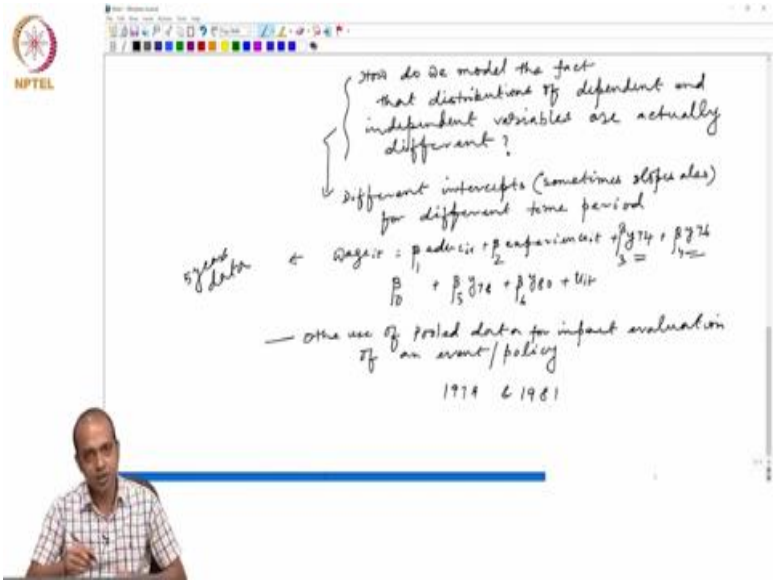
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 22
Pooled Data and Panel Data Model Estimation – Part II

(Refer Slide Time: 00:17)



Now I will show you one example of this pooled data.

(Video Starts: 00:22)

And this example is basically I will be using a specific data set. And this data is taken from Ulrich data. This is the data on women's fertility. So, this is the data. So, if you go to data then data editor, here you will be able to see this. This is the data look at this. This is a data on women's fertility, measured by number of kids. And we have data on 1972 so this is the starting year and this is 1972 then we have 1974 and 76 and 78 and 80 and then 82, 84.

So, 72 to 84 data, but in two years gapped. So, we have data on women's fertility, measured by number of kids then we have education level, then female education, then we have age, then whether the woman in black from which particular region this woman is coming from so on and so forth. Let us say that, we will be running one simple regression wherein, we say that fertility measured by number of kids.

Let us say reg $kids = f(women's\ age, education)$ and we will be using age and education, and we will be using all this year **(()) (03:17).** So, this is 74, 76, 78, 80 and 82. So,

our base year is 72. So, this is the regression what we are running and look at this. So, here interestingly what we can find is that the coefficient of education is negative and that is significant also.

It means basically that as education increases therefore women's fertility goes down, for whatever might be the reason maybe women's get married at a later age, they involve into job market because of higher education and then there is job market stress, so on and so forth which reduces women's fertility. So, there are established literature, which talks about the relationship between women's education and fertility.

That we can verify from the existing literature which is supporting the view that women's education is negatively correlated with their fertility. But what is interesting here, apart from 74 in all other years, the coefficients are negative even though not all the year term are significant. We get significant only 82 and 84. They are negative as well as significant. What does it mean?

It basically means that, compared to the base year women's fertility in 80 and 82 and 84 it was going down. So, average fertility of women in 82 is lower than 72 and what is the amount, by this amount 0.594. Similarly, 4.459, similarly 0.594, so, as I said this year dummies even though we include them in the model to recognize the fact that the distribution of the independent and dependent variables are different, at different time periods.

They also give some specific intuition or intuitive outcome and what is the additional insights? We get to know whether average fertility of the woman in other years are different from the base year 72 or not. So, in this particular example we see that in all other years that means 74, 76, 78 and 80 actually they are not different, but changes happen after 80, and reflected in 82 and 84, they are different.

So, this is how you can actually model a simple pooled data, and you can actually estimate. So, pooled data means we are still running the OLS model only. We are running the OLS but in a pooled data. There is no separate technique I am using here.

(Video Ends: 06:42)

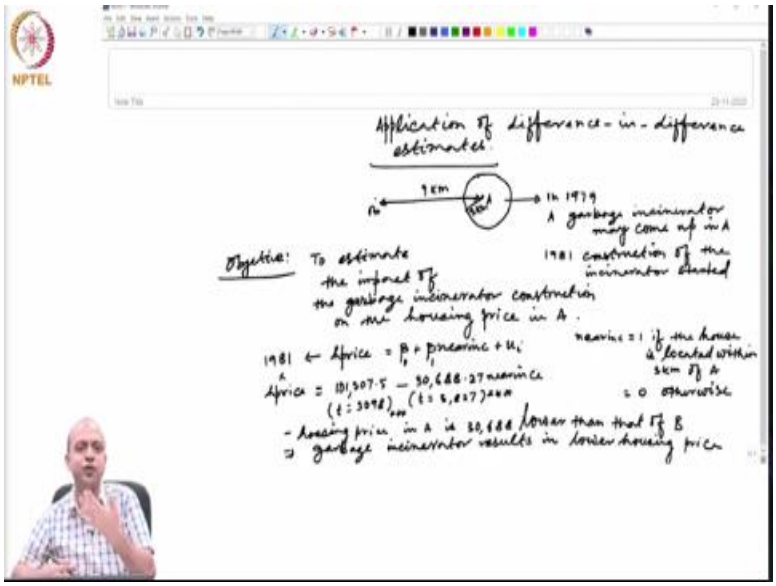
Apart from these, pooled data can also be used, other use of pooled data for impact evaluation of an event or policy. This is another use. So, if you have two periods' data, let us say I have

two periods' data on housing price and in a particular area a specific event has taken place. For example, let us say in a specific area a garbage incinerator has set up. A garbage incinerator has set up in a specific area and we can actually evaluate the impact of the garbage incinerator on housing prices, if you have minimum two period's data.

So, using again the knowledge of dummy variable we can modify, we can model this situation of two periods and then setting up a proper appropriate dummy variable model. And using a concept called difference in difference estimates, we can actually use this pooled data to evaluate the impact of that event on housing prices. That is possible. How? Let us assume that we have data on housing prices in two periods in 1979 and 1981.

Then we can actually use the pooled data for impact evaluation. What is the impact? Impact is for the setting up of a garbage incinerator and its impact on housing prices. We can actually use this data.

(Refer Slide Time: 09:06)



In 1979, in the year 1979 there is a rumour going on that in a place let us say it is called place A, this is the place. So, what we are discussing basically? The application of difference in difference estimate. There is a place called A, this is the place and in the year 1979, there was a rumour going on that, in this place a garbage incinerator may come up. What is the rumour? That a garbage incinerator may come up in A that was in 1979.

And this type of rumour was not there in previous all our previous years to 1979. And obviously you can expect that, if the garbage incinerator is set up in this place, in this

neighbourhood 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 drawing a neighbourhood around this A and this is let us say in 3 kilometres.

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

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 h price, is a housing price = beta 0 + beta 1 near incinerator + u i.

We are estimating this type of model in 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 = 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 kilometres of A 0 otherwise.

If the house is located within the 3 kilometre radius of a we will say that it = 1. That means we will consider the house is near incinerator. If the distance of the house is more than 3 kilometres 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 kilometres. So, we will say that this that house is actually far away from the place and there would not be any impact kind of thing.

Because the smell of the; garbage or whatever, that would be confined in the within 3 kilometres 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 = let us say 101,307.5 this is the value of β – 30,688.27 near inc. And t value corresponding to this is 5,800. Let us say this is 5,827 and t value corresponding to this is basically 3,098.

So, from the t value you can understand the variables are highly significant. So, that means I am putting three star here for both. Now from these estimates, what you can infer from this dummy variable estimates? As you know, the coefficient of near inc that means β_1, since the dummy variable is added in the additive format in the model what I said that coefficient of that dummy indicates the differential intercept.

So, that means if you take expectation of each price, given beta 1 = 0, then that would become only β_0 that means this value 101 point that means, 101,300 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 β_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 housing 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 30000 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 conclusion would be wrong.

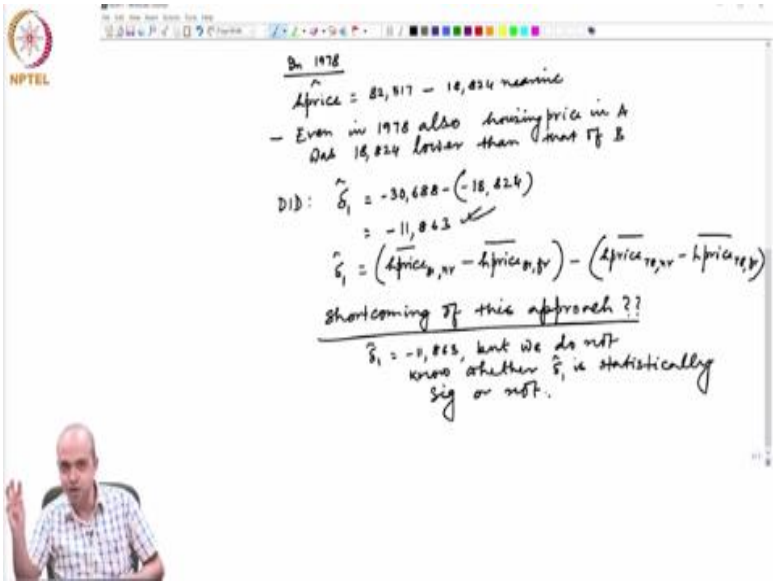
So, what I am saying you would be tempted to derive this in this type of conclusion. Housing price in A is 30,688 lower than that of B. Housing price in A is 30,000 lower than that of B, that is correct. And that implies garbage incinerator that means this impact what I am saying this lower housing price in A is due to garbage incinerator. Garbage incinerator the results in lower housing price.

So, the first thing is correct because that is straightforward coming from the dummy variable estimates. I have estimated the model and my model the results show 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 rumour 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 rumour going on, because the rumour itself started in 1979.

(Refer Slide Time: 22:08)



So, in 1978 you have estimated a similar model, and that model shows h price. Estimated value of the housing price is 82,517 – 18,824 near inc. Now what does it indicate? That means I am saying that, from this I can say 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 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 in housing price between A and B was 30,684. But the difference in 1978 it was only 18,894. So, what has happened during this year from 1978 to 1981? So, that means during this year, 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, the 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 if you denote by $\widehat{\delta 1}$ that is basically 30,000 so that is basically – 30,688 - of 18,224, that is basically 11,863 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 is basically the $\widehat{\delta 1}$ housing price in 81 near housing price in sorry this is not 18, this is 81and let us say far away.

The difference of these near minus far in 1981 - h price 78 near – h price 78 fr. So, I have calculated the difference near and far away between near and far away in 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 of garbage incinerator on housing price.

So, this way you can write two different dummy variable model and you can take the difference of the estimates, the coefficient attached with the dummy to calculate the idea. But there is a problem in this approach. What is the shortcut 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.
$\widehat{\delta 1}$ = - 11,863 but we do not know whether $\widehat{\delta 1}$ is statistically significant or not. That is the limitation of this approach when you write two different equation, 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.

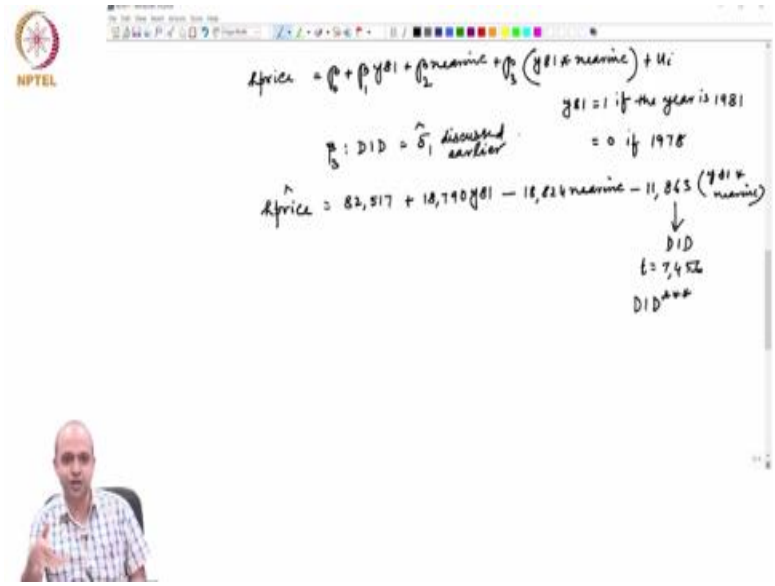
Since it is not estimated from a regression equation, we are not able to get the statistical significance of this DID estimates. And as I told you several times previously, that in econometrics what matters is basically the statistical significance, not the mathematical one.

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 another 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 kilometre 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.

(Refer Slide Time: 32:00)



So, what is that approach? Then what I will do? I will write two equations, so single equation approach. I will write housing price, $h_{price} = \beta_0 + \beta_1 +$, 1 year dummy. Let us say, this is $y_{81} + \beta_2$ near inc $+\beta_3$, I will interact these two dummies; year dummy, interacted with near inc + u i near inc I have already defined earlier. How I am defining? y 81 = 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 period's 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, β_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 kilometre 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 this, your model is like this h price = let us say 82,517 + 18,790 this is your y 81 – 18,824 near inc – 11,863, 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 7,456, what you are getting is basically shows that this variable this DID is significant at, is highly 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 β_3 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 4 alternative cases, and then treatment minus control the framework I have already given to you. This should be 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 β_3.

The $\widehat{\delta 1}$ is basically $\beta_3 = \widehat{\delta 1}$ 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 two dummies, year dummy, as well as this near inc dummy. That is the preferable approach that gives estimates of DID at one go, 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 advantages 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 there 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 apply in several contexts. This is another important application of DID. This example actually I have borrowed from Ulrich'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.