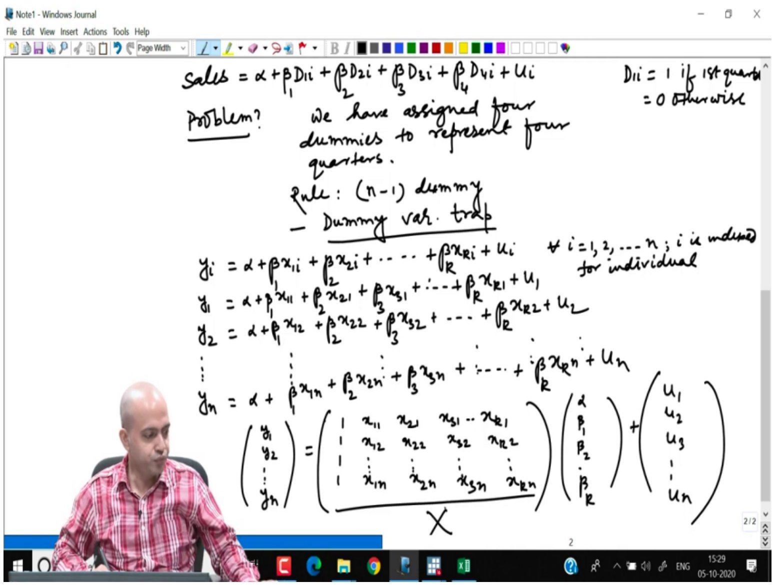
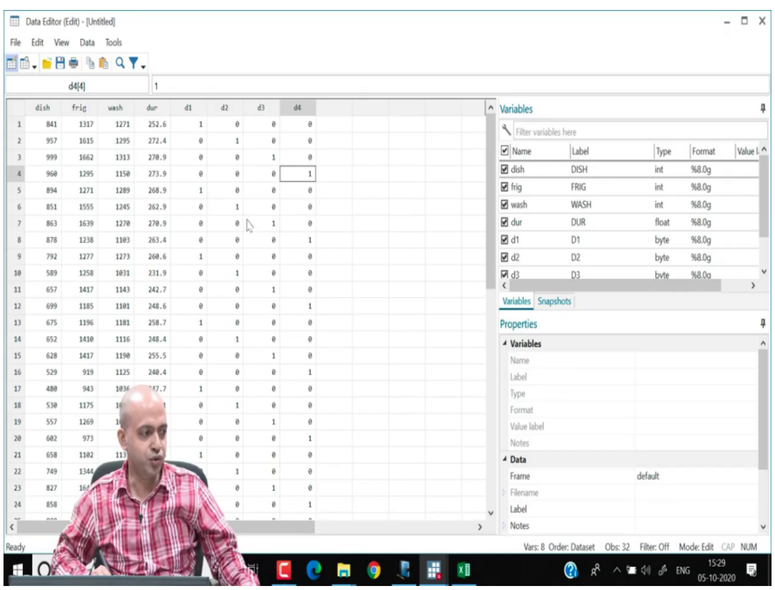
Statistical Analysis of Dummy Variable Models and Testing for Seasonal Fluctuations
Part-6
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Lecture 43
Indian Institute of Technology, Madras

(Refer Slide Time: 0:15)

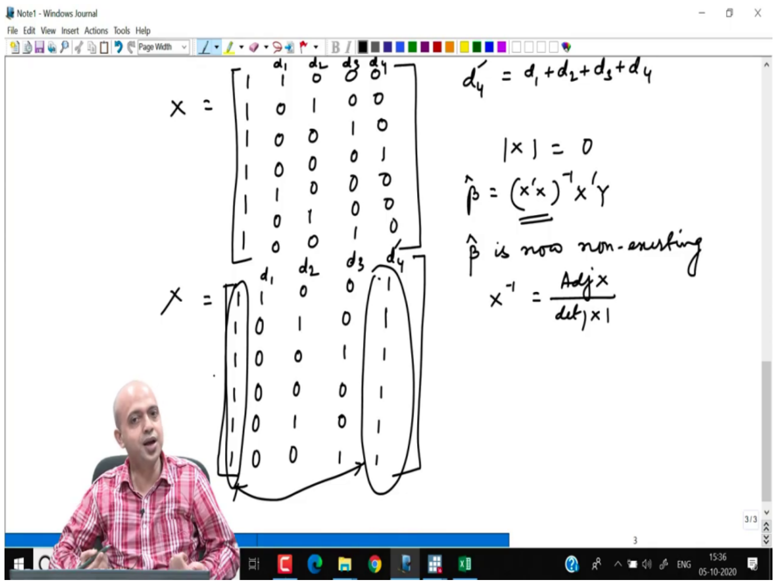


(Refer Slide Time: 0:37)



Now look at carefully the structure of the X matrix. Please remember this is my X matrix. This is X matrix. So, in the X matrix the first column is all 1 1 1 1, all 1 1 1 1. Now what I will do? I will ask you to think this X matrix in terms of this. So that means in this model I have no other quantitative covariates but only this, only this. So that means I will write the X matrix only in terms of the coefficient alpha that means intercept and this part d1, d2, d3, and d4. So, I will take this X matrix separately.

(Refer Slide Time: 1:05)



So, X is now, how is your X looks like so this is 1 1 1 1 1 1. How many 1's are there? n number of 1's. And then you have this matrix, this dummy variables coefficient you have to take, look at this. So, this is now 1 0 0 0. So that means this will look like 1 0 0 0, and then 0 1 0 0, then 0 0 1 0, and 0 0 0 1. Again, you will get 1 0 0 0, 0 1 0 0, then 0 0 1 0 likewise you will get. So, this is for d1, this is for d2, this is for d3, and this is for d4. This would be your X matrix.

Now what we can do? If you go back to your (matrix) basic sub matrix Algebra any of these columns can be transformed by a simple linear transformation. For example, let us transform this d4 column as d1 plus d2 plus d3 plus d4. This is how I am transforming this fourth column. So, if you do so what will happen?

Your matrix will change like this, this is 1 1 1 1 1 1, and then d1 would be 1 0 0 0 0 0, d2 would be 0 1 0 0 then 1 0, then d3 would be 0 0 1 0 0 1, and then d4 would be now 1 1 1 1 1 equals now I have changed the d4 into d1 plus d2 plus d3 plus d4. And if you take the summation then it will become 1 1 1. It will become 1 1 1, is that fine?

Now look at this column the changed d4 and these columns are actually identical. That means the elements of the first column which represent the intercept and elements of the transform fourth column they are similar. And if that is the case what would be the, what would be the value of this of the determinants of X? So, determinants of X what would be the value? So, if you know the determinant of this value would be 0. Determinants of this would be 0.

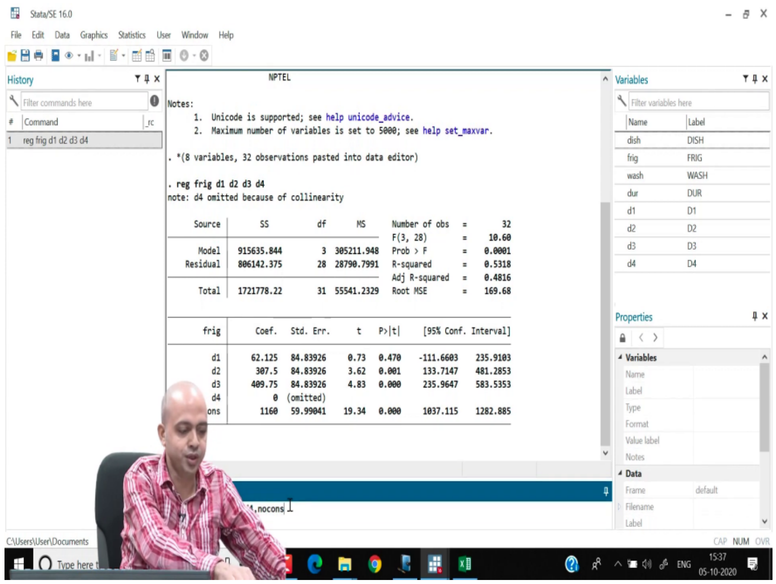
And now if the determinants becomes 0 then how we have defined our beta hat? See X prime X inverse X prime Y, for this two exist this matrix should actually exist the inverse. Now the moment this become 0. So, beta hat, beta hat is now non existing. what is the X inverse? X inverse equals to adjoint X by determinants, determinants X and because this rule now determinants X is actually 0.

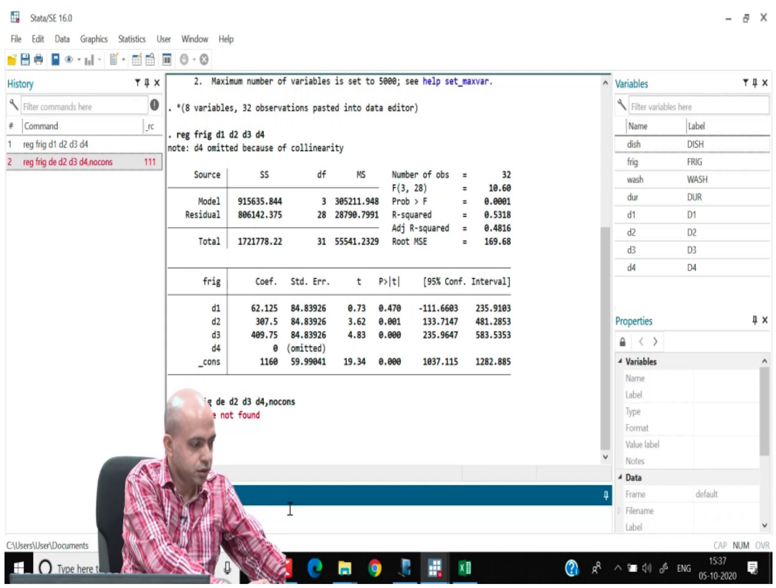
So that means instead of the same rule here we are talking about X actually the we will find the matrix of this X prime X is also does not exist because of this problem. So that means if you want to understand the actual problem of this dummy variable trap, we need to think about your X matrix. The structure of X matrix. You have the first column which represent the alpha that means intercept and since you have introduced four dummies for four categories after transformation of d4 which is d1 plus d2 plus d3 plus d4 all the elements of this transform d4 is actually, is actually 1 1 1 1 1.

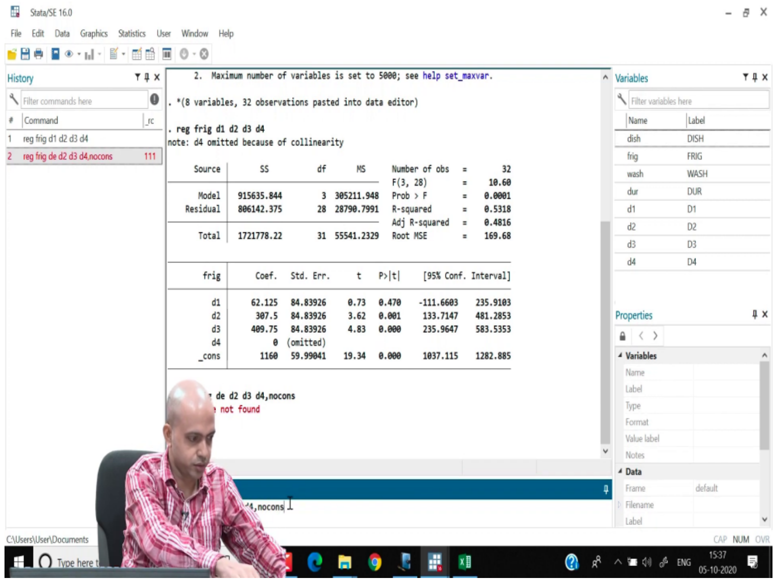
So that means there is clear dependent that mean these two columns are collinear. That is why in Stata they said that due to collinearity problem, the fourth quarter is actually dropped. Now we are in trouble. We need four dummies to be introduced in the model but the dummy variable rule says you will end you with this type of problem. So how will you solve that problem?

We solve that problem that means from the structure of the X matrix itself you can understand where there is a simple mathematical solution about this. What is the solution? Solution is you just omit this, this intercept column. So that means if you omit this intercept column then the problem is not there. That means the moment you introduce four dummies for four categories while estimating the model you need to ask Stata that I do not want intercept to be included in my model. And if you remove the intercept then absolutely no problem in estimation.

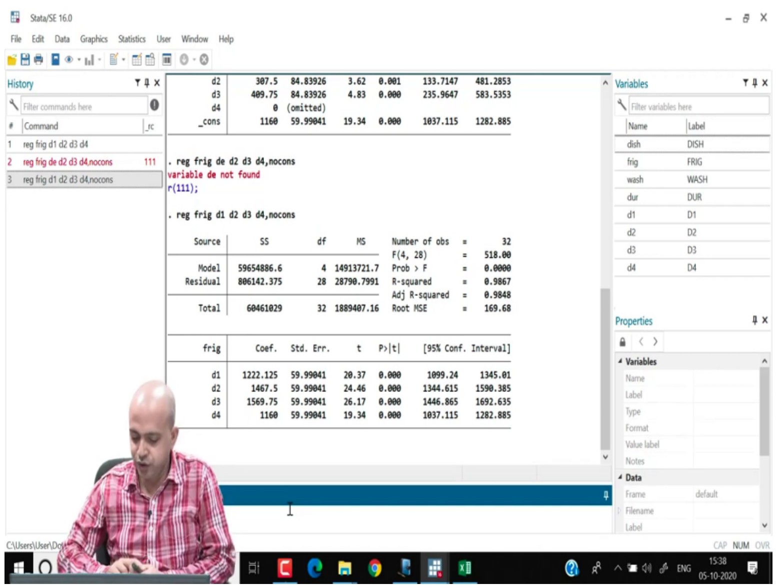
(Refer Slide Time: 8:17)







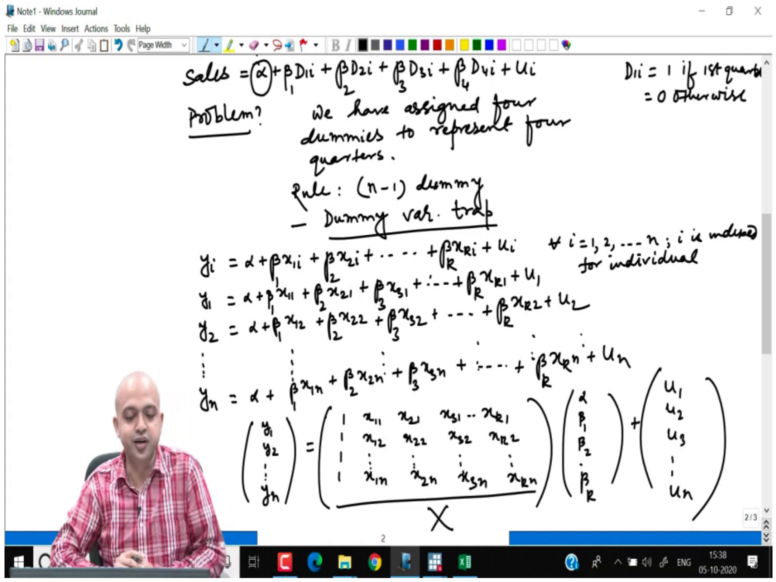
(Refer Slide Time: 9:00)



So that means what you have to do? If you want to estimate this type of model. Then the solution is what you do, you reg frig and then d1, then d2, then d3, d4. You introduce four dummies but you mention that I do not want any constant term to be included. Sorry, oh sorry this is reg then frig d1 actually d2 d3 d4 and then no constant.

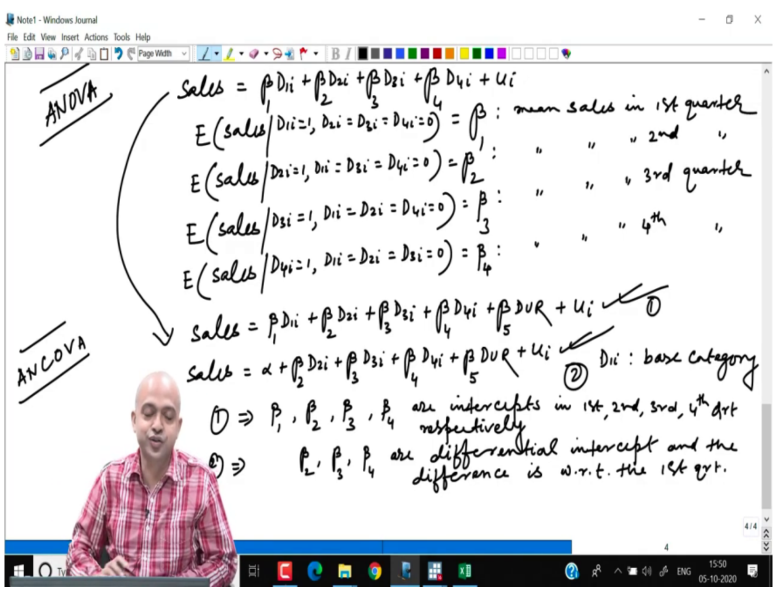
Now look at what is happening here? You have solved. Now Stata is not saying that 4 is omitted, all the coefficients you are getting. You have solved the problem by introducing four dummies. So basically, by omitting the elements of the columns first column you have so overcome the problem of dummy variable trap.

(Refer Slide Time: 9:51)



So, what would be the, what would be the interpretation of this model now? So that means in models where you have four dummies for four quarters. And no other covariates included then as I said when you have this type of model, when you have this type of model four dummies you have to omit the intercept term alpha.

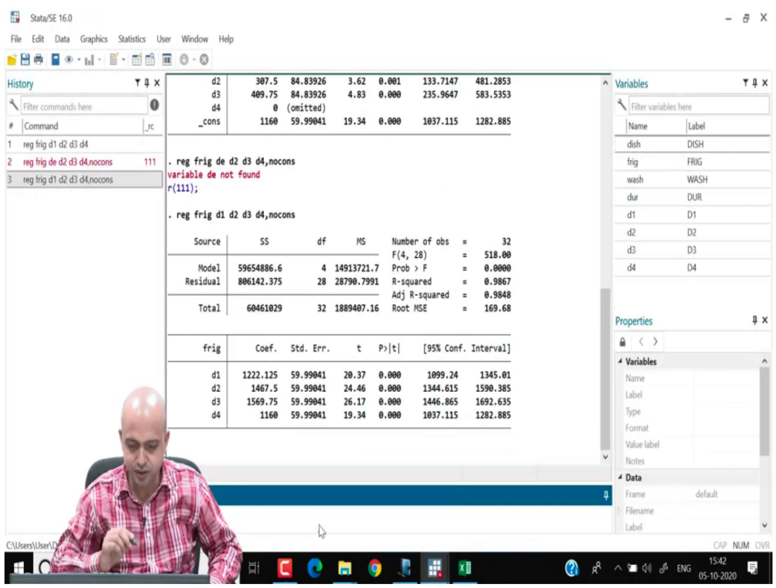
(Refer Slide Time: 10:40)

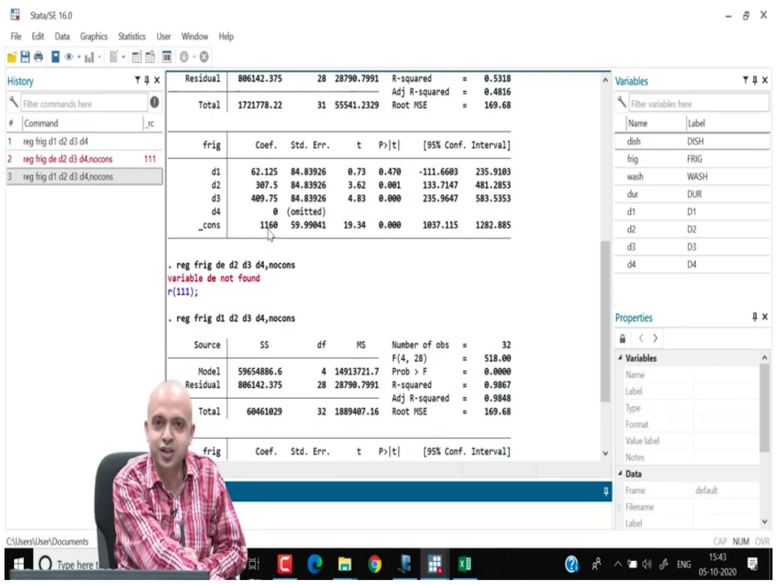


And if you do so? Then what would be the interpretation of beta 1, beta 2, beta 3, beta 4? I will show you the interpretation. So now, now you have sales equals to, sales equals to beta 1 D1i plus beta 2 D2i plus beta 3 D3i plus beta4 D4i plus ui, ui. And now if you take expectation of sales given D1i equals to 1 and D2i equals to D3i equals to D4i equals to 0 then you will get beta 1. So that means beta 1 basically indicates, beta 1 basically indicates sales in rather mean sales, mean sales in first quarter.

Similarly, expectation of sales given D2i equals to 1 and D1i equals to D3i equals to D4i equals to 0 indicates beta 2. So that means beta 2 basically mean sales in second quarter. Similarly, expectation of sales given D3i equals to 1 and D1i equals to D2i equals to D4i equals to 0. This is beta 3. And then beta 3 that means indicate mean sales in third quarter. And expectation of sales when D4i equals to 1 and D1i equals to D2i equals to D3i equals to 0. So, beta 4 mean sales in fourth quarter. So that means coefficient of this dummy themselves indicate the mean sales.

(Refer Slide Time: 13:46)





So mean sales of the first quarter is 1222, second quarter it is 1467, third quarter is 1569, and fourth quarter is 1160. And if you look at the significance of these coefficients all are significant at 1 percent level. Because the p value is 0.000 if you multiple that by 100 still it is 0.0. So, they are directly giving the sales in each quarter and all the quarters, sales in each quarter is actually significant. So, no quarter should be ignored by the salesperson. So, all the quarters are significant. So, this is the ANOVA model we have introduced, we have introduced.

And if you look at the earlier model basically now if you compare see d4 was omitted earlier so that means when d4 is omitted by Stata, Stata is basically considering fourth quarter as the base quarter. That means the constant term here 1160 indicates sales of the fourth quarter. When you introduced four dummies with intercept, Stata drop the fourth dummy that means fourth quarter and considered fourth quarter as the base quarter. And when fourth quarter is the base the constant term obviously represent the sales of the fourth quarter.

Now in this case when you have four, see this is exactly 1160 that is also 1160. And look at its standard error 59.9941 exactly matching. So that means whether you set this model or that model, that model the coefficients are coming actually the same. Only thing is that there you have to consider the fourth quarter as base since it is omitted.

Now what we will do? In this model we have introduced only qualitative information that means it was ANOVA model. So, if you consider the quantitative covariate which was sales in durables then your model will change like in this model, I have introduced only four dummies d1 d2 d3 d4. So, this is basically the ANOVA model.

Now what I will do? I will make this model ANCOVA by introducing sales equals to let say beta 1 D1i plus beta2 D2i plus beta3 D3i plus beta4 D4i plus durables sorry beta5 Dur that is the expenditure on durables plus ui. You can set either this type of model four dummies or if your objective is to consider one category as base and include dummy then you can introduce this type of model also alpha plus beta 2 D2i plus beta 3 D3i plus beta 4 D4i plus beta 5 durables plus ui.

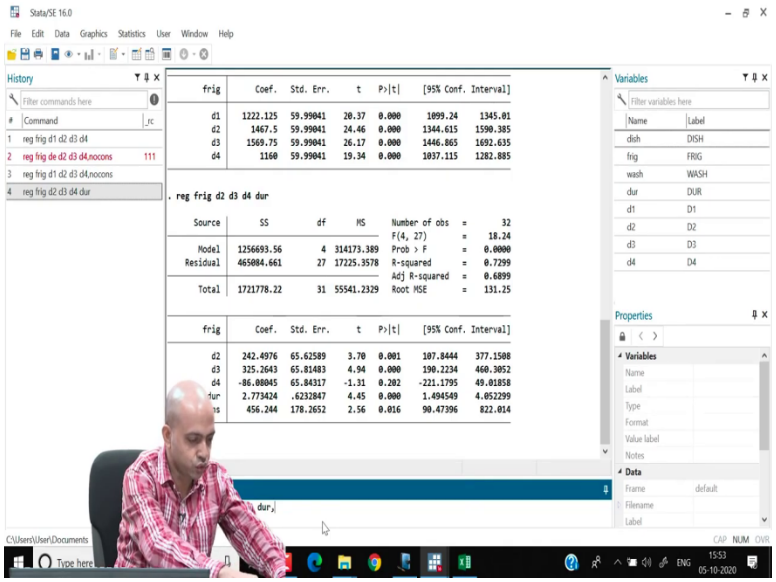
So here since we have already introduced the intercept, we consider D1i actually base category. And all these beta 2, beta 3, beta 4, and beta 5 they will indicate differential intercept, whether sales in other quarters are significantly different from the first quarter or not that we will know by the significance of beta 2, beta 3, beta 4, and beta 5.

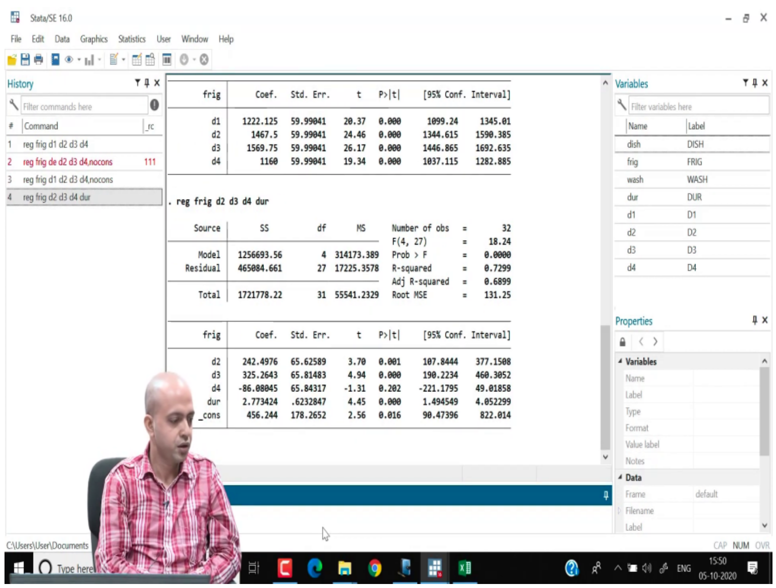
So that whether you will set this model or that model that again depends on the researcher's objective. If you want to compare other quarters with the base category, then this is the model-second model. If you want to simply test whether sales in all the quarters are significant, then you have to set the first model not the second model. Because second like the standard dummy variable model will always give you the differential intercept not the actual one.

So, you have to clearly keep in mind. So, while model one gives you in model one beta 1, beta 2, beta 3 and beta 4 they are giving direct intercept for first quarter, second quarter, third quarter and fourth quarter. So, in model one beta 1, beta 2, beta 3 and beta 4 are intercepts, intercepts in first, second, third fourth quarter respectively.

But in model two, beta 1 beta 2, sorry beta 1 is not there model 2 beta 2 beta 3 beta 4 are differential intercept and the difference is with respect to the first quarter, this you have to understand. Four dummies means they directly give you intercept, three dummies means it is always in comparison with the first quarter that is why differential intercept. And since we have introduced the other covariate, it is called the ANCOVA model.

(Refer Slide Time: 21:34)





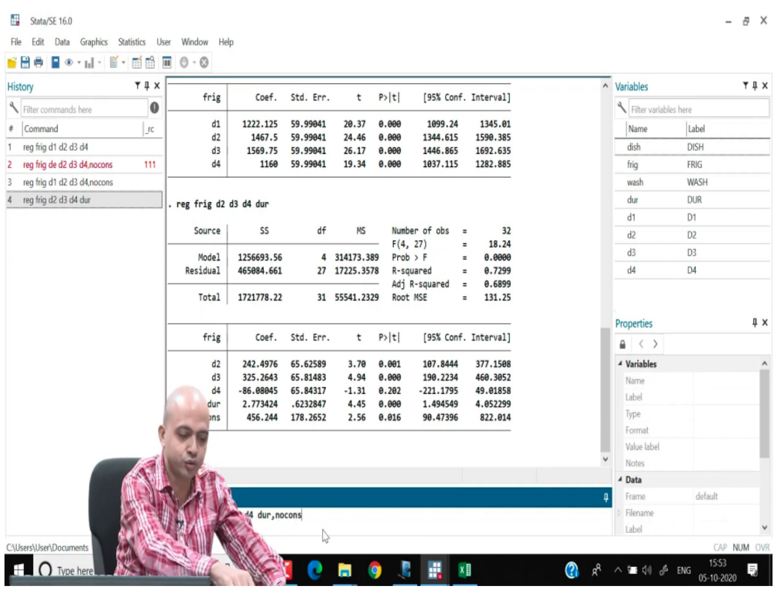
Now if you do so, if you estimate the ANCOVA model taking one as base category then what we will do? Let us say reg frig then let us say D2 D3 D4 and durables, durables we have taken and let us now this, so what you will know now? See in this ANOVA model now it shows that fourth quarter sales, sales in fourth quarter is not significantly different from the first quarter that would be your interpretation.

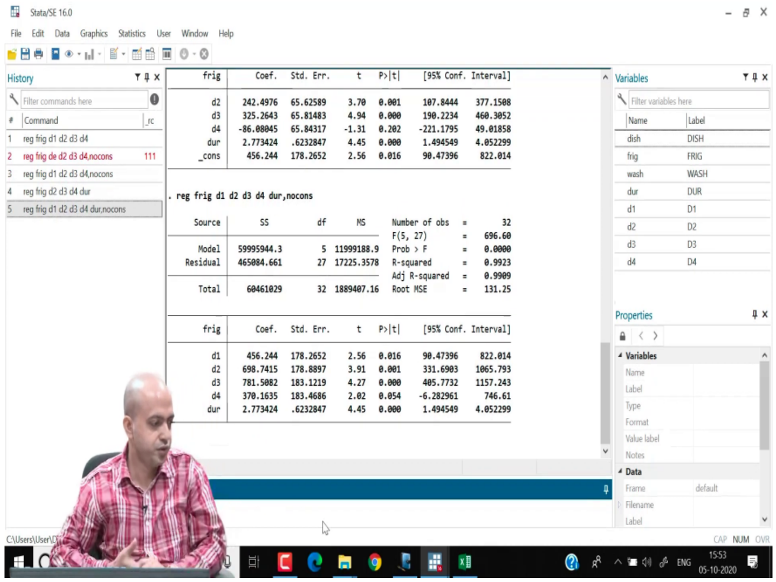
Because the corresponding p value is 0.202 if you multiply that by 100 it would become 20. So that means you are committing 20 type one errors while rejecting the null, so that means it is not significant. But D2 and D3 they are significant at 1 percent level. So that means we can say that we can say the sales in second and fourth quarter second third, second and third quarters are significantly different from the first quarter but sales in fourth quarter is not significantly different from the first quarter.

And it also shows that the impact of the quantitative covariate which is expenditure on durables is highly significant for every additional unit expenditure on durables your sales in refrigerator increases by 2.77 unit, 2.77 unit. And one more thing you have to keep in mind in this model we have not interacted the dummy variables with the quantitative covariate. What does it mean? It means that the responsiveness of sales with respect to durables are same across different quarters, across different quarters.

One durable expenditure on durable increases by 1 unit, sale will on an average increase by 2.77 unit for each quarter, for each quarter. If you believe that different quarter will respond differently for this additional expenditure then you have to interact the seasonal dummies with this covariate. But that we are not doing it here assuming they are same. We are not complicating the model. But we have to keep in mind. Every additional expenditure on durables result in 2.77 unit more sales in refrigerator.

(Refer Slide Time: 24:46)





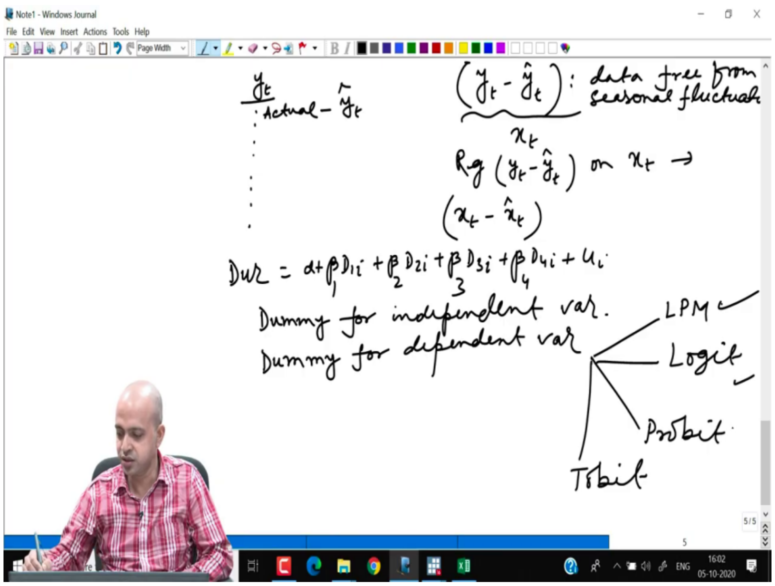
And if you introduce four dummies reg then d1 d2 d3 and durables you have to mention no cons that means I do not want intercept. And then what is happening? Here also you look at sales in all the quarters are significant but sales in fourth quarter is significant only at 10 percent level. Because the p values is 0.54. So, if you multiply by 100 it would become 5.4 which is greater than 5 but less than 10, significant at 10 percent. But as I said earlier the econometrician, they suggest that we should take significance only at 1 percent and 5 percent. So that means we will say that this particular result shows that we have very weak evidence of sales in the fourth quarter.

Then obviously since it is refrigerator sale fourth quarter is basically starts from October, November, December. And October, November, December time is not very suitable, favorable for refrigerator sale. That is that time temperature is quite low so without refrigerator also you can manage. So that is why probably this data shows that in the fourth quarter sales is, sales for refrigerator is not that significant, not that significant.

And durables is again 2.773424 that means even if you introduce four dummies also with no constant, the responsiveness of your refrigerator sales with respect to durable expenditure is almost same 2.77. This is how you have to; you have to, you have to accommodate the seasonal fluctuation in your model otherwise your model may end up giving some spurious regression.

So, we were basically trying to remove the seasonal fluctuation from the dependent variable. But what about seasonal fluctuation on the quantitative covariate, that means the expenditure for durables that we have not done. But interestingly there is a theorem which says that when you include dummies to, to eliminate the seasonal fluctuation from your dependent variable, dependent variable automatically it removes any sort of seasonal fluctuation from the independent variable as well. You do not have to do anything additional. How will you check that? There is a way by which you can check that.

(Refer Slide Time: 28:16)



See initially we said that our objective is to remove the seasonal trend from this. And how will you do that? That means you have let us say you have sales data for each and every quarter first, second, third fourth then again first, second, third, fourth you have actual data you have sorry actual data. If you remove yt hat for that quarter you will get the data on yt which is, which is completely free from the seasonal fluctuation.

So yi minus yt minus yt hat. So, this yt minus yt hat is basically your data free from seasonal fluctuation, seasonal fluctuation. And in the process your xt is also, your xt is also eliminated. Any sort of seasonal fluctuation is eliminated from xt as well, that is the beauty of this procedure. So how will you check that? If you regress this data with xt, so that means if you regress yt minus yt hat on xt then the coefficient what you will get and if you regress the same procedure if you follow then xt minus xt hat that means you have to take xt as your dependent variable your durables would be the dependent variable.

Then your alpha plus beta 1 D1i plus beta 2 D2i plus beta 3 D3i plus beta 4 D4i plus ui. In the same procedure if you now subtract the predicted xt which is dependent variable here from xt. Then what you will do? If you run that, use that as explanatory variable you will get same sort of coefficient that is proved. So, there is no need for removing seasonal fluctuation separately from the explanatory variable. The procedure itself take care of any sort of seasonal fluctuation that might exist present in the context of xt as well.

So, with this we are basically closing our discussion on dummy variable. We are closing our discussion on dummy variable. Now so far if you think about all the dummy variable models then we use dummy for some of our independent variables. For example, here we are using dummy for the season, for quarter which is independent. But it may so happen that your dependent variable also is qualitative in nature.

So that means so far, we have discussed only dummy for independent variable. But we have not discussed dummy for dependent variable and that we will discuss later on by three types of qualitative response model, linear probability model, then logit model, then probit model and the tobit model, that we have not yet discussed. That we will discuss later on.

So, with this we are basically closing our discussion on dummy variable where dummy was used to quantify some of the qualitative explanatory variable, some of the qualitative explanatory variable. We have not yet discussed anything to convert the qualitative dependent variable into quantitative one using dummy. That we will discuss later on by this four components. Thank you very much.