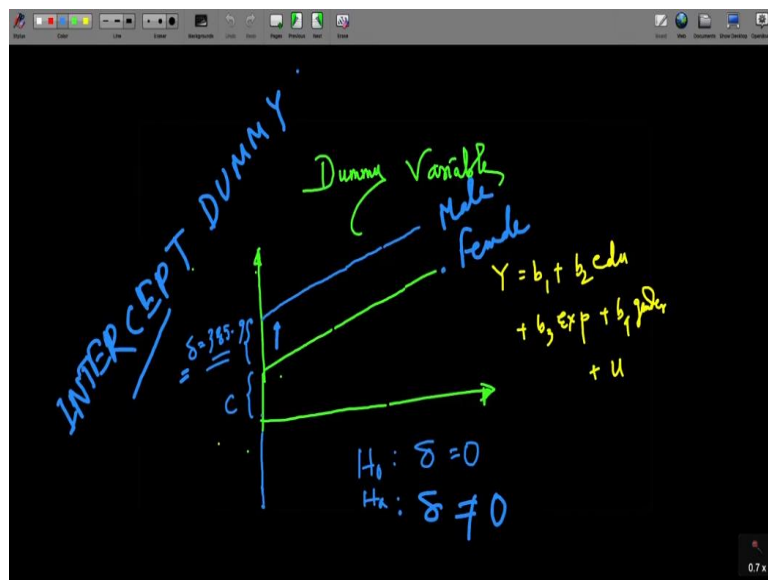


Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 7
Lecture - 59
Dummy Variable (Contd.)

Hello and welcome back to the lecture on Applied Econometrics, and we have been talking about dummy variable. And in the previous lecture, we explained what is an intercept dummy, and we have shown graphically the impact of the change in the value of the dummy variable. So, what we have shown here;

(Refer Slide Time: 00:42)



If someone is a male, so, then, that person's income is increasing by rupees 385.9 rupees every week vis-a-vis if a person is a female. Now, we have shown previously that how we really code the dummy variables.

(Refer Slide Time: 01:01)

```
11
12 gen gender = .
13 replace gender = 1 if sex == 1
14 replace gender = 0 if sex == 2
15
16 regress wagetotal genedu exp expsq gender
17
18 regress wagetotal genedu age gender
19
20 regress wagetotal genedu age gender sector
21
22 gen sectornew = .
23
24 replace sectornew = 1 if sector == 1
25 replace sectornew = 0 if sector == 2
26
27 regress wagetotal genedu age gender sectornew
28
29 regress wagetotal genedu exp expsq gender sectornew
30
31 regress wagetotal genedu exp expsq gender sectornew if prinactstatus > 21 & prinactstatus <=
32
33 replace gender = 5 if sex == 1
34 replace gender = 3 if sex == 2
35
36 regress wagetotal genedu age gender
37
```

We have shown that usually we have 0 and 1 as a value of dummy variable. So, it is for gender, it is for sector and so forth. Now, what if instead of 0 and 1, I would have, like, just because I want it, I could have marked it as 2 and 5 or 0 and 5 or 1 or 4, whatever I want. So, what would have happened then? So, let us actually try to do that. Let us say, in the previous case I had gender is equal to 1 if sex is equal to 1.

So, that is male is, I have assigned it as 1; whereas female, we have assigned it as 0. So, I will just use the same codes, just that I will change the values here. And here I am going to, let us say, if someone is male, I am going to write it as 5; and if someone is female, I am going to write it as 3. So, I will do the same regression equation, but I have changed the values here, the dummy variables. And I will just try to explain what will happen. So, I will run the full code here.

(Refer Slide Time: 02:20)

```

(12,515 real changes made)

. regress wagetotal genedu age gender

```

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

	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	192.9676	7.649897	25.22	0.000	177.9733 207.9618
_cons	-1204.076	34.56343	-34.84	0.000	-1271.823 -1136.33

```

.
end of do-file

```

And if I run it, I will see that the value has become this, 192.9. So, previously, we have seen the value was 385.9, and now it has become 192.96. Now, what is happening here? Now, you have to understand; so, this is a good example that will give us a sense of how to interpret the effect of the dummy variable. So, you have to understand one thing that in the previous case, when I assigned male and female 0 and 1, so, my Stata, my program was actually trying to explain the effect of the dummy variable for 1 unit.

So, the moment you move from 0 to 1, you are moving 1 unit. So, 1 unit change was giving me a value of 385.9. But now, what I have done here is, you see that I have valued it as 5 and 3. So you remember that always in dummy variable, we are actually measuring the relative impact. So, here, I am measuring the impact. So, 5 is the male and 3 is female. So, I am measuring the impact of male vis-a-vis female.

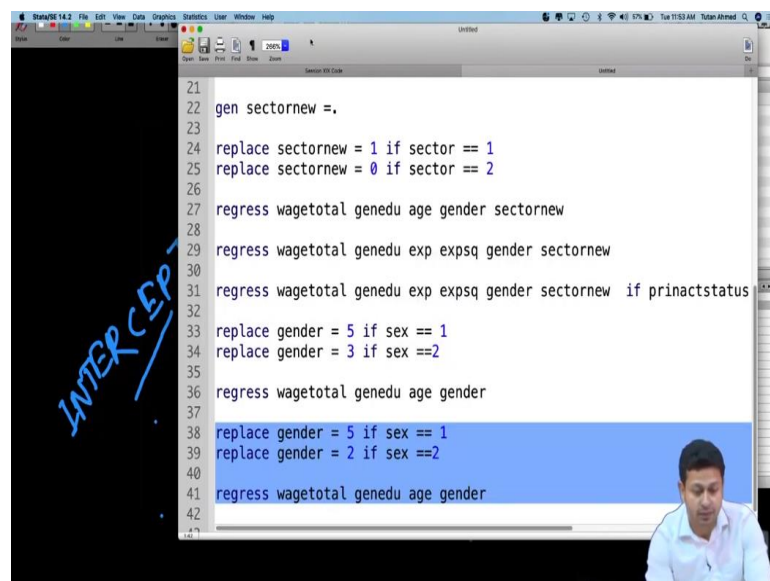
So, basically, I am measuring the impact from 3 to 5. So, I am measuring the impact of 2 units. So, I have moved from 3 to 5. So, essentially, I am representing the impact of 2 units. So, I am saying the program that okay, I have changed my dummy variable by 2 units, so, what is the impact? So, what the program has done simply? It has actually, you remember, the coefficient means the impact per unit change.

It is saying that okay, so, here you have changed 2 units, but per unit change is 192.96; so, which means that if you multiply that with 2, so, which will actually give you the full length difference between male and female, that is going to be 192.9 into 2, is basically the same

value that we got, 385.9. So, essentially, all that is saying that, it simply depends on how you choose to define the difference.

So, if the difference is of 2 units, so, then, what we will do is, like the way it interprets is, it will give you the impact for 1 unit. And then, to get the full unit, you have to multiply by the number of units you have differentiated between these 2 categories. So, essentially, you will have to multiply here by 2. Now, instead of 5 and 3, if I would have given a value of; I will write it again here; let us say 5 and 2.

(Refer Slide Time: 05:15)



```
21
22 gen sectornew =.
23
24 replace sectornew = 1 if sector == 1
25 replace sectornew = 0 if sector == 2
26
27 regress wagetotal gnedu age gender sectornew
28
29 regress wagetotal gnedu exp expsq gender sectornew
30
31 regress wagetotal gnedu exp expsq gender sectornew if prinactstatus
32
33 replace gender = 5 if sex == 1
34 replace gender = 3 if sex ==2
35
36 regress wagetotal gnedu age gender
37
38 replace gender = 5 if sex == 1
39 replace gender = 2 if sex ==2
40
41 regress wagetotal gnedu age gender
42
```

So, that means, now I have a difference of 3 between these 2 categories. And what I will get is that, I am going to get an even smaller value of the coefficient. So, one-third of 385, we are going to get, because we have to multiply 3 to reach there. So, if you go back, I will see that it is going to be 128.6.

(Refer Slide Time: 05:41)

Stata Command Window Output:

```

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

. regress wagetotal genedu age gender

```

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

	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	128.645	5.099932	25.22	0.000	118.6489 138.6412
_cons	-882.4638	24.77505	-35.62	0.000	-931.0243 -833.9033

end of do-file

Now, if you multiply 128.6 into 3; now the difference is 3, and per unit change is 125.6, so, if you have to multiply by 3 to get the full difference between male and female; so, if you do that, so, 128.6, it is going to give you the value of 385.9 that we had in the first place. So, does not matter how you define the values of the dummy variable, the interpretation remains same. So, you have to just make sure that you are interpreting it correctly; but it is just for illustration purpose, usually we always keep it as 0 and 1. So, this is something just for our understanding.

(Refer Slide Time: 06:29)

Intercept Dummy

- Concept of intercept dummy: These dummy coefficients are often called **differential intercept dummies**, for they show the differences in the intercept values of the category that gets the value of 1 as compared to the reference category.
- Why do we have a negative intercept value? - mechanical interpretation of intercept

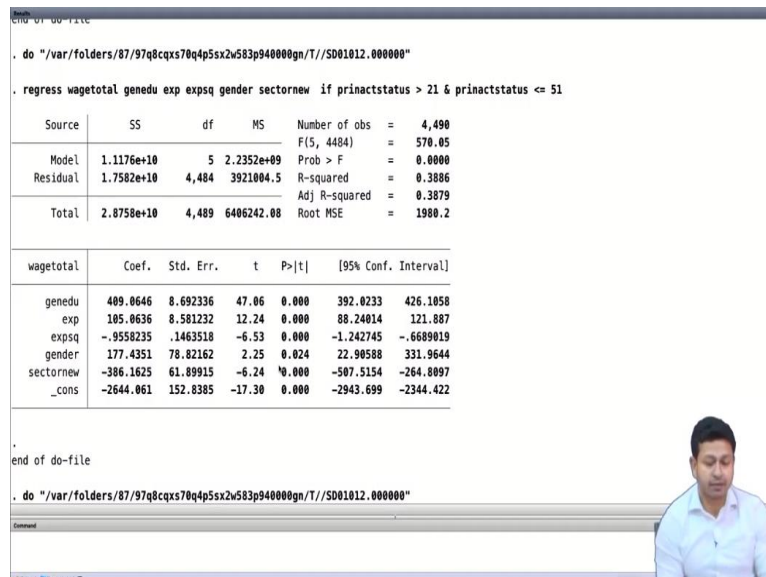
Now, let us talk about something that we have seen previously. In the previous equation, we have seen for the intercept term. For the intercept term, what you have seen? It is a negative value. So, negative 882.4, and that too significant. So, what does that mean? So, it essentially

means that, if you have everything is equal to 0, all the explanatory variable is equal to 0, a person's income is going to be minus of 882.4 rupees.

Now, how is that possible? So, often time, what happens? Previously you have seen, if your model is not correctly specified or if the functional specification is not correct, so, then, what will happen is that, you will get a constant term which is not really explainable. So, it is basically, if you actually have a better model with all the different variables included, you are likely to get a constant term which would make more sense. So, that is it.

So, at this moment, our model is not really correctly specified; so, that is why we see a constant term which may not have a very good meaning. Now, we have seen in the previous equation, if we go back to the the diagram, we have seen our equation for males and females, so, they are essentially same. The only thing that differentiates a male's equation vis-a-vis a female's equation is this delta term. So, let us actually go back to the regression equation. So, if I take the full equation here, where we have used experience, experience square and so forth;

(Refer Slide Time: 08:31)



```

. do "/var/folders/87/97q8cqs70q4p5sx2w583p940000gn/T//SD01012.000000"

. regress wagetotal genedu exp expsq gender sectornew if prinactstatus > 21 & prinactstatus <= 51

```

Source	SS	df	MS	Number of obs	=	4,490
Model	1.1176e+10	5	2.2352e+09	F(5, 4484)	=	570.05
Residual	1.7582e+10	4,484	3921004.5	Prob > F	=	0.0000
				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	78.02162	2.25	0.024	22.90508 331.9644
sectornew	-306.1625	61.89915	-6.24	0.000	-507.5154 -204.8097
_cons	-2644.061	152.8385	-17.30	0.000	-2943.699 -2344.422

```

.
end of do-file

. do "/var/folders/87/97q8cqs70q4p5sx2w583p940000gn/T//SD01012.000000"

```

So, here, if I have this full equation, so, how I will write it down? I will write it down as;

(Refer Slide Time: 08:43)

$$\begin{aligned}
 Y &= 409 \text{ Gen Edu} + 105 \text{ Exp} \\
 &+ -0.96 \text{ Exp}^2 + 177.4 \text{ Gender} - 386 \text{ sector} \\
 &\quad - 2644 \\
 &= 409 \text{ Gen Edu} + 105 \text{ Exp} - 0.96 \text{ Exp}^2 \\
 &\quad + 177.4 - 386 \text{ sector} - 2644 \text{ (Male)} \\
 &= 409 \text{ Gen Edu} + 105 \text{ Exp} - 0.96 \text{ Exp}^2 \\
 &\quad - 386 \text{ sector} - 2644 \text{ (Female)}
 \end{aligned}$$

So, my wage is equal to 409 into general education plus 105 into experience plus we have minus of 0.96 into experience square plus 177.4 into gender minus 386 into sector plus our constant term which is minus of 2644. So, this is my regression equation. Now, all that we will have from this equation is that we actually have for regression equation for a male and a regression equation for a female.

So, in this case, my gender is equal to 1 means, it is a regression equation for male. So, essentially, that would mean, if I put the value of gender is equal to 1, because that is the regression equation for male, so, it is going to be 409 into gen edu plus 105 into experience minus 0.96 into experience square. So, if I put this gender is equal to 1, it will mean that I have to add 177.4, and then I can subtract 386 sector, and then, of course the constant term which is 2644.

So, note that here the coefficient for general education is same as the first equation. The other coefficients, experience, experience square, all, they are going to remain same. So, this is the equation for let us say male. And what will happen for females? What will happen for females is that, the same thing, just that you will not include this part. So, it is 409 general education plus 105 experience minus 0.96 experience square.

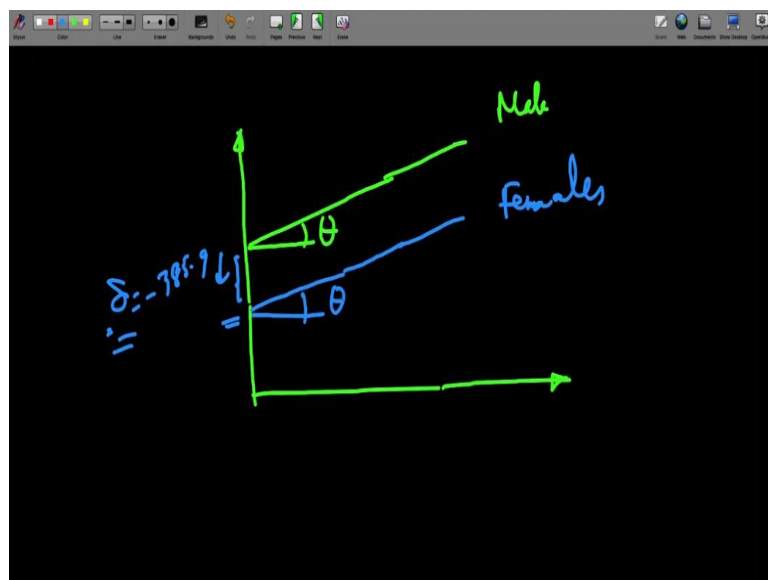
And then, you basically do not include this value here 177.4, because here, your gender term is nothing but 0. So, 0 into this value is going to be 0. So, you do not include that. And what do we have here is; let us minimise a little bit; minus 386 into sector minus 2644. So,

everything else remains the same. Only thing that is changing is that, you are basically vanishing this value of the gender coefficient for females.

So, there is no corresponding value here. So, only thing here that is changing is that, for males, I have just added the delta term which is in this case 177.4. And whereas, for females it is not there. So, for females, the base equation, you do not include any value for the gender category, the dummy category, because it is the reference category; but for males, you will simply add the impact for gender, because the way you have defined it, male is equal to 1.

So, that is how we understand the regression equation. So, essentially, the crucial point that remains is that; of course, we have explained these 2 regression line; but the crucial point that remains is that the slope; slope for the 2 regression equation is same. So, the values here, the 409 for male is same as the value for 409 for female. Similarly, experience 105, 105 and 386, 386; all other things are remaining same, remaining constant.

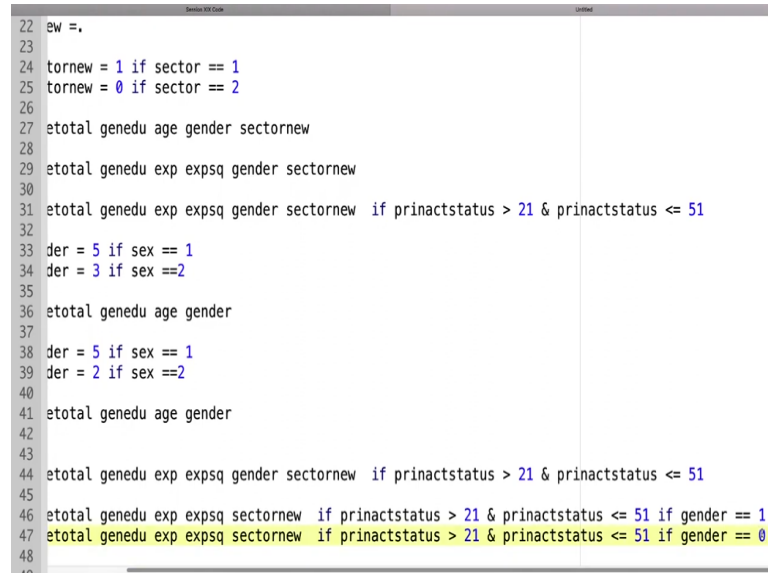
(Refer Slide Time: 13:31)



So, and that is what is quite obvious, because the slope of these 2 line, they are essentially parallel lines; so, slope of these 2 line are going to be same, because they are parallel. Now, what does that really mean? What that means is that, it is as if the impact of education and experience and experience square; so, if it is a theta, it is a theta. So, it means that, as if the impact of education and experience experience square, they are all basically same for males and females.

But if we run a regression equation actually separately for males and females, will we get the same result? So, let us see that. Now, actually, this is the equation. And what I am going to do is, I am just going to copy; I do not want to write it again.

(Refer Slide Time: 14:22)



```
22 pw =.
23
24 tornew = 1 if sector == 1
25 tornew = 0 if sector == 2
26
27 etotal gnedu age gender sectornew
28
29 etotal gnedu exp expsq gender sectornew
30
31 etotal gnedu exp expsq gender sectornew if prinactstatus > 21 & prinactstatus <= 51
32
33 der = 5 if sex == 1
34 der = 3 if sex ==2
35
36 etotal gnedu age gender
37
38 der = 5 if sex == 1
39 der = 2 if sex ==2
40
41 etotal gnedu age gender
42
43
44 etotal gnedu exp expsq gender sectornew if prinactstatus > 21 & prinactstatus <= 51
45
46 etotal gnedu exp expsq sectornew if prinactstatus > 21 & prinactstatus <= 51 if gender == 1
47 etotal gnedu exp expsq sectornew if prinactstatus > 21 & prinactstatus <= 51 if gender == 0
48
```

And what I am going to see is; so, let us actually copy paste it here. So, I had this full equation. Now, I will not have the gender dummy here; I remove it. Instead, I will run the regression equation 1 it for male, and another time it is for female. Let us actually run it first; let me actually run a separate equation here. And here, I remove gender. So, first I will run the full equation, then I will run one for male and one for female.

If gender is equal to 1; and here, again copy this regression equation; and here I will have, if gender is equal to 0. So, that means, 1 is for male and 0 is for female. Actually, let me; before that, because I have previously assigned these values here; so, Stata is reading gender is equal to 2 and 5 here. So, let me actually change the value of gender first. Here, let me run this. Now, I will get the right value of the dummy coefficient. Now, let us run it first. So, I have run the equation as it is with gender as a dummy variable.

(Refer Slide Time: 16:11)

```

do "var/folders/87/97q8cqs70q4p5sx2w583p940000gn/T//SD01012.000000"

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

```

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

	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

```

*
end of do-file
*

```

And if I actually see it, I have the regression equation here. Now, we will see this.

(Refer Slide Time: 16:32)

```

do "var/folders/87/97q8cqs70q4p5sx2w583p940000gn/T//SD01012.000000"

regress wagetotal genedu exp expsq sectornew if prinactstatus > 21 & prinactstatus <= 51 & gender == 0

```

Source	SS	df	MS	Number of obs	F(4, 769)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	1.6167e+09	4	404170558	774	147.23	0.0000	0.4337	0.4308	1656.8
Residual	2.1110e+09	769	2745082.47						
Total	3.7277e+09	773	4822316.49						

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
genedu	349.703	15.78237	22.16	0.000	318.7213 380.6846
exp	95.82445	18.04761	5.31	0.000	60.39603 131.2529
expsq	-1.021933	.3027903	-3.38	0.001	-1.616327 -.4275396
sectornew	-546.6886	122.534	-4.46	0.000	-787.2294 -306.1478
_cons	-1908.897	294.8464	-6.47	0.000	-2487.696 -1330.097

```

*
end of do-file
*

```



```

do "var/folders/87/97q8cqs70q4p5sx2w583p940000gn/T//SD01012.000000"

regress wagetotal genedu exp expsq sectornew if prinactstatus > 21 & prinactstatus <= 51 & gender == 1

```

Source	SS	df	MS	Number of obs	F(4, 3711)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	9.5037e+09	4	2.3759e+09	3,716	571.90	0.0000	0.3814	0.3807	2038.2
Residual	1.5417e+10	3,711	4154394.75						
Total	2.4921e+10	3,715	6708107.91						

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
genedu	424.5305	10.17237	41.73	0.000	404.5065 444.4745
exp	104.3936	9.687352	10.78	0.000	85.40055 123.3867
expsq	-.9074391	.1655687	-5.48	0.000	-1.232054 -.5028245
sectornew	-343.425	70.47412	-4.87	0.000	-481.5968 -205.2531
_cons	-2615.933	158.7068	-16.48	0.000	-2927.094 -2304.772

```

*
regress wagetotal genedu exp expsq sectornew if prinactstatus > 21 & prinactstatus <= 51 & gender == 0

```

Source	SS	df	MS	Number of obs	F(4, 769)
Model	1.6167e+09	4	404170558	774	147.23

We got the regression equation here; one for male, one for female, and previously we had the regression equation for both the categories together, with gender as a dummy variable. So, first thing to note, we had this number of observation 4490 for the full sample. And here, we had 3716 for males. So, there are 3716 males, because I have specified gender is equal to 1. So, this is a sub-sample regression, where we had regression only for the male category.

And for females, when gender is equal to 0; so, I got a number of observations 774. So, if I add these two, 774 and 3716, we will actually end up to this number which is 4490. Now, what you have seen previously is that, when we explained the intercept dummy, we explained that the slopes are going to be same. And if the slopes are going to be same, that means, in the regression equation, we will have the value of the coefficients for other explanatory variable, they have to be same, otherwise the regression line, the slopes are not going to be same.

So, essentially, if that is true, my sub-population regression, the sub-sample regression should actually provide the coefficient for general education as 409, experience should come as 105 and experience square should come as -0.55. So, let us see if we are actually getting that. And you see, we are getting here 424.5; experience is 104.3; and experience square has become minus of 0.9. Now, this is for the 1 regression, this is for males.

And for females, we see the impact of general education is actually 349, which is quite different from 409 that we had for the regression with both males and females. Or if you see experience square, it is also less; it is 95.8. Or experience square, it is minus of 1.02. So, there are actually, the coefficients are not same. If you run a sub-sample regression, the coefficients are not same, which we actually claimed in our regression with the dummy variable.

So, that is an assumption we are making when we are running a regression equation with a dummy variable. So, the moment you have just the intercept dummy, it is as if our regression equation for the 2 categories are going to be same only except the difference in the intercept term. So, that is a big assumption we have made when we are running a dummy variable regression with the intercept dummy. So, how to really address that?

We are going to see in the next lecture, how to really address that; but before that, from the binary category of dummy, in the next lecture, we are going to talk about a dummy variable

with multiple categories and how to actually incorporate those multiple categories. And but of course, we need to explain the other type of dummy variable apart from intercept dummy, which is only explaining the intercept term but not the slope term. So, we will see these in the next couple of lectures. Thank you.