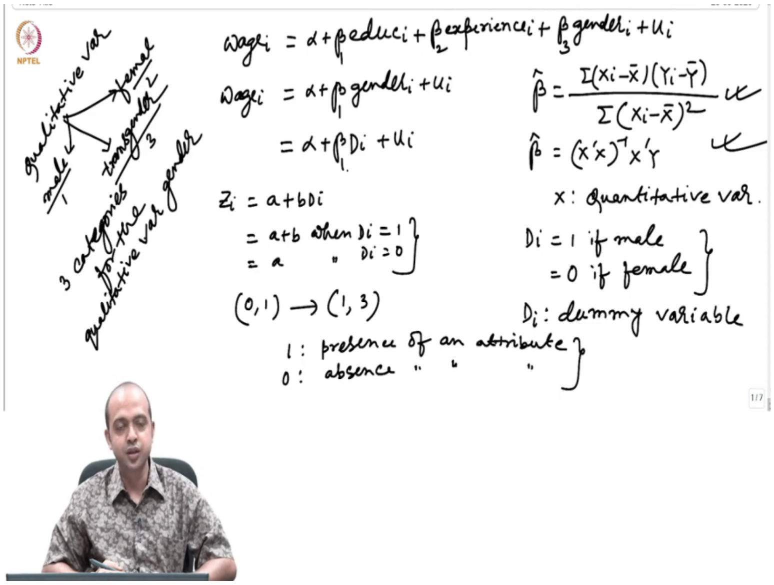
Introduction to Economics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Dummy Variable Analysis and Application of Difference- in Difference for Impact Evaluation-1

(Refer Slide Time: 00:14)



$wage_i = \alpha + \beta_1 educ_i + \beta_2 experience_i + \beta_3 gender_i + u_i$

$wage_i = \alpha + \beta_1 gender_i + u_i$

$= \alpha + \beta_1 D_i + u_i$

$\hat{\beta} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$

$\hat{\beta} = (X'X)^{-1}X'Y$

X : Quantitative var.

$Z_i = a + bD_i$

$\left.\begin{array}{l} = a+b \text{ when } D_i = 1 \\ = a \quad \text{"} \quad D_i = 0 \end{array}\right\}$

$D_i = 1$ if male
$= 0$ if female

D_i : dummy variable

$(0, 1) \rightarrow (1, 3)$

1 : presence of an attribute
0 : absence " " "

So, as we discussed yesterday, that today we are going to talk about a specific type of econometric model, where all or some of our explanatory variables in the right hand side, that we are using is qualitative in nature. See so far, whatever econometric model we have discussed and estimated, all our explanatory variables were quantitative in nature. For example, the last model we were discussing was related to saving and income.

So, here the explanatory variable income, you can easily measure quantitatively. Similarly, the previous example what we were talking about, that was child mortality rate and child mortality rate was explained by female literacy rate and per capita GNP, they are also both the explanatory variables are quantitative in nature. But sometimes, we do get a situation where some of our explanatory variable becomes quantitative in nature.

And then there is a specific technique required to quantify those qualitative variables otherwise we cannot estimate our model. So, what we will do, we will first take an example to understand what type of qualitative variable we may get in our econometric model. Let us say that we are talking about wage function where we are estimating the wage of an individual. Let us say that this wage of the i^{th} individual is actually alpha plus beta 1 when we think about wage function the first variable that comes to our mind is education.

And then you might also think about another variable, which is let us say experience, of the individual. And now, suppose you are having a research question in mind, that means you are interested to see, is there any gender discrimination in the labor market? That means with same amount of qualification and experience whether males are getting higher salary than their female counterparts? So, that means Is there any gender discrimination in the market, labor market or not? That is your research interest. So, the third variable is gender of the i^{th} individual. This is our model.

Now, for simplicity, what we will do, first we will remove the quantitative variables from the model. And we will keep only that qualitative variable gender in the model to make it simple. So, we will write this wage function once again. And we are assuming that wage is actually a function of alpha plus beta1 only gender of the individual. Now, gender is a qualitative information- male or female, it is a qualitative information. And the moment your variable is qualitative in nature, the estimation technique, what we have learned that cannot handle this type of qualitative information included in the model.

Why this is so? Because if you recall, how we have defined our beta hat, we were defining our beta hat in this way, summation xi minus x bar into yi minus y bar divided by summation xi minus x bar whole square. And this x variable is actually a quantitative variable, that is what we learn. So, this particular formulation, the way we have constructed our beta, even in matrix notation also, when you define beta hat equals to x prime x inverse x prime y even there also this x is actually a quantitative variable.

So, that means, both this formulation require some number for x, so that we can estimate beta hat quantitatively, but that is not the case here, here gender is a qualitative information. So, then what econometrician actually transformed this qualitative variable into quantitative one. So, that means, they have transformed the gender variable into some quantitative one. How did they do this? So, I can convert this qualitative information let us say, alpha plus beta1 D1i or let us say simply Di plus Ui. And how they have defined Di? Di is defined as Di let us say is defined as 1 if the i^{th} individual is male, equals to 0 if the ith individual is female. This is how we have defined Di.

So, now, the qualitative information of gender is easily transformed into a quantitative one because now it is all about 1 and 0. So, basically then this Di is actually a dummy variable. It is not the original variable, but the original variable gender, which was qualitative in nature, now transformed into a quantitative one through this Di and Di is called a dummy variable.

Now, both this formulation xi minus x bar yi minus y bar divided by xi minus x bar whole square or if you apply matrix algebra x prime x inverse x prime y both this formula can handle these quantitative. Because now it is all 0 and 1. So, this Di is called a dummy variable. And we can now easily estimate our model. Now, you might be thinking of, why all of a sudden 0 and 1? Actually there is no hard and fast rule, that Di should always be 0 and 1. Rather Di can always be transformed into this number, let us say I am defining zi equals to a plus bDi.

So, that means, where a and b are two numbers, two constant. So, now for any value of D, so that means if you put D equals to 1, then what will happen? When D equals to 1 then zi becomes a plus b and when D equals to 0, so that means zi equals to a plus b when Di equals to 1 and equals to a when Di equals to 0. So, depending on which particular value you put for Di you will get a specific value for zi. So, that means this 0 and 1, this gets converted into, let us say a equals to 1 and b equals to 2, a equals to 1 and b equals to 2.

So, that means it gets transformed into 1 and 3. So, that means 0 and 1 gets transformed into 1 and 3 depending on a specific value for a and b. Then why we are using 0 and 1? Because if you use 0 and 1 then the model becomes simple to interpret and the model becomes simple to derive certain important properties, that is the reason. So, throughout our discussion, we will use only 0 and 1 value for Di where 1 indicates the presence of an attribute and 0 indicates absence of an attribute.

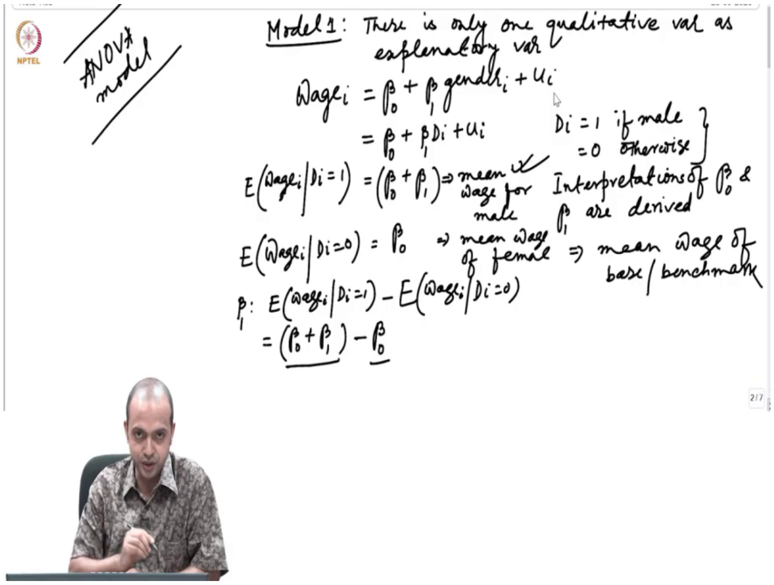
What is the attribute we are talking about? The attribute here we are talking about let us say, that male gender is that attribute. So, 1 indicates presence of that male attribute and 0 indicates absence of the male attribute, that means the person is female. You can define Di as 1 when female and 0 otherwise there is no hard and fast rule that always males would be 1 and females should be 0. Based on the researcher's intention, the researcher can decide or the econometrician can decide, which particular category you will take 0 and which particular category you will take 1.

So, only thing what I am saying that 1 indicate the presence of an attribute and 0 indicate absence of an attribute for the qualitative variable. So, here male and female, these are actually two categories. So, that means a qualitative variable will have two categories, category one is male and this is female. So, this is 1, this is 2, this is called categories. So, we have two categories. If you think that you will define the qualitative variable in terms of three categories, then obviously, you can add one more category that would be let us say become transgender category.

So, this is let us say three. So, here three categories for the qualitative variable gender. So, there are three categories of the qualitative variable gender. And we have to keep one thing in mind, if there are n categories, you have to define n minus 1 number of dummy variables.

So, here, we have two categories male and female. And that is why in our model, we have introduced only one dummy Di, but the moment you categorize the qualitative variable gender into three, male, female and transgender then number of dummy variable would be 3 minus 1, 2. So, we are just keeping that part aside. So, let us try to understand from the very simple model, so this is model number one. We will discuss several cases of the dummy variable one by one.

(Refer Slide Time: 14:09)



ANOVA model

Model 1: There is only one qualitative var as explanatory var

$$Wage_i = \beta_0 + \beta_1 gender_i + u_i$$

$$= \beta_0 + \beta_1 D_i + u_i \qquad \left.\begin{array}{l} D_i = 1 \text{ if male} \\ = 0 \text{ otherwise} \end{array}\right\}$$

$E(Wage_i | D_i = 1) = (\beta_0 + \beta_1) \Rightarrow$ mean wage for male

Interpretations of β_0 & β_1 are derived

$E(Wage_i | D_i = 0) = \beta_0 \Rightarrow$ mean wage of female $\Rightarrow$ mean wage of base / benchmark

β_1: $E(Wage_i | D_i = 1) - E(Wage_i | D_i = 0)$

$$= \underline{(\beta_0 + \beta_1)} - \underline{\beta_0}$$

So, this is let us say model one. We are talking about model one where there is only one qualitative variable as explanatory variable. And we have not included any other variable in the model. We are assuming that wage is a function of only the gender variable. And why we have assumed that? Because to keep the model simple, nothing else.

So, here, what we have to first learn is to, what is the interpretation of this beta 0 and beta 1 in this dummy variable model. So, this we will convert as beta 0 plus beta1 Di plus Ui where these Di equals to 1, if male and 0 otherwise. And we will try to understand what is the interpretation of beta0 and beta1. We have to carefully keep in mind, one thing what I will mention here the interpretation of the regression coefficients in a model involved with dummy variable are derived, there is no standard interpretation available.

What I am saying, in standard econometric model how do you generally interpret your beta0 and beta1? Beta0 is interpreted as intercept and beta1 is interpreted as slope coefficient. So, that means, if it is a saving-income relationship, you would interpret them as for unit change in your income, for 1 unit change in your income, on an average your saving changes by beta 1 amount. So, that is the standard interpretation, irrespective of whatever variable you take, your interpretation of beta0 and beta1 they do not change, but here in the context of dummy variable, I will write it here that interpretation of beta0 and beta1 are derived.

So, that means, you must derive the interpretation as there is no standard interpretation available. There is no standard interpretation available. So, that is the reason whenever you get a dummy variable model, first step is to derive their interpretation and then you estimate the model, without knowing the interpretation you should not estimate any model involving dummy variable. So, what we will do in this model also we will first try to derive the interpretation. So, like the previous case, first thing what we have to do? We have to take expectation of wage.

Given Di equals to let us say 1, for this category, equals to beta0 plus beta1. And what does Di equals to 1 indicate? Di equals to 1 indicate the male category that means, by expectation of wage given D1i equals to 1, I can say that this beta0 plus beta1 actually indicates the mean wage of the male category. So, this indicates the mean wage for male. Similarly, expectation of wage given Di equals to 0 equals to beta 0 which is mean wage of female category. And then what I said, the interpretation of beta0 and beta1 they are derived. So, you have to somehow derive beta 0.

So, how can you derive beta 0? So, if you look at beta0 and beta1 so that means beta0 is basically the mean wage, c of the female category. So, that means, in a model involving dummy variable, the intercept has a different meaning, it is not only the intercept, it is not simply the intercept, it is actually indicating the mean wage of the category for which you have not defined any dummy. So, here I have defined only one dummy Di equals to 1 for the male category, but for female category I have given 0. So, that means, intercept of a dummy variable model indicates the mean wage of the base category.

So, here female is basically the base category that means, I can say mean wage of base, or benchmark category. And the base or benchmark category is for which you have not defined any dummy. Here I have not defined any dummy for the female category it is 0, that is why female becomes my benchmark category or base category. So, that means intercept of this model beta 0 of this model indicates mean wage of the benchmark category, which is female and how will you get beta 1, see beta 1, how can you derive?

Beta1 is basically expectation of wage i given the Di equals to 1 minus expectation of wage given the Di equals to 0. So, that means, I can say that, this is beta0 plus beta1 minus beta0. So, how you are deriving beta1? Beta1 is nothing but beta0 plus beta1 minus beta0 and what is then this beta0 plus beta1 and beta0 if you know, then you can easily understand what is going to be the interpretation of beta 1. What is beta 0 plus beta 1? If you look at here, see beta 0 plus beta 1 is actually mean wage of the male category and beta 0 is the mean wage of the female category that means, beta 1 indicates that difference in mean salary between male and female.

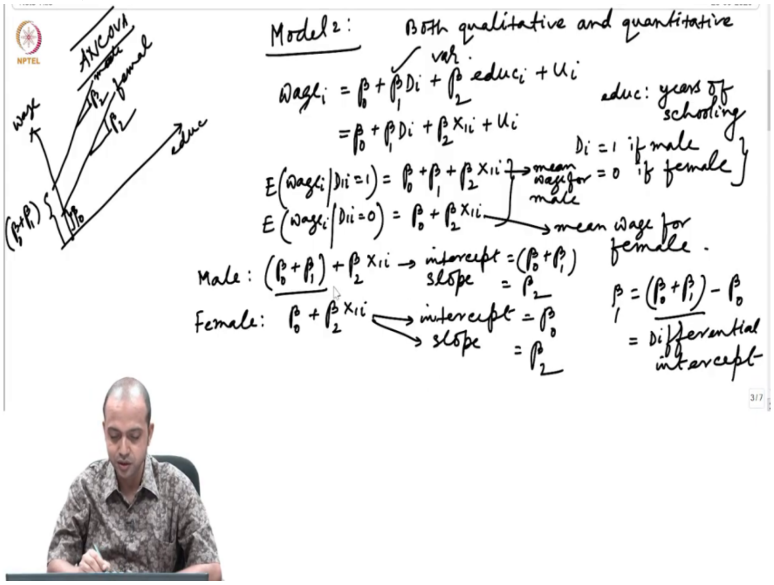
So, earlier in the standard econometric model, where beta 1 was showing the slope coefficient, that means for unit change in income, what is the change in saving on an average, the interpretation of beta 1 is totally different here in a model involving dummy variable. And what is the interpretation now, I have derived?

The interpretation is the difference in mean salary between the male and the female category. So, unless you derive beta 1 in this way, you will never get the interpretation of beta 1, that is why I have clearly mentioned that interpretation of beta 0 and beta 1 are all derived. So, this is how you have derived your interpretation.

So, this is model one. Now, in our next case, what we will do instead of involving only 1 qualitative information, let say that I am involving and this model, the model one, when you have only qualitative variable linear model that is called ANOVA model, this is ANOVA model. Because you are only discussing about the variation in wage in two different categories, you have not included any quantitative variable, which are known as covariates, x variable or explanatory variables in explanatory model they are also known as covariates.

So, this is called ANOVA model. Why? Because you are only discussing the difference in mean salary or wage between two categories. There is no quantitative variable or covariates involved in this model.

(Refer Slide Time: 24:58)



Model 2: Both qualitative and quantitative var.

$$wage_i = \beta_0 + \beta_1 D_i + \beta_2 educ_i + u_i$$

educ: years of schooling

$$= \beta_0 + \beta_1 D_i + \beta_2 x_{1i} + u_i$$

$D_i = 1$ if male
$= 0$ if female

$E(wage_i | D_{1i} = 1) = \beta_0 + \beta_1 + \beta_2 x_{1i}$ → mean wage for male

$E(wage_i | D_{1i} = 0) = \beta_0 + \beta_2 x_{1i}$ → mean wage for female.

Male: $\underline{(\beta_0 + \beta_1)} + \beta_2 x_{1i}$ → intercept $= (\beta_0 + \beta_1)$, slope $= \beta_2$

Female: $\beta_0 + \beta_2 x_{1i}$ → intercept $= \beta_0$, slope $= \beta_2$

$\beta_1 = \underline{(\beta_0 + \beta_1)} - \beta_0$
= Differential intercept

Now, suppose I am taking model two. Model two is a model where I have both qualitative and quantitative. And that model is called ANCOVA model. Because I have both qualitative as well as presence of covariates makes the model ANCOVA. And how does that model look like? Again wage equals to beta0 plus beta1 Di plus beta 2 let us say I am adding one qualitative, quantitative variable which is education.

And why I am calling education as quantitative variable? Because you can measure education by number of years of schooling, how I am measuring education? Number of years of schooling. So, these years can be easily measured that is why education is a quantitative variable. So, that means let me write this model as beta 0 plus beta 1 Di plus beta 2 let us say x1i plus Ui, and x1i indicates education. Then again, what do we have to do? We have to get the interpretation of beta 0, beta 1 and beta 2 and my Di equals to 1 if male and 0 if female or otherwise.

So, I am still assuming there are only two categories for the qualitative information gender. For the qualitative information gender, if you think more than two categories male, female and transgender then you have to simply include one more time, I am not discussing that, I will discuss in a later part of our discussion. So, now once again you have to get the interpretation beta 0, beta 1 and beta 2. So, expectation of wage i given D1i equals to 1 becomes beta 0 plus beta 1 plus beta2 x1i.

And expectation of wage given D1i equals to 0, it becomes beta0 plus beta2 x1i. And what is expectation of wage given? D1i equals to 1 this is basically the average wage for male, that means mean wage for male and what is this? This is mean wage for the female. Now, the question is how will you derive beta 1. So, that means from the wage function of the male category is beta 0 plus beta 1. I am putting both this term into bracket plus beta2 x1i.

So, that means I can easily think of that in male wage function, this is going to be the intercept and beta 2 is basically the slope. So, that means, beta 0 plus beta 1 is the intercept for the male's wage function and for the female category it is beta 0 plus beta2 x1i. So, that means for male category, this indicates intercept equals to beta0 plus beta1 and slope equals to beta2. Similarly, here intercept equals to beta0 and slope equals to beta2. Now if you compare these two that means in terms of a simple diagram will make you, let us say in this category x axis I am measuring education and here it is wage.

So, I am first plotting the male's wage function which is this, where the intercept is actually beta 0 plus beta 1. And for the female, this is the wage function. This is basically beta 0. So, it is male and this is female. Since these two lines are parallel, both males wage function and female wage function, their slope becomes beta 2. Both males wage function and female wage function, they have slope equals to beta 2, but there is a change in their intercept, while intercept for the male wage function is beta 0 plus beta 1, intercept for the female wage function is only beta 0.

Then how do you get beta 0? beta 0 plus beta 1 minus we will get beta 0. And what is this, beta 0 plus beta 1 is actually the intercept of the male and beta 0 if you look at is the intercept for the female categories. So, that means I can say, that this beta 1 what I have included in the model, this beta 1 is actually called differential intercept.

What I am saying so, that means this becomes a differential intercept. So, how do you derive the interpretation? So, when you put D1 i equals to 0 then females categories intercept is beta 0 and males categories intercept is when you put D1 i equals to 1, it becomes beta 0 plus beta 1 and the difference between these two intercept is beta 1, that is why the interpretation of beta 1 in this model is called differential intercept.

And beta 0 as usual is the intercept of the base category. Because if you have put D1i equals to 0 that becomes beta 0 plus beta 2 x 1 i. So, that means the beta 0 becomes the intercept for the male, female category. And beta 1 is the differential intercept, very simple to understand. So, the interpretation of beta 1 is the differential intercept in this model. And both male and females wage function is having slope equals to beta 2 that means there is no change in slope.