286.32   Root MSE        =    1201.8

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   78.99847   2.443705    32.33   0.000     74.20867    83.78827
         exp |   25.20398   1.075811    23.43   0.000     23.09533    27.31263
       expsq |  -.3315616   .0191142   -17.35   0.000    -.3690265   -.2940967
      gender |   394.0352   15.19955    25.92   0.000     364.2432    423.8272
   sectornew |  -132.8686     15.874    -8.37   0.000    -163.9825   -101.7546
       _cons |  -493.1023   22.71197   -21.71   0.000    -537.6191   -448.5855
------------------------------------------------------------------------------

.
end of do-file

.
```

R square will might have increased a little bit. So, it was 109; now it is 0.119; it has slightly improved. Also we need to get rid of some of the observations, because we are at this moment here basically talking about all the people in the population, but we are interested to understand the wage of those people who are actually working.

(Refer Slide Time: 16:38)

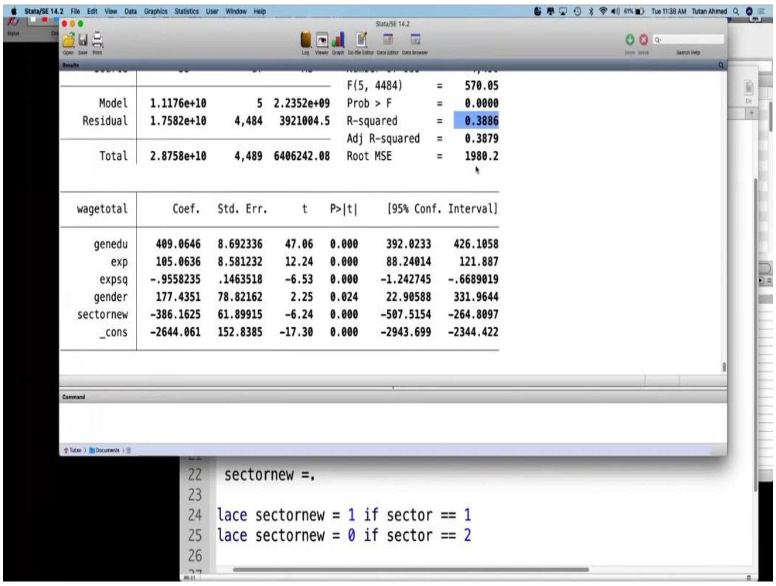


```
lace exp = age - genedu - 5

 expsq = .
lace expsq = exp*exp

 gender = .
lace gender = 1 if sex == 1
lace gender = 0 if sex ==2

ress wagetotal genedu exp expsq gender

ress wagetotal genedu age gender

ress wagetotal genedu age gender sector

 sectornew =.

lace sectornew = 1 if sector == 1
lace sectornew = 0 if sector == 2

ress wagetotal genedu age gender sectornew

ress wagetotal genedu exp expsq gender sectornew

ress wagetotal genedu exp expsq gender sectornew  if prinactstatus > 21 & prinactstatus <= 51
```

No, I do not really care about people if whether not working. So, there is no question of wage. So, I have; let me see the variable name; all I am doing is that, I am trying to see the occupational status of the people. So, there are certain certain codes in Stata that I am going to; so, what I have done, the way the national sample survey data is coded; so, if I take prinactstatus greater than 21 and less than equal to 51, it will include all those people who are in say regular salaried work or in casual work.

So, I want to understand the regression equation for this particular population. So, this is simply because I want to improve the regression equation; it has nothing to do with the dummy variable per se. So, let us see the regression equation here.

(Refer Slide Time: 17:34)



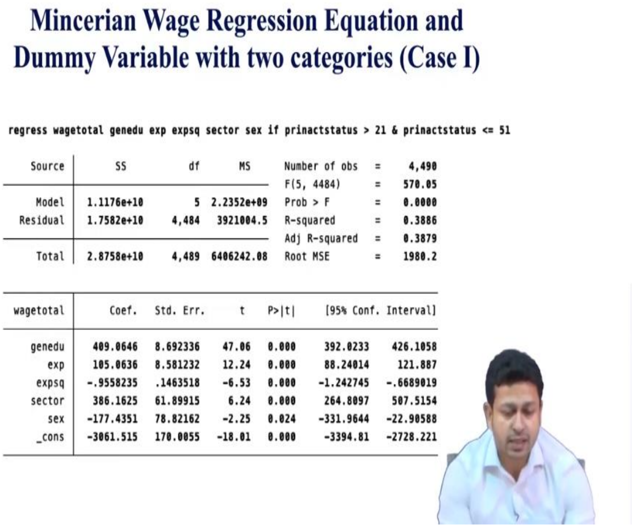
```
                                                  F(5, 4484)      =    570.05
       Model |  1.1176e+10         5  2.2352e+09   Prob > F        =    0.0000
    Residual |  1.7582e+10     4,484   3921004.5   R-squared       =    0.3886
                                                   Adj R-squared   =    0.3879
       Total |  2.8758e+10     4,489  6406242.08   Root MSE        =    1980.2


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

      genedu |   409.0646   8.692336    47.06   0.000     392.0233    426.1058
         exp |   105.0636   8.581232    12.24   0.000     88.24014     121.887
       expsq |  -.9558235   .1463518    -6.53   0.000    -1.242745   -.6689019
      gender |   177.4351   78.82162     2.25   0.024     22.90588    331.9644
   sectornew |  -386.1625   61.89915    -6.24   0.000    -507.5154   -264.8097
       _cons |  -2644.061   152.8385   -17.30   0.000    -2943.699   -2344.422
```

```
 sectornew =.

lace sectornew = 1 if sector == 1
lace sectornew = 0 if sector == 2
```

It has improved significantly. R squared value is almost like 0.4, because I am out of now to relevant sub-sample here. I am talking about prinactstatus here, the occupation status. So, people who are already in regular employment or its casual employment; so, I am talking about them. So, of course, the moment you change the sub-population or you bring new variables like experience, experience square, of course, the regression coefficients are going to change.

And we see that these are the regression equations. So, we have now this, gender and sector new, this 2 dummy variable. So, given these, let us now actually try to understand what exactly;

(Refer Slide Time: 18:20)

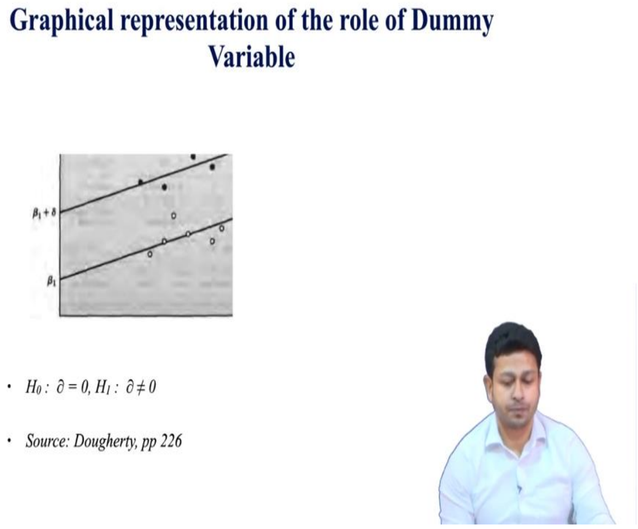


Minceriann Wage Regression Equation and Dummy Variable with two categories (Case I)

```
regress wagetotal genedu exp expsq sector sex if prinactstatus > 21 & prinactstatus <= 51

      Source |       SS           df       MS      Number of obs   =     4,490
-------------+----------------------------------   F(5, 4484)      =    570.05
       Model |  1.1176e+10         5  2.2352e+09   Prob > F        =    0.0000
    Residual |  1.7582e+10     4,484   3921004.5   R-squared       =    0.3886
-------------+----------------------------------   Adj R-squared   =    0.3879
       Total |  2.8758e+10     4,489  6406242.08   Root MSE        =    1980.2

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   409.0646   8.692336    47.06   0.000     392.0233    426.1058
         exp |   105.0636   8.581232    12.24   0.000     88.24014     121.887
       expsq |  -.9558235   .1463518    -6.53   0.000    -1.242745   -.6689019
      sector |   386.1625   61.89915     6.24   0.000     264.8097    507.5154
         sex |  -177.4351   78.82162    -2.25   0.024    -331.9644   -22.90588
       _cons |  -3061.515   170.0055   -18.01   0.000     -3394.81   -2728.221
------------------------------------------------------------------------------
```

So, this is the regression equation which is run. So, we will just see how do we really interpret the dummy variable here. So, what really is happening here?

(Refer Slide Time: 18:29)

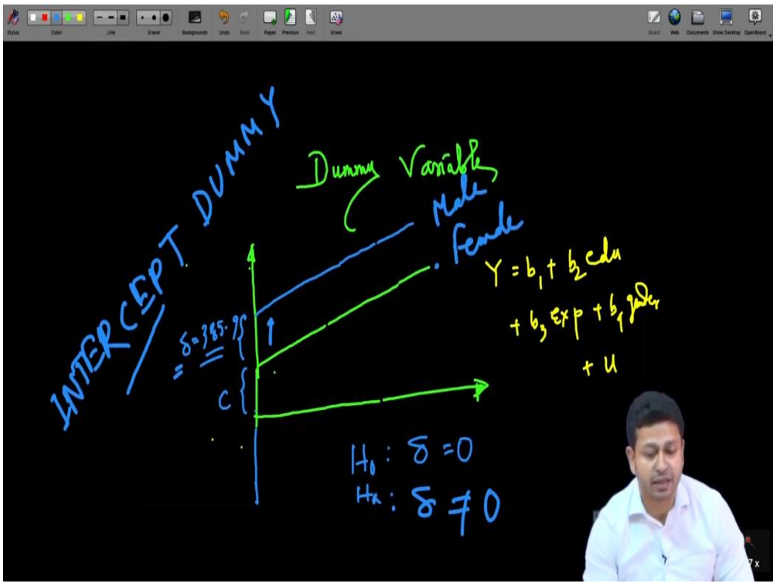


Graphical representation of the role of Dummy Variable

- $H_0: \partial = 0, H_1: \partial \neq 0$
- *Source: Dougherty, pp 226*

So, I will explain this here with this graphical representation. And I will actually draw it for you.

(Refer Slide Time: 18:40)



INTERCEPT DUMMY

Dummy Variable

Male

Female

$Y = b_1 + b_2\, edu + b_3\, exp + b_4\, gender + u$

$\delta = 385.95$

C

$H_0: \delta = 0$

$H_a: \delta \neq 0$

So, essentially, what is happening here is; let us say we have been talking about the gender dummy. Let us say we will talk about only one dummy variable here. So, let us say my regression equation initially is; let us say the reference category is female. So, let me write down the regression equation. Y, the wage variable is equal to some intercept plus, let us say beta 2 education; beta 3 experience or age, whichever you want to take plus beta 4 gender plus some error term.

Now, when I have female as a base category, all the value, all that I am getting is the coefficient for the female. So, here, for gender it is; let us say, if I go back to the previous equation where you have only gender, what we will see is, there is specific value 385.9. So, there, my female was the best category. So, by 385.9, we mean that by virtue of being male, your earning is increasing by only a very constant value, a constant value 385.9.

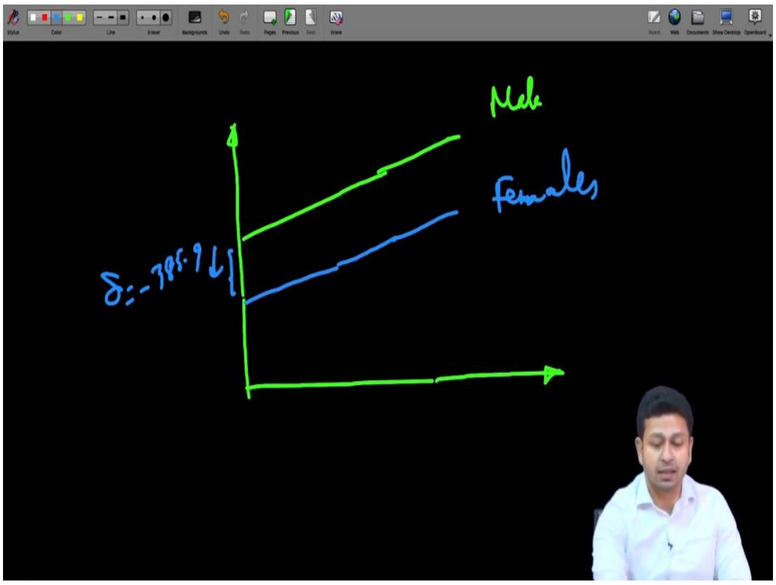
So, here, now, if I am a male, so, the regression equation is just going to be like this; a parallel regression equation, because only thing that is changing is this delta, which is in our case 385.9. And our intercept was here -625. So, it has to be on the other side of the quadrant, but the idea is the same. So, if whatever is the constant term is here, you have a delta term here which is basically differentiating the 2 regression equation, one for male and another for female.

And the male, because they are, this is the additional value that is getting added, because of the value of the coefficient; so, if I am a male, I should be earning this much more than the females. Now, what is a null hypothesis? Of course null hypothesis is the status quo, is that there is no impact. So, basically, delta is equal to 0. And whereas, my alternative hypothesis is that delta is not equal to 0.

So, any impact due to the dummy variable that is significant; so, that is my alternative hypothesis. So, this is essentially, we understand the dummy variable, and because it is only influencing the intercept, we call it as intercept dummy. So, that is basically the idea. Now, if I have, let us say, instead of gender, I would have rural and urban; so, I would have some reference dummy; let us say, if it is rural and if I add the urban here, so, it would have been like the same equation with a value of delta a different value of delta.

So, we could have very well the reference dummy, let us say; this is for female in our previous equation, and it is for male. Now, I could have very well male as the reference dummy. So, if the male is the reference dummy, this is the same diagram.

(Refer Slide Time: 22:39)



We have the same diagram; the male as the reference dummy here. This is the equation for males. And the female is the equation which we are estimating with respect to males. So, in that case, so, our delta was like, would have been minus of 385.9. So, if you subtract that; so, all that is happening is that, the intercept value, this value is actually declining for the females. And we get the equation for females.

So, this is how we essentially understand the basic concept of dummy variable, the intercept dummy. And in the next lecture, we are going to see more about the other categories of dummy variable, and we are going to explain going forward, what are the the slope dummy and how to include slope intercept dummy together in the regression equation. Thank you.