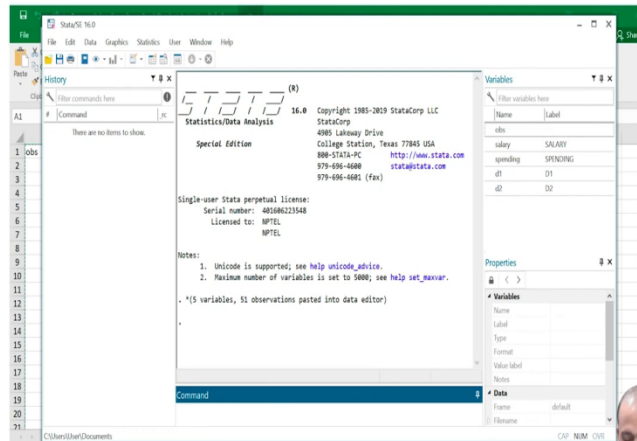


Statistical Analysis of Dummy Variable Models and Testing for Seasonal Fluctuations
Part-1
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 38

(Refer Slide Time: 0:15)



So welcome to the discussion of Dummy Variable Model once again. So far if you recall we have discussed mainly the theoretical portion of dummy variable models, that means we learned how to basically convert a qualitative information into a quantitative one. And then based on our objectives we have learnt how to set different types of dummy variable models.

Now today's class what we will do? We will learn how to estimate those dummy variable model that we have discussed earlier using some data set. And we will be using again the same statistical software, Stata for our estimation purpose. And the data set what I am going to use that is also available in your textbook, in the textbook of Gujarati.

This particular example what we are going to discuss and estimate, it is basically the data on primary school teacher's monthly salary taken from different states of the U.S. And what is our objective? Our objective here is to see whether there is any significant difference in the monthly salary of the primary school teacher who are coming from different regions of the United States.

And we have classified all the states into three regions- north east and north central, then second one is southern region, and the third one is western region. So, for the region we have three regions. So that is why as you know we will introduce two regional dummies.

And before we estimate we will first discuss how to basically set up that type of econometric model including the regional dummies wherein our objective is to examine whether there is any significant difference in the monthly salary of the primary school teachers who are coming from different regions of the United States. So, whether region has an impact on the monthly salary is what our objective is. So first we will see how to basically set up that type of econometric model.

(Refer Slide Time: 3:23)

ANOVA model

① $Salary_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + U_i$

$E(salary_i | D_{1i}=0, D_{2i}=0) = \alpha$: mean salary of west

$E(salary_i | D_{1i}=1, D_{2i}=0) = \alpha + \beta_1$: mean salary of NE/NC

$E(salary_i | D_{1i}=0, D_{2i}=1) = \alpha + \beta_2$: mean salary of south

$D_{1i} = 1$ if NE/NC
 $= 0$ otherwise

$D_{2i} = 1$ if south
 $= 0$ otherwise

$D_{3i} = 1$ if west
 $= 0$ otherwise

$\beta_1: (\alpha + \beta_1) - \alpha$
 \Rightarrow difference in mean salary b/w NE/NC and the western region

$\beta_2: (\alpha + \beta_2) - \alpha$
 \Rightarrow difference in mean salary b/w south and the western region

α : mean salary of the base category/western region

What will be the problem in estimation if we introduce 3 dummies for 3 regions??

So here let us say I am putting salary of the primary school teacher coming from the i^{th} region or i^{th} states, sorry i^{th} state equals to alpha plus beta1 D_{1i} plus beta2 D_{2i} plus U_i . This is our model. And how we have defined D_{1i} and D_{2i} , the two regional dummies? D_{1i} equals to 1 if the particular i^{th} state belongs to north east or north central region and 0 otherwise.

Similarly, D_{2i} equals to 1 if south and 0 otherwise. We have three regions and that is why we have introduced two dummies. So that means you can easily understand what is our base category here for which we have not introduced any dummy that means when D_{1i} and D_{2i} both these dummies take the value 0 that actually indicates the base region which is western region in this particular example.

$D1_i$ equals to 0 that means it is not north east north central, $D2_i$ equals to 0 that means it is north south. So, if the region is neither north east, north central nor south obviously that would be our western region that is why there is no need to introduce one more dummy for the western region, the two dummies are enough to indicate the situation.

But suppose you have introduced one more dummy $D3_i$ equals to 1, if west and 0 otherwise. If I introduce one more dummy what would be the problem in estimation, can you think of? What would be the problem in estimation if we introduce one more dummy for the western region. So, I am writing this question for your thinking purpose.

So, you can take help from the previous discussion and try to answer this question. What will happen if we introduce three dummies for three categories, what will be the problem? Problem in estimation if we introduce three dummies for three regions. So, this is the question I am giving for your thinking purpose.

Now as I told you earlier before we estimate the model in a dummy variable set up what we need to do, we need to understand what is the interpretation of alpha? What is the interpretation of beta 1? And what is the interpretation of beta 2? And as we said the interpretations are all derived so we have to derive the interpretation of alpha, beta 1 and beta 2.

Without doing anything since this model involves only qualitative variable as explanatory variables in the right-hand side that means this is an example of another model. So, you can easily understand what is the interpretation of the intercept or alpha, when there is no covariate, there is no other quantitative covariates in the right-hand side, what is the interpretation of this intercept or alpha? Can you think up, do you recall?

If you recall, if you recall you will see that when there is no other covariates quantitative covariates apart from the dummies the intercept from that model basically indicate the mean value of dependent variables for the base category. That means here in this example alpha indicates mean salary of the primary school teachers coming from the western region since west is our base category, west is our base category.

Now how will you prove that? You can easily get the interpretation, do the expectation of salary given $D1_i$ equals to 0, $D2_i$ is also equals to 0 that will give you alpha. So that means this is the

mean salary, so alpha is basically mean salary of west. Then expectation of salary given $D1_i$ equals to 1, but $D2_i$ equals to 0 that will give alpha plus beta 1.

And what would be the interpretation of alpha plus beta 1 then? Alpha plus beta 1 will indicate mean salary of north east or north central mean salary of teachers coming from north east or north central that is the interpretation of alpha plus beta 1. And expectation of salary given $D1_i$ equals to 0 but $D2_i$ equals to 1 that will give you alpha plus beta 2. And alpha plus beta 2 is then mean salary of south.

Now how will you derive beta 1 and beta 2? So that means beta 1 you can derive as alpha plus beta 1 minus alpha, alpha plus beta 1 minus alpha is actually beta 1 and alpha plus beta 1 is the mean salary of north east or north central and alpha is the mean salary of west. So that means beta 1 basically indicates the difference in mean salary between western region and north east or north central region.

What I am saying beta 1 basically indicates the difference in mean salary between north east or north central and western region. We are trying to compare mean salary of north east or north central with that of the base category which is western region, depending on the sign and significance of beta 1 that means if beta 1 is positive and significant. Then what we will say? We will say that mean salary of north east or north central is more significantly higher than that of western region.

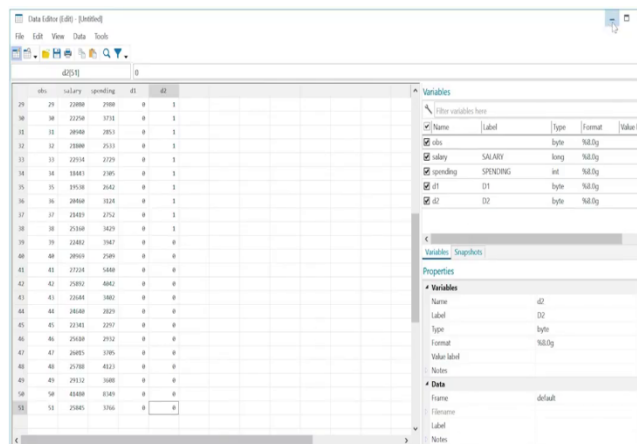
If beta 1 is negative and significant then we will say that mean salary of north east or north central region is significantly lower than the western region. And if beta 1 is insignificant then we will say that there is no significant difference between mean salary of north east or north central and west. This is how we have to interpret. So, beta 1 that means it basically indicate difference in mean salary between north east or north central and the western region, western region.

Now we will derive the interpretation of beta 2. How you can derive the beta 2? The same way alpha plus beta 2 minus alpha. So that means beta 2 indicates difference in mean salary between the southern region and the base category which is west. So, beta 2 indicates difference in mean salary, in mean salary between south and the western region. This is how we will, we will interpret the co-efficient.

I will write once again the interpretation of alpha, this is basically mean salary of the base category which is western region here. So, this is how we have to set up the model. And once again we have started with a very simple model here our assumption is that salary depends on monthly salary of a primary school teacher, depends on only one factor that is the regional dummy that means from which region you are coming that will tell you what type of salary you will get.

So regional dummy is the only one explanatory variable for the salary. This might be an unrealistic assumption, you might say that how can you say that their salary depends on the regional dummy? Yes, I understand that concern, but to start with we have set up a very simple model. After that we will set up other type of models including other covariates in the right-hand side. This is our starting point, this is our starting point. Let say this is model 1 which is a purely ANOVA model. So now what we will do? We will use the data set.

(Refer Slide Time: 16:18)



What I was talking about this is the data set look at. This is the data set and again I am telling you this particular data set is taken from the Gujarati's books only. If you go to dummy variable chapter you will get the data in tabular format and I will also share the data with you in soft copy so that you also can estimate this model using the statistical software, Stata or any other software's like R or Python.

So, I will supply you the data, you do not have to worry about the data set. All the data set that I am using here will be given to you. So here look at, our dependent variable is salary, monthly salary of primary school teachers coming from the different states of the U.S. Then we have one more explanatory variable that is the public spending for primary education. How much a particular state spends for primary education and this is per student value. So, 3,346 dollar per student per year is the public spending in that particular state.

And then we have introduced two dummies as we have discussed earlier $D1_i$ and $D2_i$. Here we have introduced only $D1$ and $D2$. So, when $D1_i$ equals to 1 that means for this particular value $D1$ equals to 1 and $D2$ equals to 0 that means this particular state belongs to the north east or north central region that is how we have defined $D1_i$ equals to 1, if the state belongs to north east or north central 0 otherwise.

So, from 1 to 21 see all these states are getting the value $D1$ takes the value 1, 1, 1, 1, 1, 1, up to 21. So, these first 21 states are basically coming from the north east or north central region that is why $D2$ is taking the value 0. After that what is happening?

(Refer Slide Time: 19:00)

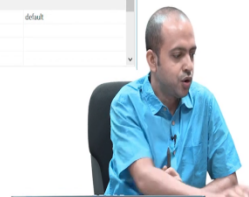
obs	salary	spending	d1	d2
17	17	23174	0	0
18	18	28021	1	0
19	19	38005	1	0
20	20	28939	1	0
21	21	23644	1	0
22	22	24024	0	1
23	23	27386	0	1
24	24	29966	0	1
25	25	23382	0	1
26	26	28627	0	1
27	27	22795	0	1
28	28	25578	0	1
29	29	22888	0	1
30	30	22258	0	1
31	31	28968	0	1
32	32	23888	0	1
33	33	22914	0	1
34	34	38443	0	1
35	35	39538	0	1
36	36	28868	0	1
37	37	23443	0	1
38	38	25348	0	1
39	39	23482	0	0
40	40	28963	0	0
41	41	27724	0	0
42	42	25892	0	0
43	43	23644	0	0
44	44	26048	0	0
45	45	23342	0	0
46	46	25028	0	0
47	47	24885	0	0
48	48	25708	0	0
49	49	23532	0	0
50	50	42448	0	0
51	51	25443	0	0



Now you see from 22, D2 takes the value 1 and it goes up to 38. So, from 22 to 38 D2 takes the value 1 that means these states are coming from the southern region, these states are coming from the southern region. And after that starting from 39 both D1 and D2 take the value 0. So that mean this is our base category as I said when both D1i and D2i they take the value 0 that particular state belongs to the western region.

(Refer Slide Time: 19:41)

obs	salary	spending	d1	d2
29	29	22888	0	1
30	30	22258	0	1
31	31	28968	0	1
32	32	23888	0	1
33	33	22914	0	1
34	34	38443	0	1
35	35	39538	0	1
36	36	28868	0	1
37	37	23443	0	1
38	38	25348	0	1
39	39	23482	0	0
40	40	28963	0	0
41	41	27724	0	0
42	42	25892	0	0
43	43	23644	0	0
44	44	26048	0	0
45	45	23342	0	0
46	46	25028	0	0
47	47	24885	0	0
48	48	25708	0	0
49	49	23532	0	0
50	50	42448	0	0
51	51	25443	0	0

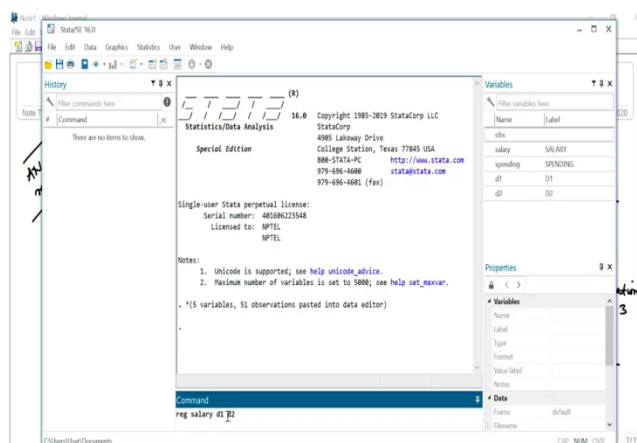


So that means 31, 39 to the end up to 51 all the states, all these states belong to the western region that is our base category here. So, given the data on salary I am spending what you have

to do in an excel sheet you have to define the D1 and D2 that means dummy variable you have to set. And you should learn how to introduce dummy variable then this is the way. You first introduce a dependent variable and independent variable and then you see the particular value is coming from which particular region.

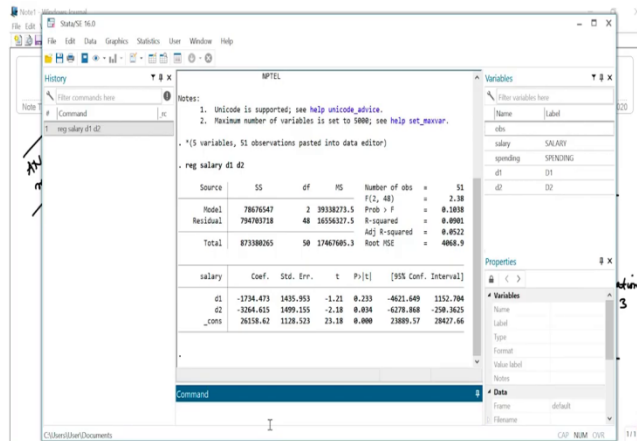
If it is coming from north east then put D1 equals to 1 and D2 equals to 0, if a particular value is coming from the southern region you have to put D2 equals to 1 and D1 equals to 0. So likewise, you have to design you have to introduce dummies and you have to give the values 1 and 0 depending on which particular region that particular ith state belongs to. It is a very simple one but you must introduce the dummy before estimation. So now what we will do once you have introduced the dummies now you can easily estimate the model.

(Refer Slide Time: 21:07)



So here what we will do? We will first run the model. So what would be our command, reg then what is your dependent variables, salary and then D1 and D2, D1 and D2. So, every time you remember that you have to give one space. So first you introduce salary space D1 space D2. And after that you have put enter.

(Refer Slide Time: 21:42)



Notes:

1. Unicode is supported; see `help unicode_advice`.
2. Maximum number of variables is set to 5000; see `help set_maxvar`.

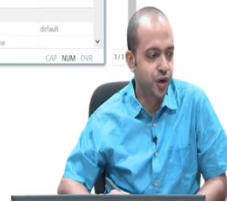
*(5 variables, 51 observations pasted into data editor)

. reg salary d1 d2

Source	SS	df	MS	Number of obs	F(2, 48)	Prob > F
Model	78676547	2	39338273.5		3.18	0.0000
Residual	794703718	48	16556327.5		R-squared = 0.0901	
Total	873380265	50	17467605.3		Adj R-squared = 0.0522	
					Root MSE = 4068.9	

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d1	-1734.473	1435.853	-1.21	0.233	-4621.649 1152.704
d2	-3264.615	1499.155	-2.18	0.034	-6278.868 -259.3625
_cons	26158.62	1128.523	23.18	0.000	23889.57 28427.66

Command



So, we have now estimated the model. And now what you have to do? You have to interpret the coefficients. And interpretation also we have already discussed. So, first of all once you estimate the model your first emphasis should be on the constant term. How to interpret the constant term which is basically the intercept or alpha here in our example?

So, the constant term is 26,150. 26,150 is the yearly salary of the primary school teacher coming from the western region. Because intercept indicates the salary of the base category which is western region. If you go back, if you go back look at alpha basically mean salary of the base category that means the western region, mean salary of the primary school teachers coming from the western region or which is a base category. So, you should immediately understand that this particular value, this 26,158 that is a mean salary of the primary school teachers from western region.

(Refer Slide Time: 23:23)

The whiteboard contains the following handwritten text:

$$\alpha = \$26,158 \Rightarrow \text{mean salary of the teachers from west}$$
$$\text{mean salary of NE/NC} = \$ (26,158 - 1,734)$$
$$\text{mean salary of south} = \$ (26,158 - 3,264)$$

The video inset shows a man in a blue shirt sitting at a desk, looking towards the camera.

So, let me write this. So alpha equals to 26,150 indicates mean salary of the teachers from western region. And then how will you interpret D1 and D2? As I told you D1i they indicate D1 means actually the beta 1 in our example. So you can easily see how will you interpret this one? Beta 1 is difference in mean salary between north east north central and the western region. So that beta 1 basically that means D1 indicates the difference.

So that means what would be the salary of the north east or north central? So that means that would be 26,158 minus since it is negative minus 1,734. So mean salary of north east, so that means mean salary of north east or north central equals to 26,158 minus what is the value minus 1,734 roughly 1,734 dollar.

And mean salary of south equals to 26,158 minus what is the value, the value is 3,264, 3,264 roughly. This is how you can determine the mean salary of the north east north central and southern region when you already know the mean salary of the west category. Because beta 1 and beta 2 they indicate the difference. So that means in this example D1 and D2 they indicate the difference in mean salary between western region and north east north central and between west and the southern region.

But one thing you have to keep in mind in this case look at the significance of D1 and D2. D1 the t statistic is 1.21 and the corresponding p value is 20.233. So, if you multiply that by 100 it

would become 23.3. So that means $D1_i$ is basically not at all significant since you are committing 20 around 23 type one errors while rejecting U null.

$D1$ is not significant. And if it is not significant what would be our conclusion? That there is no significant difference between the salary of primary school teacher coming from western region and north east or north central. There is no significant difference between north east north central and the western region. What about $D2$? $D2$ is corresponding t value is 2.18 and the p value is 0.034. So if you multiply p value with 100 it would become 3.4. 3.4 is greater than 1 but less than 5. So that means it is significant at 5 percent level. Then what would be our conclusion?

So that means salary of primary school teachers from southern region is significantly lower than that of the western region since the co-efficient is negative. So that is how you have to compare. Now one question that comes to our mind, if our objective is to compare the mean salary of the three regions and is to see whether there is any significant difference with inter regional salary why are you estimating the dummy variable model? you simply take the arithmetic mean of the salary of three regions.

And then you can just compare so that means if you compute the mean if you take the arithmetic mean of salary of the western region the value would be 26 point 26,150 and then that of southern region and north east region would be 26,150 minus 1,734 and southern region would be 26,000 minus 3,264 that would be the arithmetic mean.