

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-5

(Refer Slide Time: 00:16)

Application of difference-in-difference estimates.

$\leftarrow B \xrightarrow{9\text{ km}} A \xrightarrow{1\text{ km}} 1979$
 A garbage incinerator may come up in A
 1981 construction of the incinerator started

Objective: To estimate the impact of the garbage incinerator construction on the housing price in A.

$1981 \leftarrow \text{Price} = \beta_0 + \beta_1 \text{nearinc} + u_i$ nearinc = 1 if the house is located within 3 km of A
 $\text{Price} = 101,507.5 - 30,688.27 \text{nearinc}$ = 0 otherwise
 (t = 3098) (t = 5,827) = 4.4

- housing price in A is 30,688 lower than that of B
 \Rightarrow garbage incinerator results in lower housing price

Welcome, so, we were discussing about applications of dummy variable models, we have discussed several applications of dummy variable models so far. Today, what we will do, we will discuss one more application of difference in difference estimates. In our previous class, we already discussed the theoretical structure of difference in difference estimates.

And we said that in a dummy variable model, when you have two qualitative dummies, then interaction of those two dummies actually gives you coefficient of that interaction terms gives you an estimate of difference in difference estimate, which is useful for event evaluation or impact evaluation kind of study.

In our previous class, we discussed the difference in difference estimates using the examples of a financial hub coming in a place or to be its impact on the real estate price. And another example we were talking about when certain women they participate in a self help group, what is going to be the impact on their empowerment post joining that self-help group.

Today we will discuss that real estate example only, but in a different context, in a different context. Let us say that in 1979, in the year 1979, there is a rumor going on that, in a place let us say it is called place A. So, what we are discussing, basically the application of DID, difference in difference estimate, application of difference in difference estimates.

So, there is a place called A this is the place and in the year 1979, in 1979 there was a rumor going on that in this place a garbage incinerator may come up, what is the rumor? That a garbage incinerator may come up, a garbage incinerator may come up in A that was in 1979. And this type of rumor was not there in previous, or previous years to 1979.

And obviously, you can expect that if the garbage incinerator is set up in this place, in this neighborhood of A obviously the real estate price will go down, who wants to buy their house in a place where there is a garbage incinerator? So, let us assume that we are, we are drawing an neighborhood around this A and this is let us say in 3 km.

If a place is within 3 kilometer radius of A we will call that, that is the near, that place is nearer to our garbage incinerator or more than 3 kilometers means we will consider far away from garbage incinerator. And in 1981 actually, in 1981 the construction of the incinerator started. Now what is our objective our objective? Our objective here to estimate the impact of the garbage incinerator construction on the housing price in A.

Now if this is the objective then how can you apply a dummy variable model to test the impact of this garbage incinerator? That is the first thing, and applying a dummy variable model if you know the basics of dummy variable model, what type of model we can apply in this context? Let us say that we are defining our dependent variable as housing price, each price is a housing price equals to β_0 plus β_1 near incinerator plus u_i .

We are estimating this type of model in 1979 sorry, 1981. In the year 1981, when the garbage incinerator construction has started, we are estimating a model like this, we have collected housing price data and this is our model and near INC is basically a dummy variable how it is defined? Near INC equals to 1 if the place or let us say if the house for which we are collecting data on its price, if the house is located within 3 km of A, 0 otherwise.

If the house is located within the 3 kilometer radius of A we will say that that is equals to 1 that means we will consider the house is near i, incinerator, if the distance of the house is more than 3 kilometers from A, we will say that, for example, let us say this is a place, which is B and let us say this is the distance, this distance is let us say 9 km.

So, we will say that, that this, that house is actually far away from the place and there will not be any impact kind of thing, because the smell of the garbage or whatever that would be confined within, within 3 kilometers radius. And you have estimated this model in 1981. And then the estimated model basically, the estimated, the result of the estimates is housing price, h price, h price equals to let us say 101.30715, this is the value of beta minus 30,688.27 near INC.

And t value corresponding to this is 5800 let us say, this is 5827 and t value corresponding to this is basically 3098. So, from the t value, you can understand the variables are highly significant. So, that means I am putting 3 stars here called port. Now from these estimates what you can infer, what you can infer from these dummy variable estimates? As you know, the coefficient of near INC that means beta 1.

Since, the dummy variable is added in the additive format in the model what I say, that coefficient of that dummy indicates the differential intercept. So, that means, if you take expectation of each price, given beta 1 equals to 0, then that would become only beta 0 that means, these, these value 101 point that means, 1,01,300 and the intercept value indicates housing price for the base category, what is the base category?

The house, which is located far away from this and beta 1 basically indicates the difference in the housing price between place B and place A and since the difference is negative and significant, what I will say? That in 1981 housing price in A on an average is 30,688 less than the housing price in B. That is the interpretation you can derive.

What I am saying the intercept indicates the housing price of the base category, and the coefficient of the near dummy indicates the difference in the intercept. That means, I can say that in 1981 housing price in A is 30,688 lower than that of 1000 price of B which is basically 101,307, is that clear? So, that means you would be tempted then to say yes, this garbage

incinerator that means it has some negative impact on the housing price, because the housing price in A is 30,000 lower than that of place B.

You would be tempted to infer like this, but if you interfere this equation in this way, and then come to a conclusion that yes, garbage incinerator has negative impact on the housing price of place A, then your, then your conclusion would be wrong. So, what I am saying you will be tempted to derive this, in this type of conclusion housing price, price in A is 30,688 lower than that of B. Housing price in A is 30,000 lower than that of B, that is correct.

And that implies garbage incinerator that means these impact what I am saying this lower housing price in A is due to garbage incinerator, garbage incinerator results in lower housing price. So, the first thing is correct, because that is straight forward coming from the dummy variable estimates, I have estimated the model and my model that is also that, price of, housing price in A is 30,000 lower than B.

But whether that is due to the garbage incinerator or not that conclusion if you are drawing then that might be wrong. Why this is wrong? To understand that let us estimate a similar type of model in 1978, when there was no rumor going on about the garbage incinerator and real construction also there is no question of construction of the real garbage incinerator. So, let us estimate a similar kind of model in 1978, when there was no such a rumor going on, because the rumor itself started in 1979.

(Refer Slide Time: 14:50)



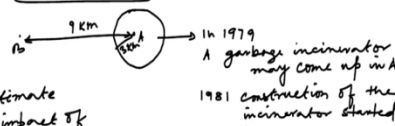
$\hat{\alpha}_{price} = 82,517 - 18,824 \text{ nearinc}$
 - Even in 1978 also housing price in A was 18,824 lower than that of B
 DID: $\hat{\delta}_1 = -30,688 - (-18,824)$
 $= -11,863$
 $\hat{\delta}_1 = (\overline{hprice}_{81,78} - \overline{hprice}_{81,78}) - (\overline{hprice}_{78,78} - \overline{hprice}_{78,78})$
Shortcoming of this approach??
 $\hat{\delta}_1 = -11,863$, but we do not know whether $\hat{\delta}_1$ is statistically sig or not.



Note Title

23-11-2020

Application of difference-in-difference estimates.



Objective: To estimate the impact of the garbage incinerator construction on the housing price in A.

$1981 \leftarrow hprice = \beta_0 + \beta_1 \text{nearinc} + u_i$ nearinc = 1 if the house is located within 3km of A
 $hprice = 101,507.5 - 30,688.27 \text{nearinc}$ = 0 otherwise
 (t = 3098) (t = 6,827)***
 - housing price in A is 30,688 lower than that of B
 \Rightarrow garbage incinerator results in lower housing price



So, in 1978 you have estimated a similar model, and that model shows h price estimated value of the housing price is 82,517 minus 18,824, 18,824 near INC. Now, what does it indicate? That means I am saying that, from this I can say even in 1978 also, even in 1978 also housing price in A was 18,824 lower than that of B.

So, housing price was already lower in price place A compared to B even in 1978 that means, we cannot ensure that this lower price is due to the garbage incinerator because the same thing was happening in 1978, also. But these two estimates actually gives you some idea about the impact,

how? In 1970, in 1981, the difference between housing, the difference in housing price between A and B was 30,684. But the difference in 1978 was only 18,894.

So, what has happened during this year from 1978 to 1981. So, that means during this year, the housing the difference in housing price between A and B has actually increased. And that difference in difference. That is basically the estimates of difference in difference. And that gives you some kind of idea of the garbage incinerator.

Do you understand what I am saying? In 1978 there was a difference between A and B, but that was only 18,824, but in 1981, when the construction of the garbage incinerator started that difference between A and B has increased from 18,824 to 30,688. And these difference in difference is basically, we can say now is the impact of the construction of the garbage incinerator.

So, that means the DID estimates, the DID estimates if you denote by δ_1 that is basically 30,000. So, that is basically minus 30,688 minus, minus of 18,224. That is basically 11,000 811,863 roughly. So, this is the impact and this is basically the DID and if you try to write it explicitly, then what will happen?

This δ_1 is basically the housing price, housing price in 81. The difference of these nr minus fr in 1981, minus h price 78 nr minus h price 78 fr . So, I have calculated the difference near and far away, between near and far away 81, near and far away 78 and that is basically your difference in difference estimates.

So this 11,863 is basically DID, which is basically the impact of construction on garbage incinerator on housing price. So this way, you can write two different dummy variable models. And you can take the difference of the estimates, the coefficient attached with the dummy to calculate DID. But there is a problem in this approach. What is the shortcoming? Can you think of this approach? What is the shortcoming of this approach?

So, the shortcoming of this approach, if you think closely, we have quantified the impact, δ_1 equals to 11,863. But we do not know whether δ_1 is statistically significant or not. That is the limitation of this approach when you write two different equations one for 1978 and another for 1981.

And then you take the difference in this way, we can quantify by looking at the value we can say that yes, the impact is 11,863 that is the DID estimates, but that is only a mathematical value. Since it is not estimated from a regression equation, we are not able to get the statistical significance of this DID estimates.

And as I told you several times previously, that in econometrics, what matters is basically the statistical significance not the mathematical one so to get the statistical significance then, what to do? we have to combine these two equations into a single equation, where we need to introduce two dummies, one is for year dummy that means there are two periods, one is 78 and other one is 81.

So, we can assign one dummy for the year and another dummy as we have already introduced near INC which will indicate whether the particular house, for which housing price data we are collecting is located within 3 kilometer radius of A or it is far away from A. Then you need to interact these two dummies and that interaction terms will basically give you the DID estimates.

And since that interaction term, you are going to estimate from the regression equation, obviously, you will get the magnitude as well as statistical significance of that. And that is basically a better approach because you need not, you do not have to estimate two different equation, one single equation will do everything for you. So, what is that approach?

(Refer Slide Time: 24:44)



$$r_{price} = \beta_0 + \beta_1 y_{81} + \beta_2 nearinc + \beta_3 (y_{81} \neq nearinc) + u_i$$

$y_{81} = 1$ if the year is 1981
and onward
 $= 0$ if 1978



$$r_{price} = \beta_0 + \beta_1 y_{81} + \beta_2 nearinc + \beta_3 (y_{81} \neq nearinc) + u_i$$

$y_{81} = 1$ if the year is 1981
and onward
 $= 0$ if 1978

β_3 : DID = $\hat{\delta}_1$ discussed earlier

$$\hat{r}_{price} = 82,517 + 18,790 y_{81} - 18,824 nearinc - 11,843 (y_{81} \neq nearinc)$$

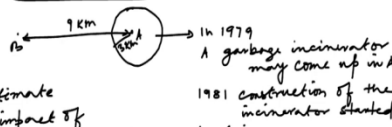
↓
DID
 $t = 7,452$
DID***





$\hat{\beta}_1 = 82,517 - 18,824 \text{ nearinc}$
 - Even in 1978 also housing price in A was 18,824 lower than that of B
 DID: $\hat{\delta}_1 = -30,688 - (-18,824)$
 $= -11,863$
 $\hat{\delta}_1 = (\overline{hprice}_{81,17} - \overline{hprice}_{81,87}) - (\overline{hprice}_{78,17} - \overline{hprice}_{78,87})$
 shortcoming of this approach??
 $\hat{\delta}_1 = -11,863$, but we do not know whether $\hat{\delta}_1$ is statistically sig or not.



Application of difference-in-difference estimator.

 Objective: To estimate the impact of the garbage incinerator construction on the housing price in A.
 $1981 \leftarrow hprice = \beta_0 + \beta_1 \text{nearinc} + u_i$ nearinc = 1 if the house is located within 5 km of A
 $hprice = 101,307.5 - 30,688.27 \text{nearinc}$ = 0 otherwise
 (t = 3098) (t = 5,827)***
 - housing price in A is 30,688 lower than that of B
 \Rightarrow garbage incinerator results in lower housing price



Then what I will do, I will write two equations. So, single equation approach, I will write housing price h price equals to β_0 plus β_1 , one year dummy, let us say this is y_{81} plus β_2 near INC plus β_3 , I will interact these two dummies, near dummy interacted with near INC plus u_i . Near INC I have already defined earlier, how I am defining y_{81} , equals to 1 if the year is 1981 and onward.

If you have more periods of data, if you have only two periods 1978 and 81 then this is only, if the year is 1970, 1981 and 0 if 1978, let us say that I have only two periods data, if the year is 1970, 1981 and 0 if 1978. Two periods data is enough to do this kind of impact evaluation. And

near INC, I have already defined. So, in this case, what will happen if you estimate, then beta 3 is basically called the DID, beta 3 is basically the DID.

And you can apply the framework, what we have discussed earlier, that means you have treatment and control. Treatment is the houses which are located within the 3 years, 3 kilometer radius of A and control is beyond and then pre and post, 1981 and before that, then if you take the difference in that, you will get the difference in difference estimates which is denoted as beta 3 following that framework, we have already discussed earlier.

Suppose after estimating these your model is like this, h price equals to let us say 82, 557, 517 plus, plus 18,790 this is your y 81 minus 18,824 near INC minus 11,863, 63 this variable y 81 which is interacted with near INC. So, that means this value is actually the DID. So, that means, if you compare the previous case that was also around 11,863 and let us say that this is 11,863.

So, magnitude wise there is not much of a change whether you apply single equation method or double equation method, but the advantage of this is, you are getting a t value, which is let us say 7456. This is the advantage, this 7456 what you are getting is basically source that this variable, this DID is significant at, is highly significant, significant at 1 percent level.

That is the advantage of this model. So, I am giving you as an assignment at home, as I said you try to derive this beta3 DID following the framework, we have already discussed in our previous class. So, you have treatment and control, you have pre and post, before and after. Then you need to calculate four alternative cases and then treatment minus control the framework I have already given to you.

This is, this would be the framework. This is the framework housing price 81 near, housing price 81 far, then housing price 78 near and housing price 78 far. Then you have to take difference in difference to get this delta 1 hat in this here it is basically beta 3. The delta 1 hat is basically beta 3 equals to delta 1 hat discussed earlier.

So, this is an application of difference in difference estimates. So, today, I have discussed both the approaches, single equation method and double equation method, both are giving almost similar type of magnitude, but advantage of this method is we are not only getting the mathematical value, but also statistical significance of that.

And this approach, we follow because in econometrics, we are not only interested in value, but also its statistical significance, that is why the single equation method, where we are basically interacting to dummies, year dummy as well as this near INC dummy that is the preferable approach that gives estimates of DID at one, you need not calculate it mathematically.

However, the beauty of this approach is that, it is also giving you almost similar value compared to the double equation method, but advantage is you are getting the statistical significance. So, this way if you learn this DID technique as I told you, it has many important and interesting applications, whenever you get a situation, where you are interested in impact evaluation kind of study.

So, you need to basically minimum two periods data is required to get that, to for applying this DID framework pre, post and then near and far away, minimum two periods data, more than two if you have then also it is fine, minimum two periods data if you can, you can apply in several contexts. This is another important application of DID. This example actually I have borrowed from Woolridge's book which I have referred introductory econometrics a modern approach.

This is the example what you will get, when not in the context of dummy variable, but when they are discussing pooled data, panel data, all these things that is beyond the scope of our discussion, we have not discussed pooled data and panel data, but you can go to that particular chapter. And you can get this particular example to understand, thank you very much.