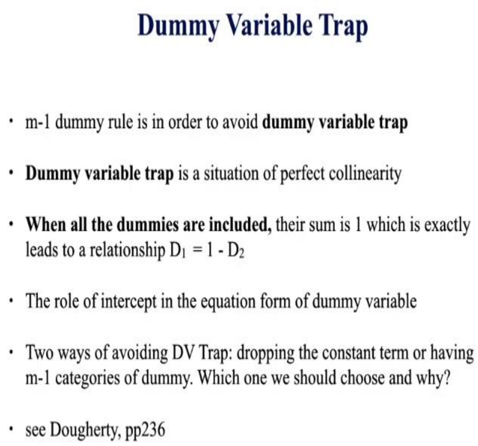
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

Module - 8
Lecture - 61
Dummy Variable (Contd.)

Hello and welcome back to the lecture on Applied Econometrics, and we are talking about dummy variable. And we have been talking about different types of cases that we may face when we are dealing with dummy variable.

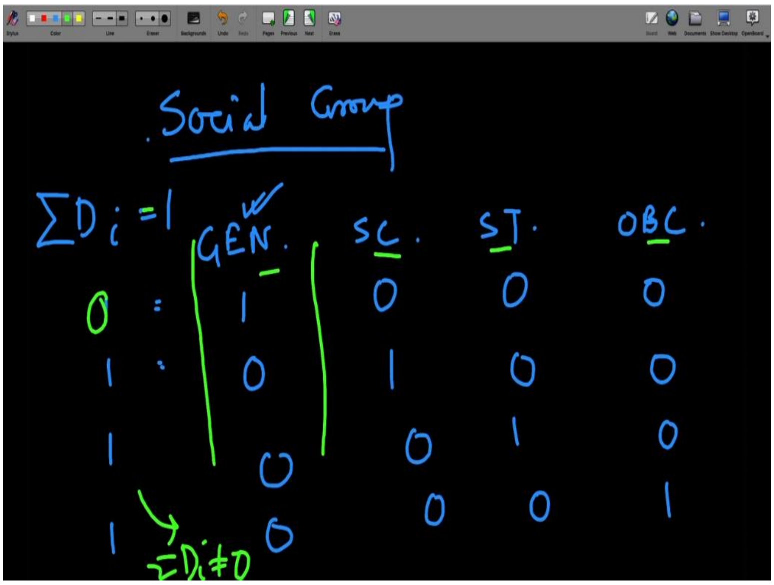
(Refer Slide Time: 00:37)



And in the previous lecture, we spoke about a possibility that if we want to include all the dummy variables in the regression equation, what is going to happen. And we said that there is a critical problem called dummy variable trap that we are going to face, and we need to understand what is dummy variable trap and why is it happening. So, let us actually try to explain that.

Now, we said that if I have m categories of a categorical variable, so, we will include m - 1 dummy variables in the regression equation to represent that particular category. Now, if I have all the categories of the dummy variable included, so, it is going to give me a situation called perfect collinearity. And the perfect collinearity will happen because there is an intercept term we have in the regression equation. I am just going to explain how these 2 things are related.

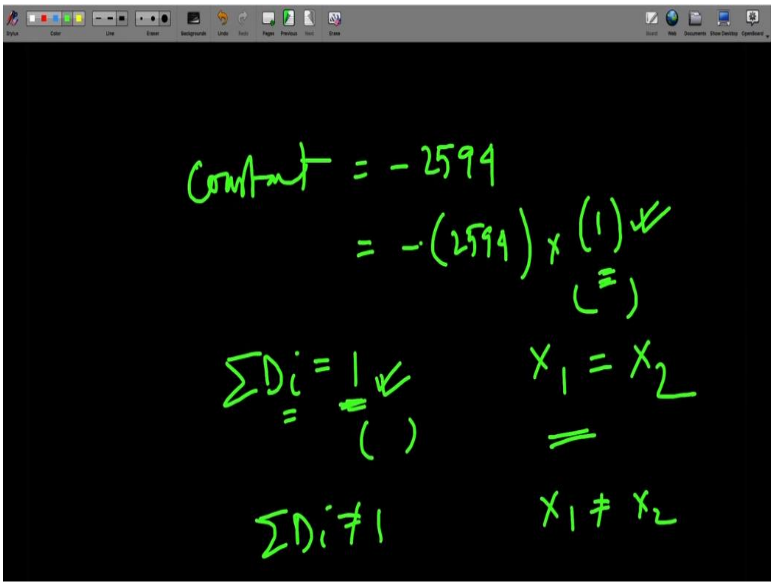
(Refer Slide Time: 01:40)



Now, let me first show you why this dummy variable trap is coming and wherefrom it is coming. So, the moment I include all these different, in this case, General, SC, ST and OBC; I have all these 4 categories and I have seen the summation of D i is always going to give me a value of 1. So, if I have all these different categories included, and if I sum them up, they are always going to give me a value of 1, value of that particular variable social group is going to be always 1, because if I include all the 4, so, these are adding up to 1.

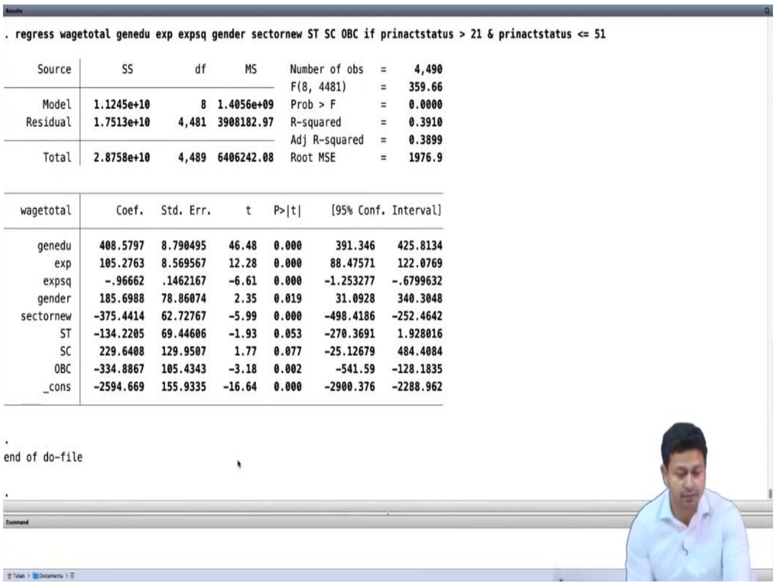
So, if you do not have the reference category, you really do not see the variations. So, if I have one category as 0 and other categories as 1, then I can see the variation from 0 to 1; but here, all I am having is a constant term 1, because I have included all the subcategories. Now, if you look at the constant term, so, constant term actually could be explained; it is a constant value.

(Refer Slide Time: 02:52)



So, in our case is, maybe -2 something, whatever value it has, it really does not matter; but yeah, -2594.

(Refer Slide Time: 03:00)



Now, if that is the constant value, so, it could actually be expressed as a; let us say this is the value of the coefficient, 2594, which is multiplied with the value of the variable which is 1. So, constant term as well; or for all the observations; if I actually think about all the different observations, so, they will have the value of the variable always 1. And if I have this summation of D i, that is also going to be 1.

So, it would mean that, when my software is trying to read the values of the variable, it will always read the value of the sum total of the dummy variable; of that particular variable, all the different dummy, if I basically take a sum total, then it is going to give me a value of 1.

And at the same time, for each of the different observation, my program will read the value of the variable is always 1.

So, it will try to understand wherefrom the variation is coming; the variation is not coming. So, it will basically see that these 2 variables are essentially same, because they are always having the same value, they are always having the value is equal to 1, for the constant term as well as for the sum total of D i. Now, because it is always same, so, it will not be able to differentiate the 2 different variables.

It will essentially think that this variable for the constant term essentially and this one, they are essentially the same variable. So, when I have, like, this is the X 1 = X 2; so, my program will say that, okay, so, these variables are perfectly collinear. So, then, I can actually explain one variable with respect to the other without any sort of deviation. But if I have in some cases, this is equal to 0 and this is equal to 1, so, then, what I will have?

I will have some variation. And X 1 is not going to be equal to X 2, if I have summation of D i not equal to 1. And that is only possible if I actually exclude a category. If I exclude a category, let us say if I exclude the General category here, and if I include only SC, ST, OBC, so, then, this value is going to be 0 and these values are going to be 1. So, in this case, summation of D i is not equal to 0 for all the cases, or is not equal to 0 for some cases.

So, that is where the variation is coming. So, essentially, we understand that, if I have this summation of D i is equal to 1, so, there it will have a perfect collinearity with a constant term. But if I allow one particular category to be a reference category and exclude that, then I include some variations here, actually I will be able to interpret the dummy variable or basically the other categories vis-a-vis the reference category.

Now, how to avoid the dummy variable trap? So, there are 2 ways to actually avoid the dummy variable trap. And the first way you can actually avoid the dummy variable trap is with the simple way where you actually can create like m - 1 categories of dummy variable; you actually choose a reference category. And that is the most convenient one, because, if you choose 1 as a reference category, you can always explain the other categories vis-a-vis the reference category; but there is another way of doing it.

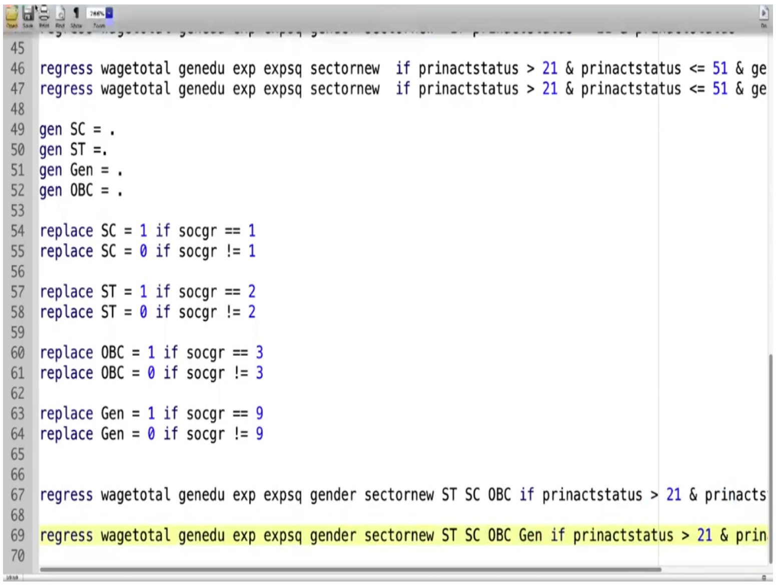
And that other way is, if you actually make the intercept term is 0. So, you can basically, in your regression equation, you can actually specify that you do not want an intercept term, and you will get a regression equation accordingly. But then, you are actually assuming that, if you do not have any intercept term, so, your regression line is passing through the origin. So, that is a big assumption you are making.

So, usually, you probably do not want to do that. But more importantly; so, what will happen if you actually make the intercept term 0? So, this perfect collinearity problem will not arise and you will see that you are actually able to run the regression equation with m dummies. If you have all the m dummies included, there is no problem; it will still run. But there is a problem in the sense that when I have all the m dummies included, how do I really interpret the values of the dummy variable with respect to each other?

Because, a dummy variable is always explained as a, is a relative concept; it is explained in terms of, with respect to something, but if you do not have any reference category, so, usually, it is very difficult to explain also. So, because of that, we normally do not go for the route of making intercept term 0, but rather what we do is, we include all these different categories of dummy with keeping one category as a reference category.

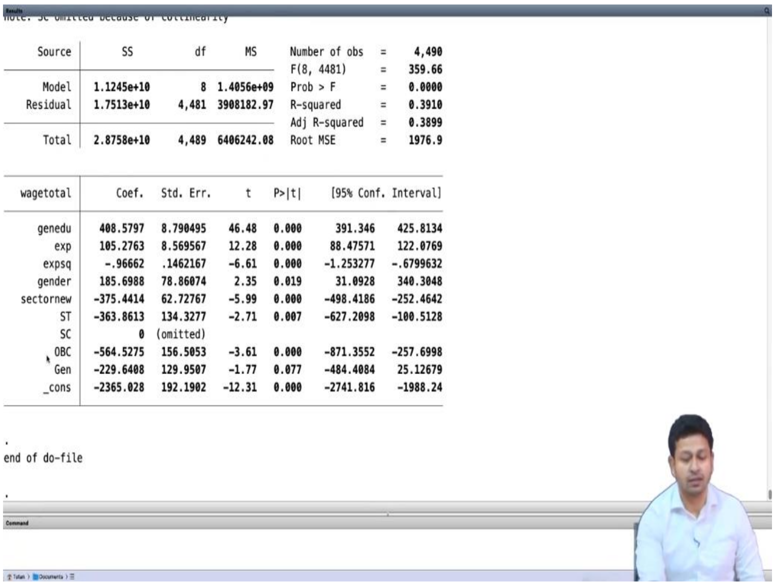
So, this is how we essentially address the dummy variable trap. Now, suppose in our regression equation, we include all the dummy variables, so, what will happen then? Let us see. Stata is a very smart software, and let us see what will happen if I include all the different dummy variables here. So, here I had SC, ST, OBC; and let me actually include in the regression equation, let me also include General.

(Refer Slide Time: 08:38)



And then, let us see what happens here. So, the same equation I am going to copy and I am going to paste also. So, what I will do is, I will also include General. And we will see what happens here if I have all the 4 categories included. Now, it will definitely give me a collinearity problem; SC, ST, OBC and General; but sometimes what will happen, you will see your Stata or any other software is actually giving you a error message, but here, Stata is really smart, and what it will do is, it will automatically choose a reference category and it will omit it.

(Refer Slide Time: 09:12)



And it will say that SC is omitted because of collinearity, the problem that we just said; because, when you add all these 4, SC, ST, OBC, General, the value of the variable is becoming 1; and when it is equal to 1, it is becoming perfectly collinear with the constant term. It could be interpreted as the multiplication of this coefficient into a value of the variable, which is equal to 1.

So, that is why my software has already omitted a particular category which is SC, and it will choose at its own convenience. So, it has decided to have SC as a reference category and it is omitted. And then, when it is omitted, then you can again explain the impact of other, categories vis-a-vis the SC category. So, this is how we should understand the dummy variable trap concept. So, you can actually look, and it is in the chapter 5.

And with this, we will end the lecture on dummy variable trap. And in the next couple of lectures, we are actually going to see the concept of reduced form regression; we have already kind of touched upon that; reduced form equations when we talk about dummy variable. And we are also going to see the other type of dummy variable, that is slope dummy variable in the next couple of lectures. Thank you.