

**Introduction to Economics**  
**Professor Sabuj Kumar Mandal**  
**Department of Humanities and Social Sciences**  
**Indian Institute of Technology Madras**  
**Dummy Variable Analysis and Application of Difference- in Difference for Impact Evaluation Part-4**

(Refer Slide Time: 00:14)



Application of DID for impact evaluation

$$y_t = \alpha + \beta_1 D_t + u_t$$

(A) 2000

1990  
 ⋮  
 +2000 → financial hub was established  
 ⋮  
 2005

$D_t = 1$  for year 2000 and onward  
 $= 0$  for 1990 to 1999  
 $\beta_1$ : sig & positive

Application of DID for impact evaluation

$$y_t = \alpha + \beta_1 D_{it} + \beta_2 D_t + u_t$$

$$y_t = \alpha + \beta_1 D_{it} + \beta_2 D_t + \beta_3 (D_{it} \times D_t) + u_t$$



	Post 2000	Pre 2000	
Treatment	$\alpha + \beta_1 + \beta_2 + \beta_3$	$\alpha + \beta_1$	$\beta_2 + \beta_3$
Control	$\alpha + \beta_1$	$\alpha$	$\beta_2$
T-C	$\beta_1 + \beta_3$	$\beta_1$	$\beta_3$

(B) controls

Even before setting up financial hub in 2000 real estate price in A was higher than B.  
 DID

1990  
 ⋮  
 +2000 → financial hub was established  
 ⋮  
 2005

$D_t = 1$  for year 2000 and onward  
 $= 0$  for 1990 to 1999  
 $\beta_1$ : sig & positive  
 $D_{it} = 1$  if A  
 $= 0$  if B

So, application of DID, this I will say application, application of DID for impact evaluation. For this, what we need, minimum two years of data point, two years of data point for impact evaluation kind of study, if you have panel data that is also fine. So, first I will give you an

example. Suppose, in 2000, in 2005, in 2000 this is let us say a place called A, this is a place called A, where a financial hub actually was established.

Let us say if you go to (01:29) site in Chennai, you see the near the Seruseri, there is one financial hub, that means a lot of IT companies and financial companies, they are there. And all those companies they set up their operation in the year 2000. And, and we assume that because of these financial harbor IT hub which was established at place A in 2000, the real estate price gone significantly high after year 2000.

So, that means if you have let us say data from 1990 to let us say 2005 and this is the year of 2000, where the IT or financial hub was established. So, you have these 16 years data. And in 2000 there was a financial hub. So, that means if you collect the real estate data for this particular area, then you can say that, let us say I have another, this is 1990 to 2005, 16 years data available, and you are thinking that in 2000, there is a setup of financial hub.

Because of that the real estate price gone up significantly compared to the other places, compared to the pre financial hub prices. Now, what, what type of model you will say, you will say this type of model, let us say, let us say  $y_t$  is the real estate price equals to  $\alpha + \beta_1 D_t + U_t$ . And how I have defined  $D_t$ ?  $D_t$  equals to 1 for year 2001 and onward and 0 for 1990 to 1999.

And then, if you define the dummy, let us say if your  $\beta_1$  turns out to be significant and positive, let us say I will change the model a little bit, let us say that I have, I have placed A and this is my place B, place B where they, there is no new financial hub, let us say this is simply your this Velachery or some other area, where there is no financial hub.

That mean this place where the financial hub was set up in econometric literature, we give a different name called treatment, treatment and this is control because here you do not have financial hub. So, we have basically when you set up your model  $y_t$  equals to  $\alpha + \beta_1 D_{1t} + \beta_2 D_{2t} + U_t$  and  $D_{1t}$ ,  $D_{1t}$  is equals to 1 if A, if place B.

So, that means you have a real estate data for both the areas, you have real estate data for A, real estate data for B, B is actually my control area, because there was no financial hub set up in area B, there is financial hub established only in area A. This is my objective. So, you have set up this

type of model and then you when your beta 1, sorry, this beta 2, this would become now beta 2, your beta 1 is significant beta 2 is significant and positive, then you will say that real estate price in A has gone up significantly compared to the area B because of this financial hub.

So, that means you are trying to estimate the impact of setting up a financial hub in place A compared to the place B which is my control, which is my control. So, this is, this setup is actually called impact evaluation. Now, if you set up this type of model and try to conclude that yes, beta2 is positive and significant that means, the financial hub has a positive impact on real estate price for place A then your conclusion would be wrong, why this is wrong?

Because it may so, happen that even before setting up the financial hub at place A in 2000, real estate price in place A was higher pre financial hub set up also, this is possible you do not know. So, that means what I am saying that even before setting up financial hub in 2000 the real estate price in A was higher than B, because you have not collected that type of information.

So, it is, if that is the case I cannot say that yes, it is because of setting up the financial hub, the real estate price in A is higher than real estate price B. So, what basically then we you have to check? You have to compare, you have to first estimate the difference in the real estate price between A and B in pre financial hub setting and then that means pre 2000, then post 2000 then we have to see whether there is a change in the difference in difference.

That means, I understand there is a difference in real estate price between A and B in pre 2000, there is also a difference in the real estate price between A and B post 2000 but my hypothesis says, that post 2000 difference between A and B is significantly higher than pre 2000. And if that becomes true, then only I can say that this financial hub has actually positive impact on real estate price for the area B.

Then how you have to modify this, like the previous example? Instead of having only the dummy variables added in the model, additively you have to interact these two dummies. So, that means your model, if you want to get a DID set up, would become  $\alpha + \beta_1 D_{1i} + \beta_2 D_{2t} + \beta_3 D_{1i} \text{ interacting with } D_{2t} + U_{it}$ . Now, again I will, what I will do?

I will do this is, let us say post 2000 and this is post 2000, this is pre 2000 this is area A which is treatment, treatment and this is my control. That means area B. So, once again what you have to

do, that expectation of, that means you have to get treatment. So, I am not showing the expectation, you have to calculate this and put it in the value here.

So, what I will do, now 2000 onwards and for the treatment group, you will have  $D_{1i}$ , sorry,  $D_{1i}$  equals to 1 and  $D_t$  is also one. So, that means, when you put both 1 and 1 this would become  $\alpha + \beta_1 + \beta_2 + \beta_3$  and pre 2000 for the treatment group, that means your  $D_t$  equals to 0, but  $D_{1i}$  is equals to 1.

So, that means, this would become  $\alpha + \beta_1$ , this would become 0 this would also become 0. So,  $\alpha + \beta_1$  is the pre 2000 real estate price for the treatment group, because this would become, since it is treatment  $D_{1i}$  equals to 1. So,  $\alpha + \beta_1$ , these would vanish and if you put 0 this will also vanish.

Similarly, for the control category post 2000 would be  $\alpha + \beta_2$  since it is control  $D_{1i}$  equals to 0 and you will have  $D_t$ . So, that means  $\alpha + \beta_2$ . Similarly, for the control and pre, that means both  $D_{1i}$  equals to 0  $D_t$  equals to 0 that is my benchmark category, you will have  $\alpha$ ,  $\alpha$  and then this is also 0, this is 0 everything is 0. So, now these I would say that treatment minus control for post.

So, that means if you take difference in these  $\alpha$  will get cancelled,  $\beta_2$  will get cancelled, this would become  $\beta_1 + \beta_3$ . And this would become  $\beta_1$ . So, if you take difference in these then again you will get your  $\beta_3$ . So, treatment minus control, the difference in the real estate price between A and B post 2000. And then I have also taken difference in real estate for pre 2000, then again I have taken the difference that is why  $\beta_3$  is DID here.

Similarly, you can take this way, row wise difference as well. So, then  $\alpha + \beta_1$  will get cancelled, you will get  $\beta_2 + \beta_3$ , here you will get better 2, that also you will get the difference in differences  $\beta_3$ . So, what you have, what you have to do? You should have two sets of data one for, actually there are four sub samples, one for the treatment, one for the control, one for pre, one for post.

So, this particular setup, if you can understand, these has very, very interesting and profound implication in the context of policy evaluation. So, you should not only introduce the two

dummies, but also interact them to get the impact evaluation or DID. So, this you can apply for many cases, for example, let us say you are thinking of impact of climate change on migration.

So, that means what you have to do, you have to take one treatment group, one treatment area, one control area, what is the treatment area where climate change is significantly higher. And then you have to get data from pre and post, let us say climate change became significant that means in terms of let us say temperature or rainfall, let us say that is the year 2000. In 2000, there was a drastic change in your climate. And that happened, let us say in place a.

So, what I have to do? I have to calculate the rate of migration for area A and area B and then I have to take the difference in rate of migration between A and B in pre 2000, when the climate was not changed drastically, then in post 2000 also, post 2000 also, I have to calculate the change in migration rate between A and B. Then I have to take difference in difference.

So, that means difference in rate of change of migration pre 2000 and difference in rate of change of migration in post 2000 whether they are significantly different, that could be determined by your beta3. And then you can say that, yes, climate change has an impact on migration. Because it may so happen that in place A even before climate change, also migration rate was significantly higher than B, I do not know.

Similarly, let us say another example quite frequently we do this type of study; let us say estimating woman empowerment, impact of joining self-help group on women empowerment. Let us say that in 2000, a group of women participated in self-help group, and you assume you hypothesize post 2000, the group of women, who participated in self-help group, their empowerment, and in the literature, there is a list of literature how to measure empowerment, I am not going into the detail of measurement of empowerment.

Basically, let us say how women are empowered in decision making process at the household. And when you join self-help group that will give you extra income and with economic freedom, it is assumed that woman gets empowered. Now if you assign one dummy for those who participated in the self-help group and say that, this dummy is significant. And as a result of which I am saying that self-help group participation gives more empowerment to the women that may be wrong.

Because it means so happen those group of, those, that group of woman, what you were considering they were already empowered, even before joining, they were more empowered compared to the control group, that mean those who have not participated even before joining the self-help group. Then what you have to do?

You have to identify those women who participated in the self-help group and you have to calculate their impairment with the control group, those who did not participate, then you have to calculate the difference in empowerment pre and post to joining in the self-help group. And you have to take difference in difference once again.

And if that difference in difference coefficient is significant, then only I can say that yes, joining self-help group has given a significant higher impact on woman empowerment, otherwise, your analysis will be wrong. Because it may so happen that those women were already more empowered compared to the others who did not participate in the self-help group. So, that means there are several cases you try to understand this case, then you can easily apply this particular technique DID in several other contexts as well, depending on your requirement, is that fine?

So this is another case, case 4, which is basically the application of dummy variables, when you have two dummy variables and the interpretation of the interaction between two dummies, that is quite important and interesting. So, we are closing our discussion with this today. And in our next class, we will take data set, and then we will try to estimate different type of dummy variable model.

We will see how to set up the data, how to define the dummy variable in your original data set. And then we will discuss the estimation of dummy variable model. And we will try, also try to understand the interpretation of the different estimates involving dummy variables, thank you.