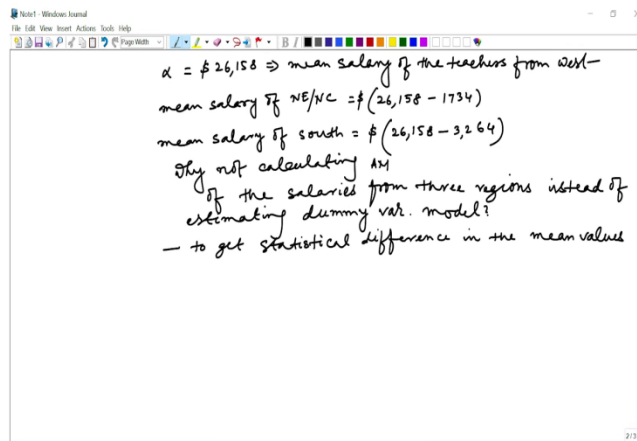


Statistical Analysis of Dummy Variable Models and Testing for Seasonal Fluctuations
Part-2
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Lecture 39
Indian Institute of Technology, Madras

(Refer Slide Time: 0:15)



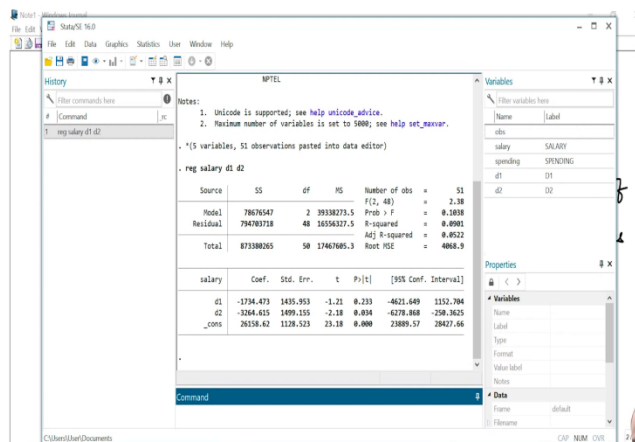
$\alpha = ₹ 26,158 \Rightarrow$ mean salary of the teachers from west -
mean salary of NE/NC = $₹(26,158 - 1734)$
mean salary of south = $₹(26,158 - 3,264)$
Why not calculating AM
of the salaries from three regions instead of
estimating dummy var. model?
- to get statistical difference in the mean values



So, the question what I am asking why not calculate the arithmetic mean of the salaries from three regions instead of estimating dummy variable model. Can you think of why we are not taking arithmetic mean and just compare? Yes, you can always compute arithmetic mean and comparing the mean you will get to know whether there is any mathematical difference between the mean values. But what you will not know is whether the difference is statistically significant or not.

So, in econometric analysis, always our objective is to get the statistical difference not the mathematical one. So arithmetic mean will only tell you whether there is mathematical difference in the mean values but this dummy variable model will tell you the statistical significance. So that means the answer to this question is to get statistically significant difference in the mean values. That is why dummy variable model.

(Refer Slide Time: 2:28)



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	R-squared	Adj R-squared	Root MSE
Model	78676547	2	39338273.5		3.18	0.0818			
Residual	794703718	48	16556327.5			0.8901			
Total	873380265	50	17467605.3			0.8522			4068.9

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d1	-1734.473	1435.853	-1.21	0.233	-4621.649 1152.704
d2	-3264.635	1499.155	-2.18	0.034	-6278.868 -259.3625
_cons	26158.62	1128.523	23.18	0.000	23889.57 28427.66

Command

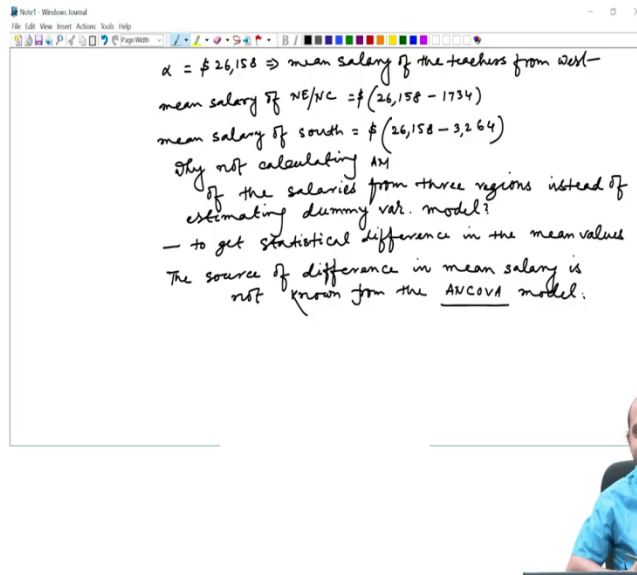


If you go to the output once again and compute mean value of the north east or north central, then that will also apparently show that the mean value of north east or north central is also different from that of western region because that value would be 26,158 minus 1,734. But this 1,734 difference, is it statistically significant? That the arithmetic mean cannot tell you. That we will know just by looking at the significance of the D1 coefficient and it is not significant.

So even though mathematically it is different, statistically it is not. Similarly, the D2 coefficient is statistically significant as well as mathematically. So that is the advantage of setting the dummy variable model and that is why we are setting dummy variable model even though our objective is to examine the statistical difference. Objective is to see the statistical difference in their mean salary.

But one more thing that we have to keep in our mind, whatever analysis we did so far, we only know there are differences in mean salary of the three of the other two regions compared to the base category. But we do not know what is the reason for that difference.

(Refer Slide Time: 4:05)



$\alpha = \$26,158 \Rightarrow$ mean salary of the teachers from west—
mean salary of NE/MC = $\$(26,158 - 1734)$
mean salary of south = $\$(26,158 - 3,264)$
Why not calculating AN
of the salaries from three regions instead of
estimating dummy var. model?
— to get statistical difference in the mean values
The source of difference in mean salary is
not known from the ANCOVA model.

The NPTEL logo is visible in the top right corner of the whiteboard area.

The source of difference in mean salary is not known from the ANCOVA model because in the right-hand side there is no explanatory variable. So, we do not know what is the reason for these differences.

Now can you think of what could be the reason? Why the other regions are earning less salary compared to the western region? One possible reason would be it might so happen that in western region the state government is spending more. Public spending per student is more compared to the other two regions.

And when public spending is more, obviously the salary of the teachers from that particular region will also be more. So public spending is one reason or might so happen that the cost of living is also higher in the western region for which the government is giving extra salary. So, what we have to do? Since we have data on the public spending as well, let us see whether there is any significant difference in the mean public spending of these three regions like the way we examine the difference in mean salary.

(Refer Slide Time: 6:35)

The whiteboard contains the following text:

$$\text{spending}_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + U_i$$
$$E(\text{spending}_i | D_{1i}=0, D_{2i}=0) = \alpha : \text{west}$$
$$E(\text{spending}_i | D_{1i}=1, D_{2i}=0) = (\alpha + \beta_1) : \text{NE/NC}$$
$$E(\text{spending}_i | D_{1i}=0, D_{2i}=1) = (\alpha + \beta_2) : \text{south}$$

Definitions of D variables:

$$D_{1i} = \begin{cases} 1 & \text{if NE/NC} \\ 0 & \text{otherwise} \end{cases}$$
$$D_{2i} = \begin{cases} 1 & \text{if south} \\ 0 & \text{otherwise} \end{cases}$$

An inset image shows a man in a blue shirt sitting at a desk, looking at a device.

In the same way we can examine the difference in mean public spending per student. So that is also possible. So how will you do that? In the same model instead of, so that means what we will do here, we will say that spending of the i^{th} state equals to alpha plus beta1 D_{1i} plus beta2 D_{2i} plus U_i .

So, in this model if you say this type of model then you will know depending on the value and say magnitude and the significance of beta1 and beta2 how I am defining beta 2 the definition of D_{1i} and D_{2i} is same. So same way I am defining D_{1i} equals to 1 if north east or north central and 0 otherwise, D_{2i} equals to 1 if south and 0 otherwise.

So, you will get expectation of spending. That means expectation of spending given D_{1i} equals to 0, D_{2i} equals to 0. So that would become alpha which is the spending for the base category of west. Likewise, you can derive the interpretation of beta1 and beta2 in the same way we have derived in the context of mean salary.

So, what we will do? We will now estimate the model. So, this is expectation of spending given D_{1i} equals to 1, D_{2i} equals to 0 that is alpha plus beta1. So, this is for the base category west. This is for the north east or north central and if you take expectation of spending when D_{2i} equals to 1 and D_{1i} equals to 0 that would become alpha plus beta2 which is basically for the southern region.

(Refer Slide Time: 9:25)

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	R-squared	Adj R-squared	Root MSE
Model	78676547	2	39338273.5		3.18	0.0818			
Residual	794703718	48	16556327.5			0.9901			
Total	873380265	50	17467605.3			0.9522			4068.9

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d1	-1734.473	1435.953	-1.21	0.233	-4621.649 1152.704
d2	-3264.635	1499.155	-2.18	0.034	-6278.868 -250.3625
_cons	26358.42	1128.523	23.18	0.000	23895.57 29427.66

Command
reg spending d1 d2



Now we will estimate. So same way now we will put reg for regressions space with you have to give a space then spending D1 and D2. And then after that you have to put enter.

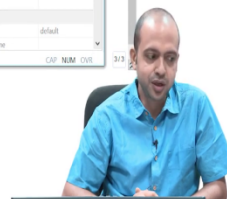
(Refer Slide Time: 9:55)

```
. reg spending d1 d2
```

Source	SS	df	MS	Number of obs	F(2, 48)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	4548121.39	2	2274060.7		3.14	0.0818			
Residual	51877878.8	48	1084122.43			0.9918			
Total	56425999.2	50	1124519.96			0.9435			1051.6

spending	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d1	-18.5348	364.844	-0.05	0.969	-750.4944 713.4248
d2	-644.7421	388.8971	-1.70	0.096	-1408.918 113.4341
_cons	3919.154	296.3941	13.19	0.000	3143.193 4634.495

Command



And same way you can interpret these coefficients. What does the constant term indicate? Constant term 3,919 tells you that this is the per student spending in the western region. So, government in western region spends 3,919 dollar per student yearly and that of north east or

north central is 3,919 minus of 18 point something and then for the southern region it would 3,919 minus 644.

(Refer Slide Time: 10:49)



(Refer Slide Time: 11:29)

Source	SS	df	MS	F(2, 48)	p	R-squared
Model	4548121.39	2	2274060.7	2.14	0.1291	0.0818
Residual	51877876.8	48	1080789.1			
Total	56426000.2	50	11285200.0			

Variable	Coeff.	Std. Err.	t	P> t	[95% Conf. Interval]
d1	-18.5348	3.46044	-0.89	0.380	-25.4944 733.4248
d2	644.7421	300.8072	1.79	0.0818	-1400.318 1311.8312
_cons	3919.154	286.1841	13.70	0.000	3343.393 4494.495



So that means you can say that alpha equals to 3,919. That is the mean spending. That means this is the spending per student in west. Now look at the coefficient D1 is not significant, D2 is significant. Of course, it is significant only at 10 percent level because if you multiply the p value with 100 that becomes 9.6 which is greater than 5 but less than 10.

So, one thing is clear since in the previous example we saw that there is no significant difference between the mean salary of the north east and north central with west. You see that there is no significant difference in spending as well. So, when there is no significant difference in spending obviously you can understand there will not be significant difference in salary as well which is true by looking at the results of the spending equation.

(Refer Slide Time: 12:56)

ANCOVA model

$$\text{Salary}_i = \alpha + \beta_1 D1i + \beta_2 D2i + \beta_3 X1i + U_i$$

No interaction b/w the dummies with $X1i$ (spending)

⇒ Responsiveness of salary w.r.t. public spending is same across regions

⇒ every additional spending results in same amount of increment in salary for all the regions.

$D1i = 1$ if NE/NC
 $= 0$ otherwise

$D2i = 1$ if south
 $= 0$ otherwise

$X1i$: spending per student

Now in the third case what we will do? In the first model we introduce only in the ANOVA format without introducing any explanatory variable in the right-hand side. Let us now introduce the covariate which is spending and modify the model. If you introduce the covariate in the right-hand side that would become ANCOVA model and it will look like salary equals to alpha plus beta1 D1i plus beta2 D2i plus beta3 X1i plus Ui.

And how we have defined X1, D1i and D2i? Their definition is same. So that means D1i equals to 1 if north east or north central and 0 otherwise, D2i equals to 1 if south and 0 otherwise. And X1i is actually the public spending per student. Now in this example what we did? We have introduced the dummies but we have not interacted the dummies with the covariate X1.

What is the assumption behind this type of model? When there is no interaction between the dummy variable and the explanatory variable which is spending here, what does it indicate?

What is the implicit assumption? First of all, what would be the interpretation of beta 3? Beta 3

is basically the responsiveness of salary with respect to public spending. So, when public spending increases by one-unit, on an average the salary increases by β_3 unit.

And there is no interaction between the dummy variable and the covariate. There is no interaction between the dummies with X_{1i} which is spending. And this basically indicates that responsiveness of salary with respect to public spending is same across different regions. So, one extra unit of spending results in β_3 amount of increment in the salary of the primary school teacher irrespective of whether the teacher is coming from southern region, western region, or north east, or north central region. That is the assumption. That is why no interaction between dummy and the covariate indicates.

This is actually not the same when public spending increases. Different states salary responds differently then you have to interact the dummy with the quantitative variables spending. We are not doing it here for simplicity sake but if you believe that then you have to interact the dummy variable with this. So, every additional spending results in β_3 amount of increment in salary for all the regions. That is the assumption we maintain here.

(Refer Slide Time: 19:36)

The screenshot shows the Stata command window with the command `reg salary d1 d2` and its output. The output includes a table of coefficients and statistics for the dependent variable 'salary'.

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d1	-1734.473	1435.953	-1.21	0.233	-4621.649 1152.704
d2	-3264.635	1499.155	-2.18	0.034	-6278.868 -250.3625
_cons	26158.62	1128.523	23.18	0.000	23895.57 29427.66

Below the coefficient table, the ANOVA table is displayed:

Source	SS	df	MS	Number of obs	F(2, 48)	Prob > F
Model	4548121.39	2	2274060.7	51	11.2251	0.0000
Residual	51877876.8	48	1080789.1			
Total	56425998.2	50	1128519.96			

The command window shows the command: `reg salary d1 d2`



The screenshot shows the Stata command window with the command `reg salary d1 d2 spending` and its output. The output includes a table of coefficients and statistics for the dependent variable 'spending'.

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
d1	-18.5348	364.044	-0.05	0.969	-750.4944 713.4248
d2	-644.7421	388.0671	-1.70	0.095	-1408.918 119.4341
_cons	3929.154	296.1041	13.29	0.000	3343.983 4494.405

Below the coefficient table, the ANOVA table is displayed:

Source	SS	df	MS	Number of obs	F(3, 47)	Prob > F
Model	631161483	3	210387161	51	49.82	0.0000
Residual	242238782	47	5153995.11			
Total	873380265	50	17467605.3			

The command window shows the command: `reg salary d1 d2 spending`



Now what I will do? I will estimate this third model. And for your sake this particular example is an ANCOVA model. So now we will estimate this ANCOVA model and what would be the command for this? `reg then your salary then D1, D2 and spending`. This is your complete model spending and this is the result.

Now if you compare this result with the earlier one, this result is quite different from the result that we derived earlier. In earlier cases in the ANOVA model the intercept value was something around 26,000 but now it is 13,269. That means mean salary of the western region is now 13,269 which is quite different from the earlier models.

Why? Obviously earlier model was the ANOVA model because in earlier there was no quantitative covariate included in the model, spending was absent. In absence of spending variable, it might so happen that the earlier ANOVA model was miss-specified. So that is why your claim that all the difference in salary is coming just because the teachers are coming from different regions is quite observed.

So that is the reason why when you include the other covariates spending your results are changing. So, the D1 variable which was not significant earlier is now significant at 5 percent level. D2 variable which was significant earlier is now insignificant. That is the thing. So likewise, you can estimate the mean salary of the other two regions by subtracting 1,673 and 1,144 from the constant term.

But one striking point here is that the spending variable is highly significant and that is positive also. So that means when average spending increases, that results in significant increase in monthly salary. And the estimate of the spending coefficient is 3.28. So that means every additional unit of spending results in 3.28-unit increment in salary.

(Refer Slide Time: 22:28)

ANOVA model

← Salary = $\alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_{ii} + U_i$

No interaction b/w the dummies with X_{ii} (spending)

⇒ Responsiveness of salary w.r.t. public spending is same across regions

⇒ every additional spending results in β_3 amount of increment in salary for all the regions.

$\beta_3 = 3.28$ unit increment in salary

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

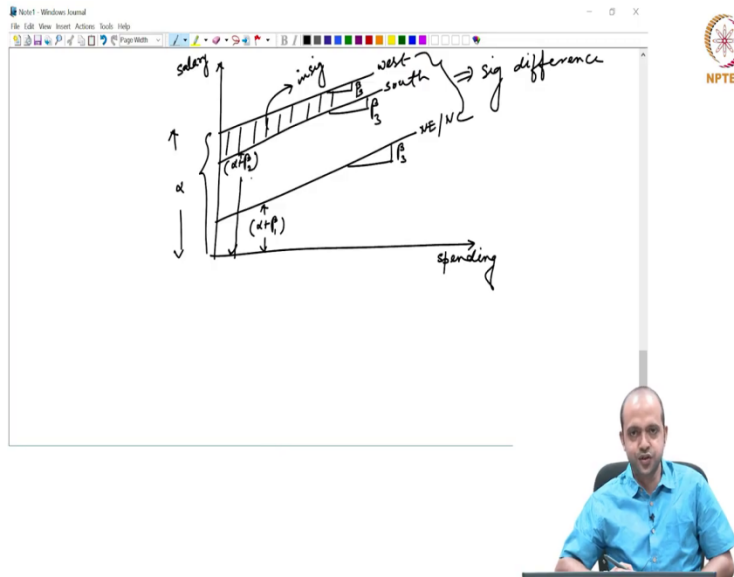
$D_{2i} = 1$ if south
 $= 0$ otherwise

X_{ii} : spending per student

NPTEL

So, this beta 3 equals to 3.28. So, every additional spending results in 3.28 unit increment in salary. Now what we will do? We will try to represent this result in a diagram. That will make the things much more clear.

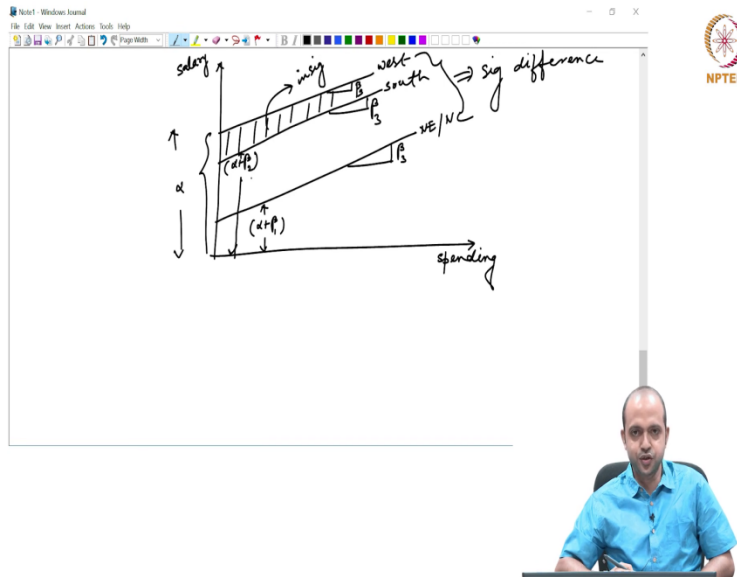
(Refer Slide Time: 23:25)



So, we will go to the next page and what we will do now? In x axis we are measuring spending and in the y axis this is salary. Let us say that this is for the west, the base category. So, this is basically alpha the coefficient the intercept term,. So, the intercept indicates the mean salary of the western region.

And then you see from the result that D1 is significant but D2 is not. So that means salary of the southern region is not significantly different from the western region. So, what I will do? I will put the southern regions value like this, this is for south and this is for, this is for north east or north central. Since all this regression have same slope and that is beta 3 all these lines are parallel with same slope.

(Refer Slide Time: 23:24)



So $\alpha + \beta_1$ would be intercept for north east or north central. And $\alpha + \beta_2$ would be for intercept for the south. So, since $\alpha + \beta_2$, that means β_2 is insignificant, I will say that since β_2 is insignificant so that means this is basically $\alpha + \beta_2$. And this is $\alpha + \beta_1$ that is significantly lower than that of the western region.

So, all the three regions have same slope. So that means we have to clearly keep in mind that every additional spending results in same amount of increment in their monthly salary or western region, southern region and the north east or north central region. Since the D1 is significant that means $\alpha + \beta_1$ if you see from the diagram $\alpha + \beta_1$ this is significantly lower than the west. So, the gap between west and north east or north central is significantly different.

But southern region is not significantly lower. So that means this difference is not significant. But difference between north west and north east or north central is significant. So once the value of α is given, you can easily calculate the mean salary of this. Given specific value of public spending. This is how you have to interpret the coefficients.

So, what we have learned so far? How to estimate the model and what is the interpretation of the coefficient. If it is ANOVA model, then the intercept will directly tell you the average value of the dependent variable which is salary here. If it is ANCOVA that will only tell you the intercept.

For example, here I cannot say that α is basically mean salary of the western region because I have one more covariate which is X_{1i} . So, you have to specifically put one value for X_{1i} to get

the mean salary of the western region. Similarly, for southern region and north east region. So $D1_i$ and $D2_i$ in this context indicate the difference in the mean salary for every additional unit of spending.

In mean salary between the western category, which is the base category, with the other two categories. So now what you have to do? You just read your textbook, they have nicely explained this particular example for your convenience for your understanding. I have taken the same example and I have explained how to interpret the coefficient, I have also given the demonstration how to estimate the model using Stata.

So, once I give you the data set you also can easily estimate the models and interpret. You first try to interpret of your own and then you see whether your interpretation is matching with what is given in the textbook.

With this we are closing our discussion today. And in next day tomorrow we will take another example and we will see how to estimate that particular model and interpret the coefficients. So, we are closing our discussion with this today.

Thank you very much.