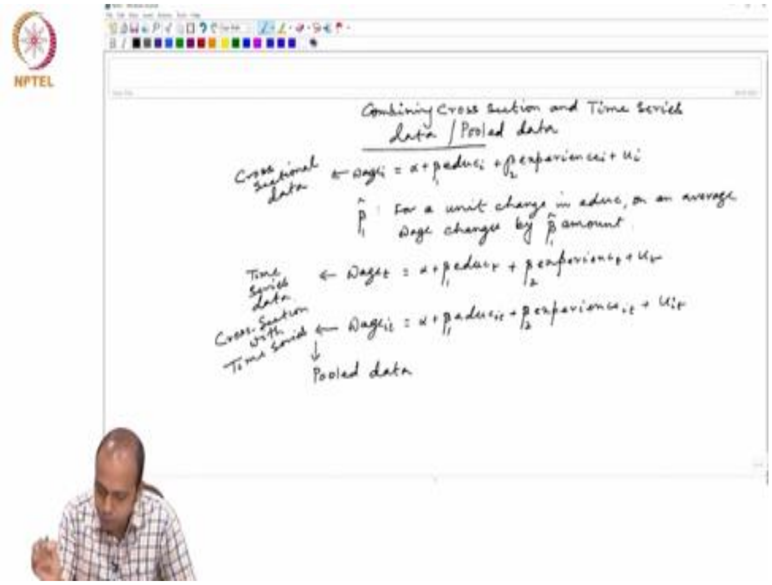
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 21
Pooled Data and Panel Data Model Estimation – Part I

(Refer Slide Time: 00:16)



Combining Cross section and Time series data / Pooled data

Cross sectional data ← $wage_i = \alpha + \beta_1 educ_i + \beta_2 experience_i + u_i$

$\hat{\beta}_1$: For a unit change in educ, on an average wage changes by $\hat{\beta}_1$ amount

Time series data ← $wage_t = \alpha + \beta_1 educ_t + \beta_2 experience_t + u_t$

Cross-Section with Time series ← $wage_{it} = \alpha + \beta_1 educ_{it} + \beta_2 experience_{it} + u_{it}$

↓

Pooled data

Welcome to our discussion on applied econometrics. And in our last class we were discussing about simultaneous equation model. Now today we will discuss another important and interesting topic in econometrics. Particularly econometric models when we combine cross-section and time series data. So, today's discussion is on combining cross section and time series data. So, we need to understand what is the specific need of combining cross section and time series data.

And what should be the appropriate econometric modelling to deal with the situation, so when you actually combine this type of two data sets. So, let us take an example let us say if you recall in our instrumental variable estimation, we were discussing about in wage function where wage is a function of $\alpha + \beta_1$ let us say education + β_2 experience plus let us say u_i. This is our model, we were talking about this type of specific model.

Now this is a purely cross-sectional data because see here I am using the subscript i that means all these variables it varies across cross section. So, what would be the interpretation of β_1? When you actually estimate your $\widehat{\beta_1}$, then your interpretation would be for a unit change in education on an average wage changes by $\widehat{\beta_1}$, amount. Now generally change the concept change we understand over time.

So, that means we think any change happens over a period of time. If that is the case in this particular model when you are saying for a unit change in education since this is a cross-sectional data you have to clearly keep in mind that there is no change over a period of time, rather what I am saying that it is a change across individuals. So, that means if you have different individual with different level of their education then, how their wage changes.

So, it is a change across several cross section. Since it is a cross-sectional data there is no scope for education to change over time there is no scope for wage to change over time. That is what we say that it is a change over a cross section not from this one. Now the same equation if we run in with time series concept. Then it would become $wage_t$ = α +$\beta_1 education_t$ + $\beta_2 experience_t$ +u_t . So, that is what we mean by time series data.

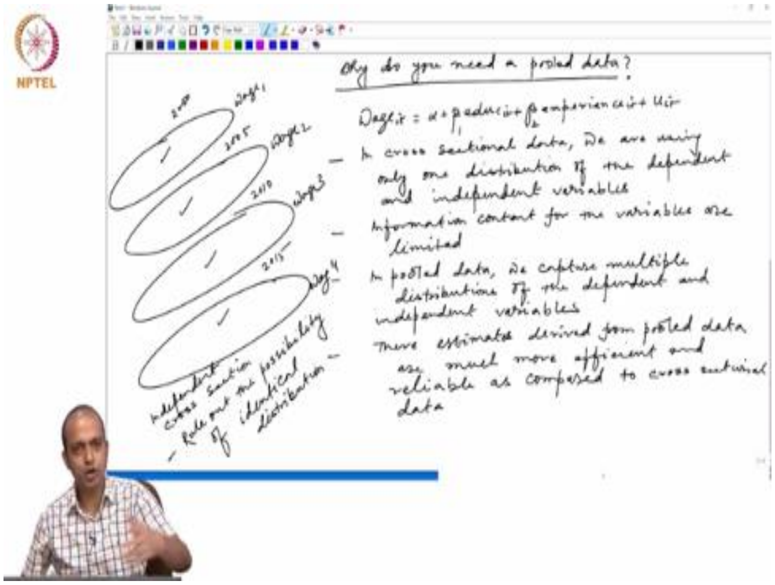
So, what do we mean that when you collect this type of data as a location changes for the unit for one unit change in education on an average wage changes by again $\widehat{\beta_1}$, amount but this change is basically over a period of time. But these data we have collected for a single individual maybe for a particular person. So, if you collect an individual's education for a period of let us say 10, 15 or 20 years and wage and experience maybe education may not change that much.

But you can regress this wage and experience because for a single individual it is quite unlikely that individual will after entering into the job market the education will change in each and every year. Probably you can run wage on experience itself. So, this is called a time series data. So, this is time series data this is cross-sectional data. Now suppose I am combining these two. So, that means what will happen if I combine these two?

Our equation would be $wage_{it}$ = α +$\beta_1 education_{it}$ + $\beta_2 experience_{it}$ +u_{it}. So, what I have done in this case? In this I have basically combined cross section with time series. So, here what I did I have combined cross section with time series. So, that means here the variables are changing over a period of time as well as across individual. So, let us say I have collected wage experience and education data for several individuals over a period of let us say 10 years.

Then that would become a cross sectional time series data. Sometimes it is also called as pooled data.

(Refer Slide Time: 09:23)



independent cross section

- Rule out the possibility of identical distribution

Why do you need a pooled data?

$Wage_{it} = \alpha + \beta_1 educ_{it} + \beta_2 experience_{it} + u_{it}$

- In cross sectional data, we are using only one distribution of the dependent and independent variables
- Information content for the variables are limited
- In pooled data, we capture multiple distributions of the dependent and independent variables
- These estimates derived from pooled data are much more efficient and reliable as compared to cross sectional data

Now next question what I am going to ask, why do you need a pooled data? Now sometimes if the question is asked to the students. Then they say that since I have many individuals' data for many years, I have used pooled data. Since I have many forms data on for many years, I have used pooled data but that should not be your justification. The justification for using pooled data comes from the limitation of a cross-sectional data or time series data.

So, let us try to understand what is the limitation of cross-sectional data. So, when you run this equation $wage_i$ = α +$\beta_1 education_i$ + $\beta_2 experience_i$ +u_i. let us say this is your education. So, in cross section what are you doing actually? In cross sectional data, we are using only one distribution of the dependent and independent variables. What is the distribution? That means if you collect these data for a particular period of time.

Then each of this variable will give you a particular distribution means what is the mean variance so on and so forth. So, one particular distribution is used for wage, one particular distribution is used for education, one particular distribution is used for experience. So, using a single distribution for all dependent and independent variables we are using this particular models. So, if that is the case then what will happen the information content for each and every variable is very less.

So, information content for the variables are limited because it is purely based on only one single distribution. Now what happens over a period of time the distribution of wage changes, the distribution of education changes, distribution of experience changes from the same population let us say we are talking about the wage function of Tamil Nadu. So, if you compare the distributional pattern of wage education experience of the working people in 2021 which is quite different from what was there in 2000.

Last 2021 there might have been many changes taken place in the labour market resulted in different time types of distributional pattern in wage, education and experience. Now if I combine the data of the cross section let us say 100, 200 or 500 individuals' data over a period of 20 years then in pooled data what I am trying to get is multiple distributions of the dependent and independent variables to estimate the relationship.

So, the estimates this $\widehat{\beta_1}$ $\widehat{\beta_2}$ which we derive from that type of pooled data is much more reliable and efficient compared to the cross-sectional data, because the information content is very limited in cross section. In pooled data we are able to capture multiple distributions of that dependent and independent variables. And therefore, what happens therefore estimates derived from pooled data are much more efficient and reliable as compared to cross-sectional data.

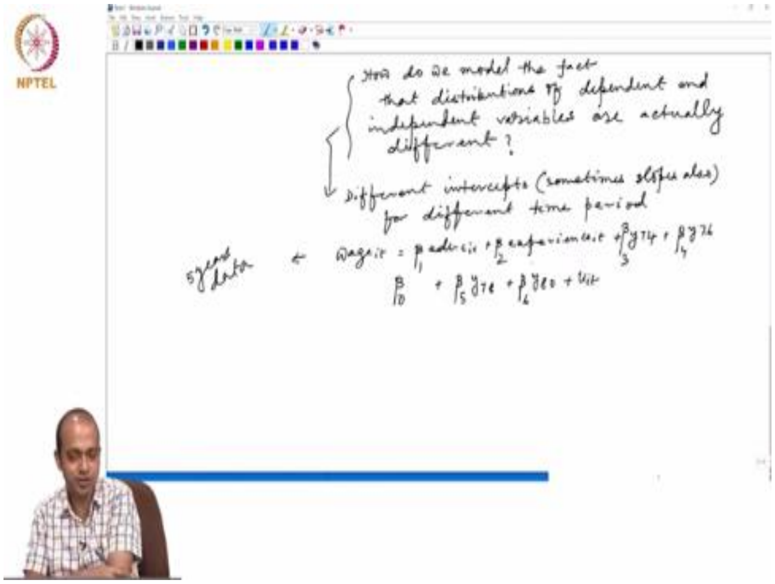
So, that means what we are doing in pool data? Basically, we have several distributions let us say this is the distribution of wage. This is wage let us say 1 this is wage 2 this is 3 this is wage 4 and all these distributions are actually independent to each other. So, pooled data means combining independent cross sections over a period of time. That means the people who are interviewed let us say the data collected on this particular cross section is different from this cross section to this cross section to this cross section.

So, all these cross sections are actually independent to each other, there is no relationship among these distributions. So, you have to keep in mind, when you specifically we are talking about pooled data is pooling independent cross section over a period of time. Let us say this is in 2000 this is let us say 2005 this is 2010 this is 2015. It may not be 5 consecutive years it may or may not be consecutive years.

But one thing we have to keep in mind, pooled data means it is a independent cross section there is no relationship. So, that way we can actually rule out the possibility of these error terms of different, when I am writing it is a pooled data so, it would become it, it. So, since this is independent cross section there is no correlation among the error terms, that we have to keep in mind. So, these are all by pooled data independent cross section.

That is why we rule out the possibility of identical distribution. This distribution may not be identical.

(Refer Slide Time: 20:05)



How do we model the fact that distributions of dependent and independent variables are actually different?

Different intercepts (sometimes slopes also) for different time period

5 year data ← $wage_{it} = \beta_0 + \beta_1 educ_{it} + \beta_2 experience_{it} + \beta_3 y74 + \beta_4 y76 + \beta_5 y78 + \beta_6 y80 + u_{it}$

Now next question is econometrically how do you model the fact that actually we are getting different distributions in different years. So, how do you the next question that we are going to ask how do we model the fact that distributions of dependent and independent variables are actually different? This is the challenge. How do you model econometrically the distributions of dependent and independent variables are a different in different time periods.

Here our knowledge of dummy variables becomes handy. So, what does this different distribution mean different distribution mean actually at different time period. The intercept or sometimes the slopes are actually different. So, that means the answer to this is how do you model this different slope different intercept sometimes loop for different time period. So, if you have a five years data that means we need to include four year dummies to reflect that in each year the intercepts are actually different.

We use four year dummies because one year we use as the base category. As you know from our knowledge of dummy variable if there are five categories then we need to define only 5 - 1 = 4 such damage. So, in this case suppose we have a data on let us say we have a data on wage and education for five years starting from 70 to let us say 76. So, in that case what we need to do? We need to include your dummies four year dummies except for the base year which is let us say 1972.

So, your model would become wage it equals to let us say β_x it or you can write $\beta_1 education_{it}$ + $\beta_2 experience_{it}$ + if you have data on 72 to 76 that means it would be year 72 + β_3 year 72 or let us say 74 + β_4 year 76 + β_5 year 78 + β_6 year 80 1, 2, 3, 4 these four years data and we have used 72 as the base category plus u it. So, we have five years data. So, by adding these five variables we can actually say that, so if you include one specific intercept here let us say this is beta 0.

So, that means that β_0 capturing the average wage of the base category which is 72 here. So, these year dummies inclusion of this year dummies makes it clear that we econometrically model a situation where distribution of wage education and experience they are different in different time periods. Sometimes the coefficient attached with this year dummies they itself they themselves are of specific interest.

For example, we would be knowing whether on an average 9 in 1974 wage is different than 72. On an average wage in 76 is different from 72 on an average 78 wage is different from 72 so and so forth. So, that means these intercepts the other advantage you have added these dummies to recognize the fact the distributions of this independent and dependent variables are different at different time period.

Additional advantage what you are getting by including this year dummies basically you are getting an additional advantage, what is that? You will be knowing on an average 74 wage is different from 72 or not depending on their sign and significance. So, whenever you have this type of pooled data combining cross section and time series you must remember that we need to econometrically model the situation by including your dummies to recognize the fact that our distributions are different.

Otherwise, the modelling itself is wrong. There is no way we can differentiate several distribute different distributions at different time periods.