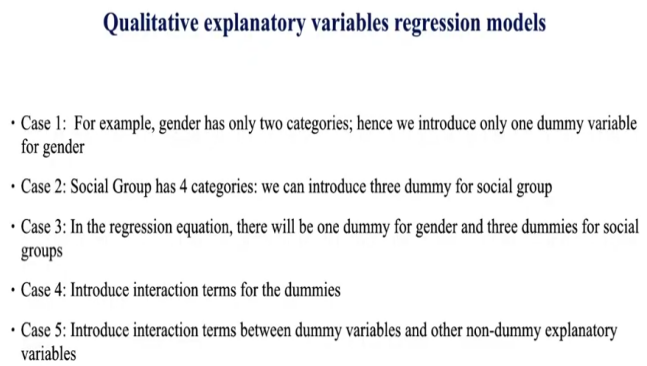
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 8
Lecture - 60
Dummy Variable (Contd.)

Hello and welcome back to the lesson on Applied Econometrics, and we have been talking about dummy variable. So far, we have covered a binary dummy variable. We have seen how we can incorporate a binary dummy variable in our regression equation. Now, as we said at the beginning that all dummy variables are not necessarily binary, so, we can have multiple categories. For example, region, religion, social group. So, they can have like 4 or 5 categories. Now, how to really include 4, 5 categories into a regression equation. This is basically going to be the topic of this lecture.

(Refer Slide Time: 00:56)



For example, we have social group which is 4 categories; we have seen, right? So, how do we really address that and that is what we are; this is basically our case 2.

(Refer Slide Time: 01:03)

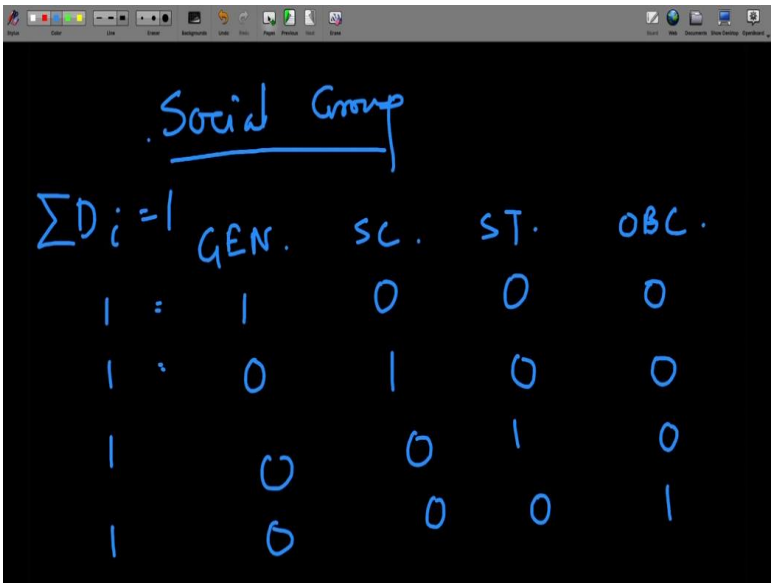
Reference Category of a Dummy Variable and its choice

- First, if an intercept is included in the model and if a qualitative variable has m categories, then introduce only (m – 1) dummy variables.
- The concept of Reference category or omitted category
- The choice of the Reference category
- *Can we include all the m categories of a dummy variable?*

Now, we have seen something called reference category already. And reference category, we have seen, if I have like a reference; in any dummy variable, I have to have a reference category. So, if I have a binary dummy variable, so, I have like gender; so, I have, say male or female; and I have female reference category. So, I include only 1 dummy variable for that, and that is representing the males.

And when I include the dummy variable, the impact due to the gender change is reflected in the coefficient for that dummy variable. Now, what happen if I have 5 categories? How do I really define the reference category? And how do I really define the other categories? So, essentially, how we do is, we actually define one particular variable name for each of these different categories of dummy variable. For example, let me show you.

(Refer Slide Time: 02:05)



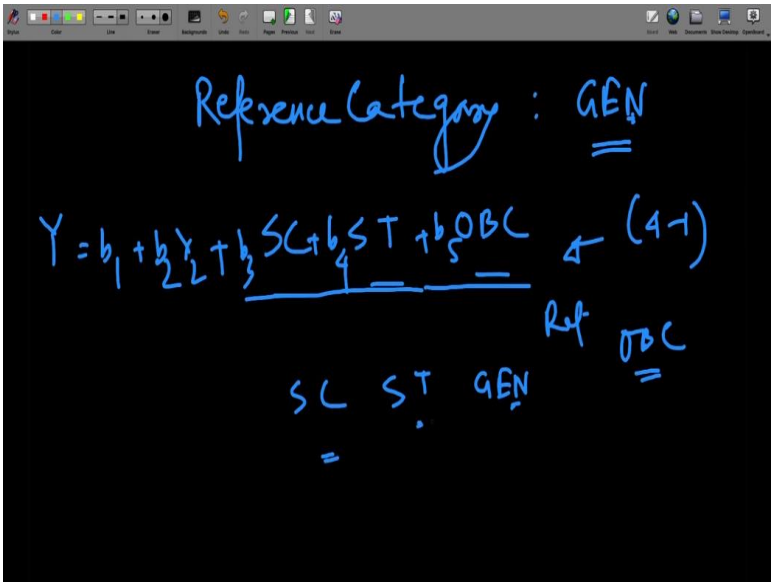
If I have the social group case, and I have let us say these 4 social groups. So, this General, SC, ST and OBC. And I have to assign a value for all of these different categories. Now, if someone is General, let us say the first person is General; so then, by virtue of being General, he is not SC, he is not ST or he is not OBC. Similarly, let us say a person is basically SC. Then, if someone is SC, then he is not General, he is not ST, he is not OBC.

Or if a person is ST, then that person is not any of these. And similarly, if a person is OBC, then he is not ST, he is not SC or he is not General. So, essentially, what we will do is, we will create dummy variable for each of these different categories. So, there could be a dummy variable for General; there could be a dummy variable for SC; there could be a dummy variable for ST, and there could be a dummy variable for OBC.

But note here, the moment you sum them up, if you basically take a sum, so, the sum value, here it is going to be 1. If I just want to take a sum of all these different values, it is going to give me a value of 1, because someone is 1, and then, automatically by default, he is going to be 0 for all other categories. If I sum up, so, essentially I can write, summation of this particular dummy D i for all these different categories; it is going to give me a value of 1.

So, this is how I will first create the dummy variables, for General, for SC, for ST and OBC. Now, I know from the previous example that I have to have a reference category.

(Refer Slide Time: 04:02)



So, which one is going to be my reference category? Now, in general, we take the reference category which is most prominent, or I know I have a lot of information about that category,

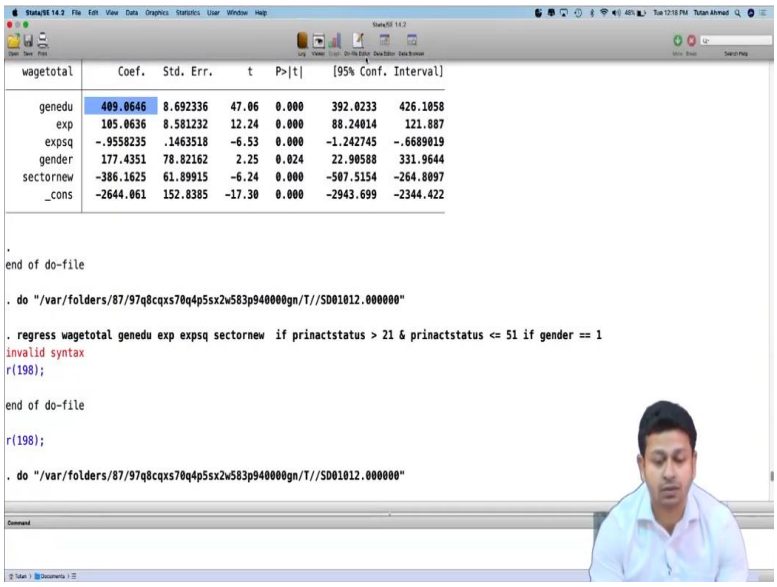
or which has the most number of observations, because, then I can measure the other categories with respect to the reference category. So, here, let us say in this case, I will take General as a reference category.

You can choose anything you want, but usually, it is easier for us to explain if I take the category which is like the most prominent, most well-known. So, I take that as a reference category. Now, when I take that as a reference category, in my regression equation, what I include is that, I include these variables: I include a variable for SC; I include a variable for ST; I include a variable for OBC.

So, let us say my Y is equal to some beta 1 plus some beta 2 X 2 plus beta 3 SC, beta 4 ST and beta 5 OBC. Of course, you can add any other variable you want, but just because the social group, what I have done is, because I have 4 categories, I have taken 1 reference category that is the most well-known category, that is General, and I have added all these different variables as my dummy variables into my regression equation.

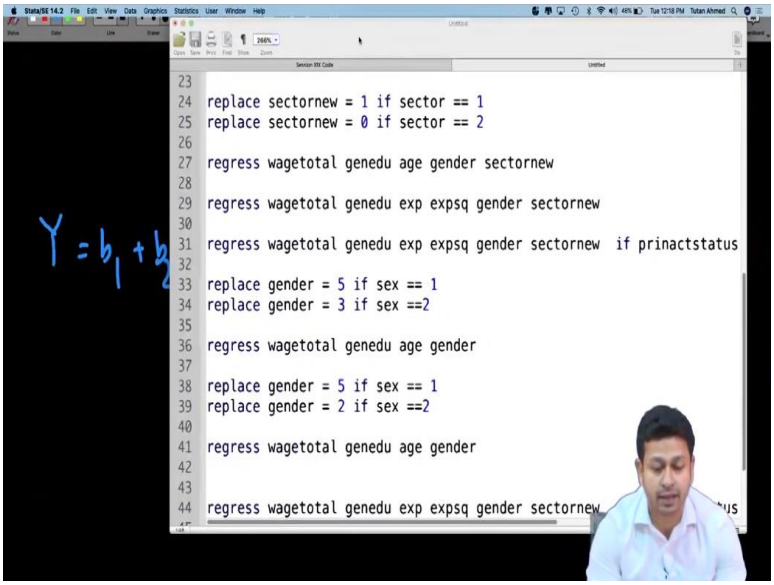
So, because there are 4 variables, so, I have taken 4 - 1 dummy variable in the regression equation; 3 dummy variables. Now, as I said, you could have like very well OBC as a reference category. And then, you could have added SC, ST and General instead of OBC. And you have to explain the regression with respect to the OBC category. So, SC with respect to OBC, ST with respect to OBC and General with respect to OBC. Now, let me actually run the regression.

(Refer Slide Time: 06:01)



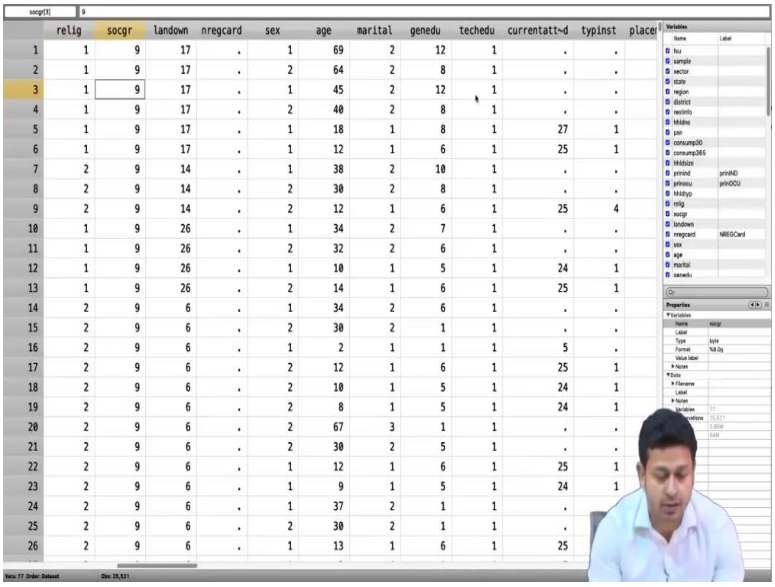
So, let us actually create the categories here.

(Refer Slide Time: 06:06)



So, in my data set, I think, how we have named it as?

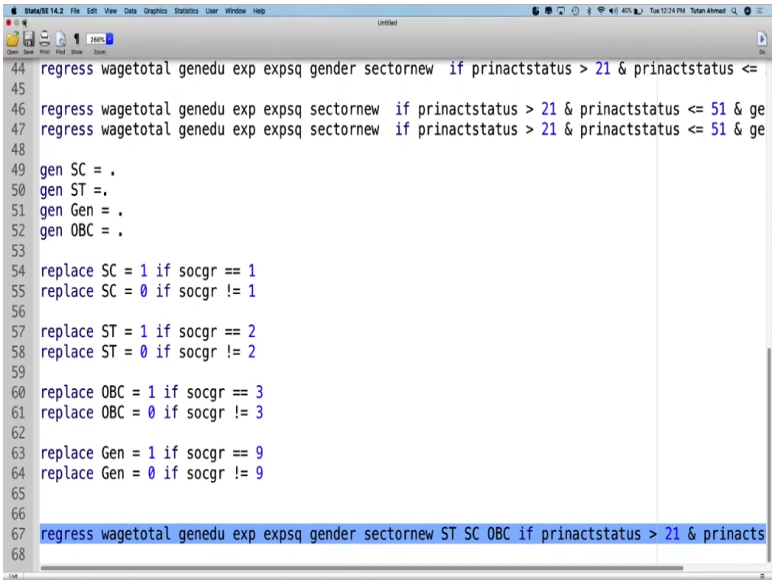
(Refer Slide Time: 06:15)



	relig	socgr	landown	nregcard	sex	age	marital	genedu	techedu	currentatt~d	typinst
1	1	9	17	.	1	69	2	12	1	.	.
2	1	9	17	.	2	64	2	8	1	.	.
3	1	9	17	.	1	45	2	12	1	.	.
4	1	9	17	.	2	40	2	8	1	.	.
5	1	9	17	.	1	18	1	8	1	27	1
6	1	9	17	.	1	12	1	6	1	25	1
7	2	9	14	.	1	38	2	10	1	.	.
8	2	9	14	.	2	30	2	8	1	.	.
9	2	9	14	.	2	12	1	6	1	25	4
10	1	9	26	.	1	34	2	7	1	.	.
11	1	9	26	.	2	32	2	6	1	.	.
12	1	9	26	.	1	10	1	5	1	24	1
13	1	9	26	.	2	14	1	6	1	25	1
14	2	9	6	.	1	34	2	6	1	.	.
15	2	9	6	.	2	30	2	1	1	.	.
16	2	9	6	.	1	2	1	1	1	5	.
17	2	9	6	.	2	12	1	6	1	25	1
18	2	9	6	.	2	10	1	5	1	24	1
19	2	9	6	.	2	8	1	5	1	24	1
20	2	9	6	.	2	67	3	1	1	.	.
21	2	9	6	.	2	30	2	5	1	.	.
22	2	9	6	.	1	12	1	6	1	25	1
23	2	9	6	.	1	9	1	5	1	24	1
24	2	9	6	.	1	37	2	1	1	.	.
25	2	9	6	.	2	30	2	1	1	.	.
26	2	9	6	.	1	13	1	6	1	25	.

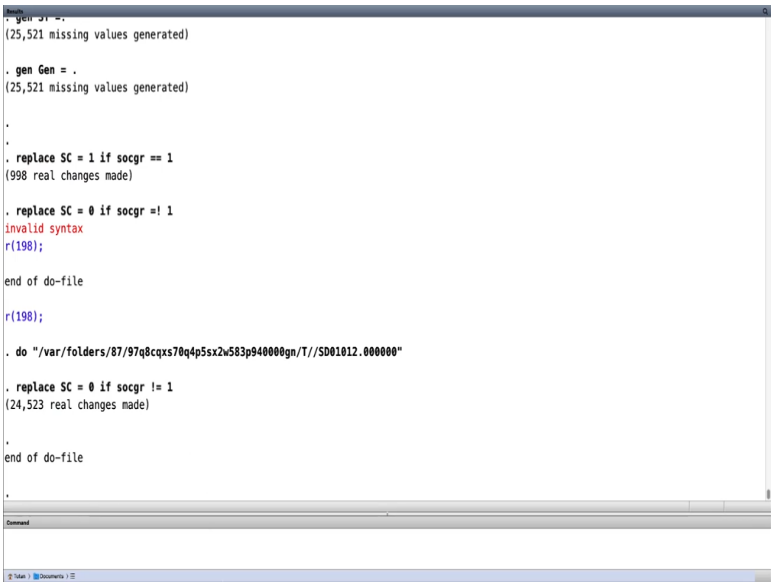
socgr is the name of the variable. So, what I am going to do is, where I am going to rename the variables; so, I am going to write it as;

(Refer Slide Time: 06:31)



So, let us say I have to generate a variable for SC; I have to generate a variable for ST; I have to generate a variable for General. And I will replace SC is equal to 1 if the social group, I think it is 1. I replace SC is equal to 0 for all other categories if social group is not equal to; I think this has to be like this. Now, it should be fine.

(Refer Slide Time: 07:26)



So, what I have done? Simply, I have; so, there are 998 people who are SC and there are so many other, 24,523 people who are not SC. So, SC is equal to 0 for all other categories; exactly what we have seen here. So, SC is equal to 0 for all other categories. So, for this person, only for if social group is equal to 1 is SC; for all other categories, is not SC. Now, similarly you can create other variables also.

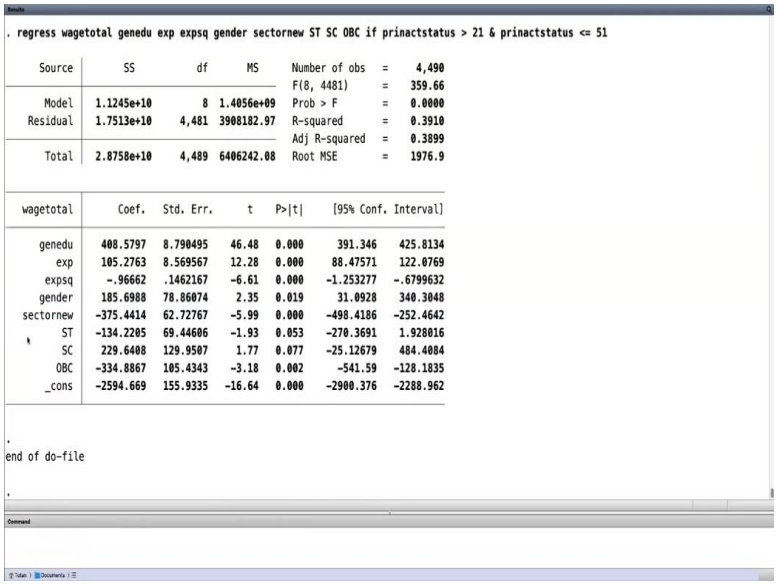
Let us say replace ST is equal to; of course, I have to create another category; this is OBC dot. ST is equal to 1 if social group is equal to, let us say 2. I think this is the values as I am assigning from my memory, but it could be like, ST could be 1 and SC would be 2. I do not recall, but it does not matter; it is just for an illustration purpose. So, if ST is equal to 1, for social group is equal to 2; so, in my original data, if social group is equal to 2 means ST.

So, I am simply replacing ST is equal to 1 when the person is really ST. And when the person is not ST, that is, social group is not equal to 2, then I am replacing the value ST is equal to 2. Similarly, I can have, replace OBC is equal to 1 if socgr is equal to 3, let us say; and replace OBC is equal to 0 if socgr is not equal to 3. And I think the General category, they categorise General as 9; General is equal to 1 if social group is equal to 9; and General is equal to 0 if social group is not equal to 9.

So, anything which is not 9 is not General. Let us say this one to do; and then there is no problem. I will get all the different values replaced here. Now, in my regression equation; so, let us say I run the regression equation for; in the previous case, we had gender and sector; so, now I want to include all these different categories. So, what I will do? I can, now I can choose what reference category I want.

Let us say, from my prior knowledge, I will be better off by defining General category as a reference category. So, what I am going to do? I am going to include all these other variables, ST, SC and OBC, all these other variables as a dummy variable in my regression equation. So, I am not including General here, but rest 3. So, let us see what we get here.

(Refer Slide Time: 10:58)



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

      Source |       SS           df       MS      Number of obs   =     4,490
-------------+----------------------------------   F(8, 4481)      =    359.66
       Model |  1.1245e+10         8  1.4056e+09   Prob > F        =    0.0000
    Residual |  1.7513e+10     4,481  3908182.97   R-squared       =    0.3910
-------------+----------------------------------   Adj R-squared   =    0.3899
       Total |  2.8758e+10     4,489  6406242.08   Root MSE        =    1976.9

------------------------------------------------------------------------------
   wagetotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      genedu |   408.5797   8.790495     46.48   0.000     391.346    425.8134
         exp |   105.2763   8.569567     12.28   0.000     88.47571   122.0769
       expsq |   -.96662    .1462167    -6.61   0.000    -1.253277   -.6799632
      gender |   185.6988   78.86074      2.35   0.019     31.0928    340.3048
   sectornew |  -375.4414   62.72767     -5.99   0.000    -498.4186   -252.4642
          ST |  -134.2205   69.44606     -1.93   0.053    -270.3691    1.928016
          SC |   229.6408   129.9507      1.77   0.077    -25.12679    484.4084
         OBC |  -334.8867   105.4343     -3.18   0.002     -541.59   -128.1835
       _cons |  -2594.669   155.9335    -16.64   0.000    -2900.376   -2288.962
------------------------------------------------------------------------------

.
end of do-file

.
```

So, if I run it, I will get value, the coefficient for all these different categories. So, what I am getting here? So, I have already explained gender and sector new; these are binary dummy; but here, I have more dummy variables ST, SC, OBC. So, what I see here is that, for ST, it is -134.2, and it is significant at 10% level, not alpha is equal to 0.1, not at alpha is equal to 5%. So, it says that a person, if that person is ST, he is likely to earn basically 134 rupees less than a person who is General.

So, just by virtue of the social group, his earning is rupees 134 less. Remember, when I am saying this, I am saying that effect of all other variables are constant, only for the change in the social group from General to ST, not other. So, basically, everything else remains constant here. If a person is SC, it is interesting, it shows that a person, if he is SC, so, he is going to earn rupees 229.6 more than a person who is General.

So, it is little counterintuitive, but this is the result. Of course, the significance level, the P value is 0.07. That is, it is significant if I have alpha is equal to 10%; but if I have alpha is equal to 5%, then it is not significant. And for OBC; now, OBC person earns rupees 334.8 less than a General candidate, General person. And this result is significant because the P value is very less.

So, that means, the OBC category actually is earning rupees 334.8 less than a General category. So, that is how we will interpret the different dummy variables for social groups. So, now, going forward, if I have to include the dummy variable for region; so, let us say

east, west, north, south and centre. Let us say I know that people in, say east, they earn lot more money than rest of the other areas.

So, if that is the case, then, what will happen is, maybe I will choose east as the reference category, and I will see, okay, how much less people in other regions are actually earning vis-a-vis the people in the eastern category. So, that is how I will choose based on my knowledge, which reference dummy I am going to choose. So, depending on how many categories you have, you always have m - 1 dummy included in the regression equation.

So, if I have like 7 categories, let us say I have 7 different religions, and I will have like 6 variables included in the regression equation. So, for India, since Hindus are the majority, so, I will have like Hindu religion as the reference dummy. And then, I will take Muslims, Buddhists, Christians, Jains, Persians and so forth in my regression equation. So, all the different religions will be represented by different dummies.

So, remember, if I have m categories, then I will include m - 1 dummy, because of that particular variable in my regression equation. So, that is how we will take into account a dummy variable which has multiple categories. Now, one very relevant question here is that, can we include all the different variables, all the different categories? Can I really include all the different categories in my regression equation?

What if I include, let us say, in case of the social group, General also into the regression equation? What if I include General, SC, ST, OBC all together? Or what if I include, say in the first case, male, female together; I have both male and female; so, if I have 2 dummy for a particular variable? So, we will see that there is a specific problem that we are are going to face, and that problem is known as the dummy variable trap.

And that is a very important problem that we need to understand. And we need to understand the concept behind dummy variable trap; why it appears? And in the next lecture, we are going to talk about this very important concept of dummy variable trap. With this, we will end this lecture here. Thank you.