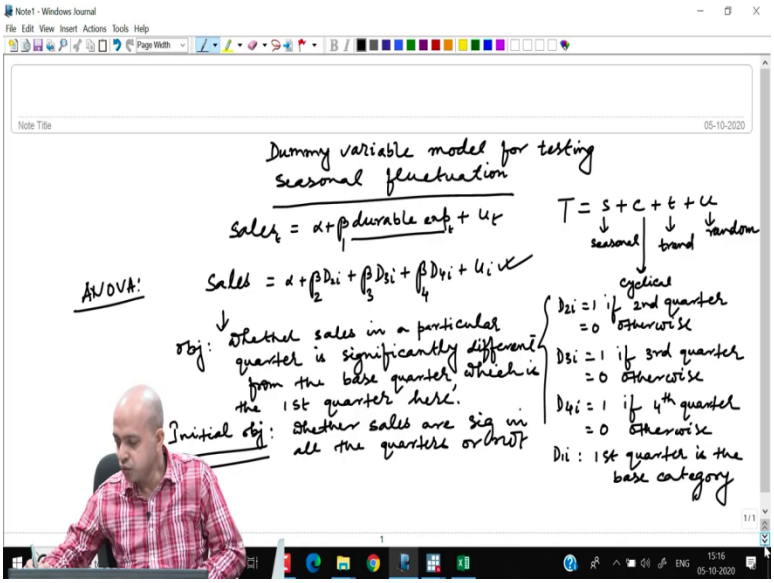
Statistical Analysis of Dummy Variable Models and Testing for Seasonal Fluctuations
Part-5
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 42

Welcome once again to our discussion on dummy variable models basically we were discussing about the applications of dummy variable models. So, in our last class if you recall then what we were doing? We were trying to understand basically whether a dummy variable model can be applied in a context where our variable, so firstly we discussed about on the saving income relationship.

And then later on we talked about how to use that saving income relationship in a dummy variable context. That means to test the structural break how the dummy variable model can be an alternative of the Chow test that we learned earlier. And we discussed that dummy variable model can actually overcome the several limitations of Chow test model.

(Refer Slide Time: 1:56)



Today again we will discuss another application of dummy variable model. And this application is basically to test whether there is any seasonal fluctuation in a time series data. So dummy variable model for testing seasonal fluctuation.

Now many times what we see that the time series variable give some kind of fluctuation based on which particular season we are observing that particular variable. For example, if you think

about the sales data let us say we have sales data on refrigerator, sales data on refrigerator and other house hold appliances, let us say washing machine, dish washer then refrigerator so on and so forth.

And the many a times what we observe that the sales of these house hold appliances do show some kind of seasonal fluctuation. For example, sales of refrigerator will be quite high during the summer season and quite low in winter. Similarly, the home appliances like washing machine, T.V so on and so forth, you might be seeing that the sales of these appliances are quite large during the Diwali season.

So, if you have let us say quarterly data that means in a year basically you have four quarters. And your objective here is to see which particular quarter in a year gives you the significant sale. Whether all the quarter sales in all the quarters are significant or it is happening in some of the quarters.

And based on that as a salesman, as a sales person you may devise some kind of policies, that is our objective. So, we have data on sales for these household appliances and you want to see whether sales in in a particular quarter is significant. Now why do we do so? This is very important for time series data.

For example, here if you look at the sales basically sales is actually a function of let us say alpha plus beta 1 is the expenditure on durables. Let us say this is my durable expenditure plus this is let us say t ut. This is sales which is sales on household appliances which is a function of the expenditure on durables.

And if you try to run this type of regression then what will happen? Firstly, what you need to do, you need to remove some kind of seasonal fluctuations from this data. Why this is so? Because a time series data t is actually a summation of three types of fluctuations.

What are those? Firstly, a time series data will show you some kind of seasonal fluctuation. So, let us say that is called seasonal, then second is cyclical, then third is t and the fourth is the random one. So, this is seasonal, this is cyclical, then this is trend any time series data will show you some kind of trend over a period of time and this is purely random.

Now this random component what you are already adding in the model you see there is ut but s, c and t the seasonal, cyclical, and trend this variable should be removed from the time series variables. For example, here sales or durable expenditure otherwise you may get a spurious regression in time series data. And there are specific ways by which you can remove the cyclical and the trend component of the time series.

In this example we are basically interested in how to remove the seasonal component from the time series data. So, if you do so sales is the function of durable expenditure that means to remove the seasonal component from the data, we need to consider that these four quarters are actually different. And that is why while modeling, we need to accommodate the fact that these quarters are actually different. If you do not consider that fact then your model will suffer from the seasonal fluctuation.

So, what we need to do here? Here let us say for the time being I am removing the durable expenditure part, this variable so we basically will start our analysis with a simple ANOVA model, simple ANOVA model. So, as you know that ANOVA model will not consider any quantitative covariate as explanatory variable. So, in that sales is a function of let us say you have four quarters, then you have alpha plus let us say beta, beta 2 D2i plus beta 3 D3i plus beta 4 D4i plus ui.

And how we defined D2, D3 and D4? So D2i equals to 1 if second quarter and 0 otherwise. D3i equals to 1 if third quarter and 0 otherwise; and D4i equals to 1 if fourth quarter and 0 otherwise. This is how we have defined the dummy. And we have not defined what is our base category here. Since we have not assigned any dummy for the first quarter that means D1i is a base category.

So that means all this beta 2, beta 3, and beta 4 they are differential intercept. So that means they will basically tell you that whether the sales in second quarter is significantly different from first quarter. D3 will tell whether the sales in the third quarter is significantly different from the first

quarter. And D4 will be beta 4 will tell whether the sales in fourth quarter is significantly different from the first quarter.

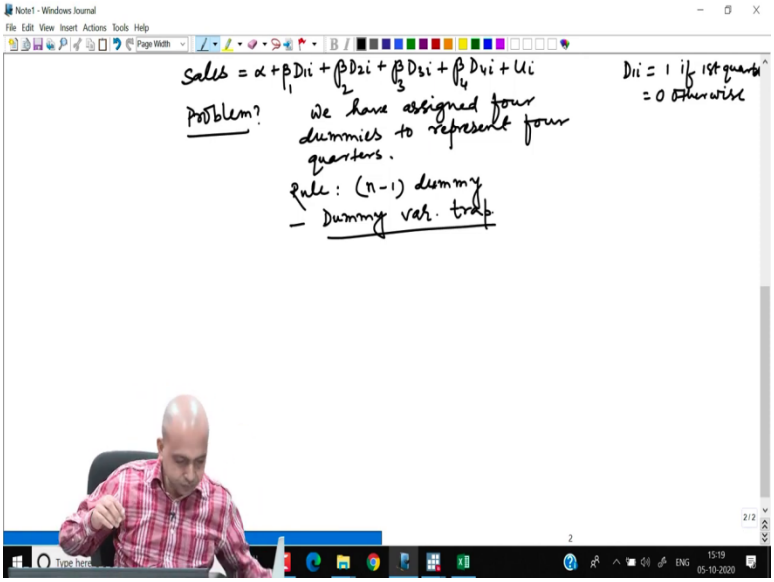
But see our objective initially when we pose the problem our objective was not to examine whether sales in a particular quarter is different from any of the other quarters any with the base quarter which is first quarter here. Rather we were interested in out of these four quarters which particular quarter gives you the significantly higher sale. It might be second quarter, third quarter, fourth quarter, first quarter or all the quarters together.

Are you getting my problem? Our initial research question was whether the sales in refrigerator or likewise any other household appliances, sales in household appliances are significant in all the quarters or in one or two quarters. So basically, we were interested in only seasonal presence of seasonal fluctuation on the sales of the durables. That is our objective.

If that is the case if you said this type of model, dummy variable model, then your objective is not fulfilled because in this model they will only tell you whether the sales in a particular quarter is different from the base quarter. So, what is this model tell you? This model's objective is whether sales in a particular quarter is significantly different from the base quarter which is the first quarter here.

But our initial objective, what was our initial objective? Our initial objective was to see whether sales are significant in all the quarters or not, all the quarters or not. So, we were not interested in checking the significance of sales in a particular quarter with the base. Now if this is your objective then you cannot set this type of model.

(Refer Slide Time: 15:58)

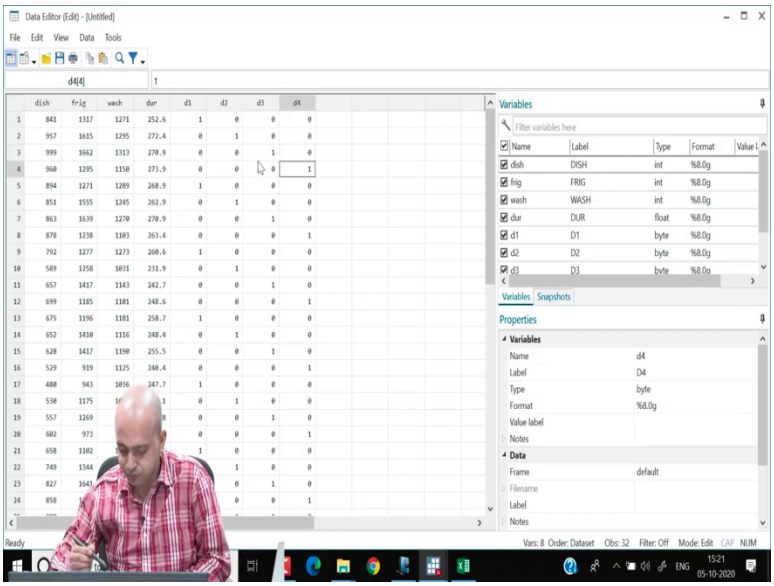


So, what type of model then you have to set? You have to basically set this type of model. Sales equals to alpha plus beta1 D1i plus beta2 D2i plus beta3 D3i plus beta4 D4i plus ui. And now beta D2, D3, D4 we have already defined. And D1i is now equals to 1 if first quarter and 0 otherwise. So, this model should be specified. But if you specify this model what would be the problem in estimation? What would be the problem in estimation? Can think of? If I set this type of model what would be our problem in estimation?

The problem is, see here we have assigned four dummies to represent four quarters. But what is the rule says? The rule says that there should be n-1 dummy, that means if you have n categories, here it is a four categories, four quarters so you should have introduced four minus one, three dummies.

Now our problem is if we assign three dummies then that model will tell you what would be the interpretation of the dummy variable coefficient. They will only tell you whether sales in a particular quarter is significantly different from base quarter. So that cannot serve our initial purpose and if we fit this model to serve our initial purpose that means to see whether sales in all the quarters are significant. Then we will end up with dummy variable trap.

(Refer Slide Time: 19:21)

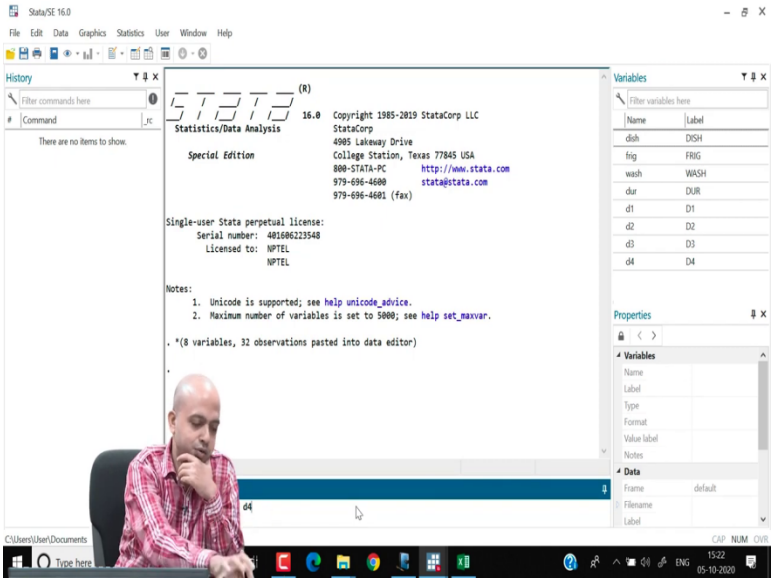


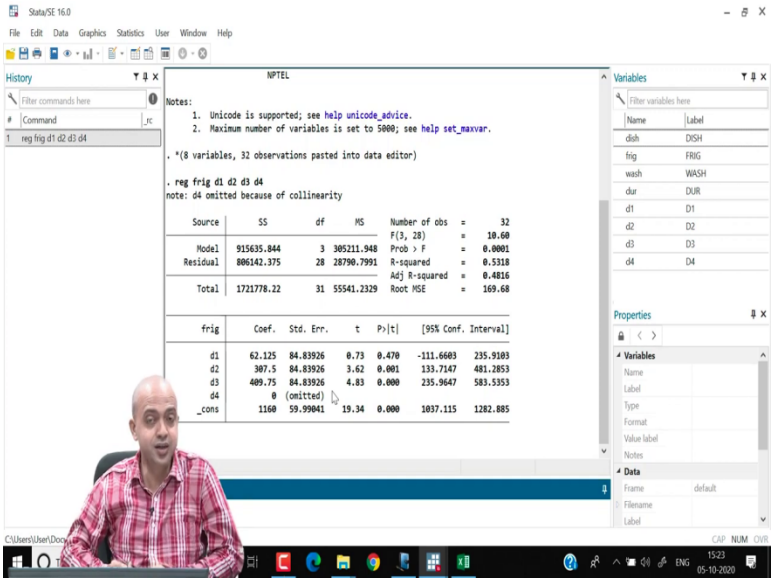
Now what is the exact problem of dummy variable trap? That first we will try to understand using the data. So, look at the data set. This is our data. So, we have quarterly data. So, in each year we have four data points. And look at, see here dish means dishwasher sale 841 all in thousands. This data is also taken from the United States and this is taken from your textbook Gujrati that basic econometric textbook and this is table number 9.3. This data is from table 9.3.

So, this is the sales data of refrigerator which is represented as frig, and then this is washing machine and this is actually quantitative covariate which shows the total expenditure on the durables. And then what we did? We have introduced four dummies. So, this data is from first quarter that is why D1 equals to 1 and see all other quarters are 0. D2, D3, D4 are 0.

Similarly, the second data point is from the second quarter that is why first, third and fourth quarter dummies are 0. This data is from the third quarter that is why first, second, and fourth quarter they take the value 0. And similarly, the dummy for the fourth quarter is 1 and that is why first, second, and third quarter take the value 0. So here we have assigned four dummies for four quarters. And let us see if we do so that means if we violate the rule of assigning dummy what type of estimation problem, we may end up due to dummy variable trap.

(Refer Slide Time: 21:24)



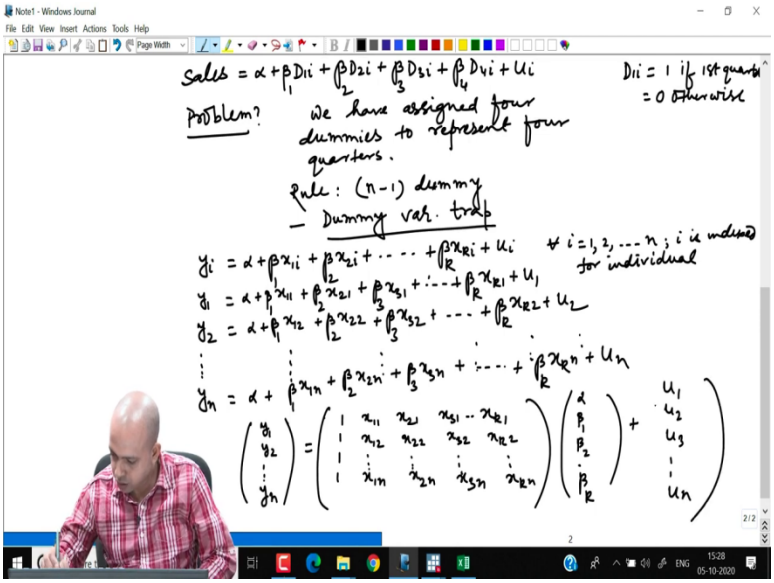


Let us see we will try to estimate. So, what we will do? We will regress reg let us say frig means sales in refrigerator which is a function of D1, D2, D3, and D4, D1, D2, D3 and D4 four dummies we have introduced. And then enter. Now look at this, look at this Stata has reported that D4 omitted because of collinearity. And here also in the, in the, in the output set we see the coefficient corresponding to D4 is 0 and it is written in the bracket omitted, omitted.

Now what is the problem that means we have assigned four dummies but while reporting the result Stata is basically omitting one dummy here, we have omitted the fourth dummies. Now why Stata is dropping one dummy that means is there any problem in estimation. If we introduce

four dummies for four quarters to understand that we have to once again go back to the way, we have represented our multiple linear regression model using that matrix Algebra. That will make things clear. So, you will know what exactly is the problem of dummy variable trap.

(Refer Slide Time: 23:07)



So, for that we will go back to our model. So, let us say that this was yi equals to alpha plus beta 1 x1i plus beta2 x2i plus dot dot dot beta k xki plus ui and where i is basically equals to 1 to n. So that means this equation is for i^{th} individual, i is indexed for individual.

So, what will happen? You will see y1 if I substitute the values of i for 1, 2, 3, n what will happen? This would become alpha plus beta1 x11 plus beta 2 then you will get x21 plus beta 3 x31 plus dot dot betak xk1 plus u1. Similarly, y2 would be alpha plus beta1 then x12 plus beta2 x22 plus beta3 x32 plus beta k xk2 plus u2 this will happen.

So that means ultimately you will have if you represent this, what will happen? yn would become alpha plus beta2 x1n plus beta sorry this is beta 1, this is beta 1, so this is beta 1, beta 1 x1n plus beta 2 x2n plus beta 3 x3n plus beta k xkn plus un for the nth. So, in matrix notation how we represented this?

We represented as y1 y2 dot dot dot yn equals to what we did? So, we will take this alpha as common. So, this will become 1 1 1 1 and then we will have x x11 x21 x31 and then xk1 then you have x12 dot dot dot x1n x21 sorry this is x22, x22 then x2n then x32 then you have x3n and

xk2 then xkn. And here you will have alpha and then beta 1, beta 2 dot dot dot beta k, beta k plus you have u1, u2, u3 dot dot dot un. So, this we have discussed.