Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 53
Panel Data Modelling

Hello, everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will start with a new concept that is on Panel Data Modelling. We have already discussed various types of you know econometric models that too in a cross sectional setup and that too in a time series setup. The difference is that you know we have cross sectional data and then we apply regression modelling to the cross sectional data and which is called a type of you know cross sectional modelling, where the sample is a exclusively cross sectional type and the other part is the using time series data that too the structure is called as a time series modelling.

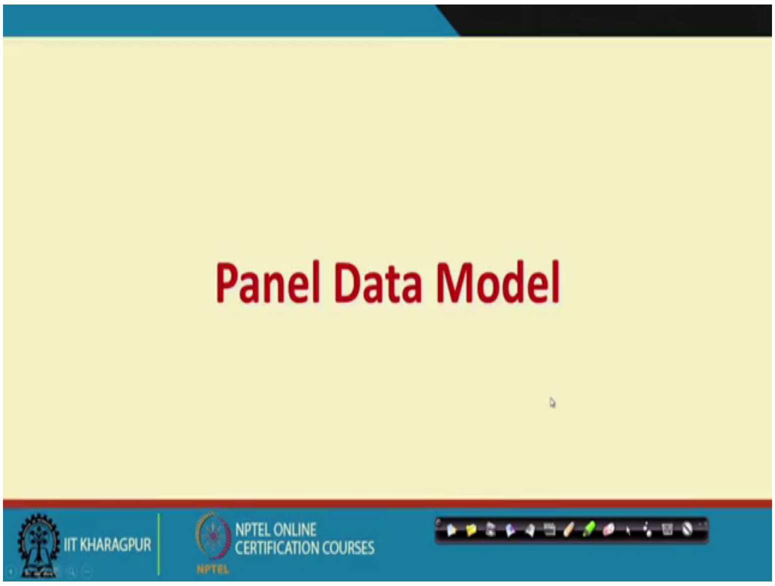
Now, in the real life scenario we have a couple of problems where the sample size may be very small and that too with actually with respect to either cross sectional data or with respect to time series data. So, now, we like to discuss a concept here where we can do the pooling; that means, bringing both cross sectional data and time series data and then you know have a kind of you know new data structure or the kind of you know new data set, where by default we will have a more and more sample.

Because, we are clubbing time series data with you know cross sectional data. The same engineering problem and same kind of you know modelling that too the kind of you know simple regression modelling. And, then when we use this kind of you know data where we are pooling the time and cross then we will have a kind of you know new data structure that is what simply called as you know panel data structure.

So, when we are pooling the data time with cross sectional types, then in the first end we call it a pool data and on the on the basis of you know pooled data we try to integrate some kind of you know specifications and that too there is a chance of you know structural kind of you know difference and when we bring that kind of you know structural difference then we call as you know panel data model.

So, now let us see what is exactly the that particular concept and how we can use this panel data models to solve some of the engineering problems and as per the decision making you know requirement. So, let us see how is the kind of you know structure.

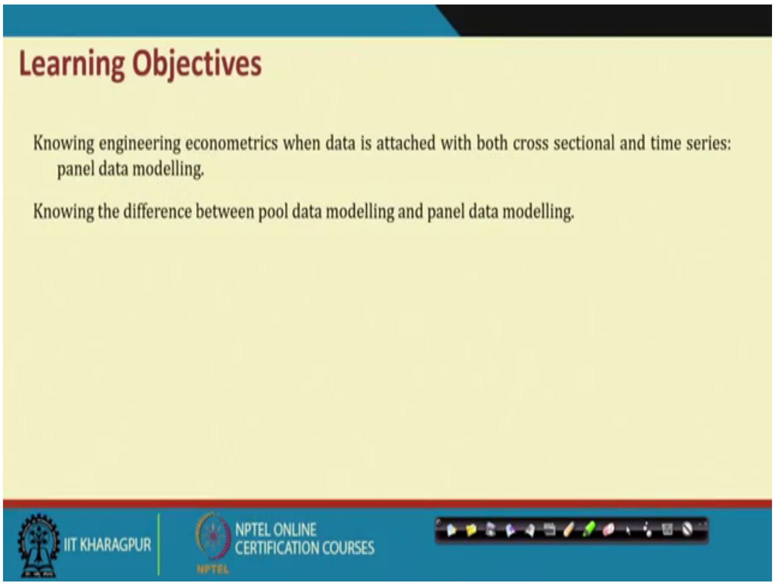
(Refer Slide Time: 03:44)



Panel Data Model

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

(Refer Slide Time: 03:46)



Learning Objectives

Knowing engineering econometrics when data is attached with both cross sectional and time series: panel data modelling.

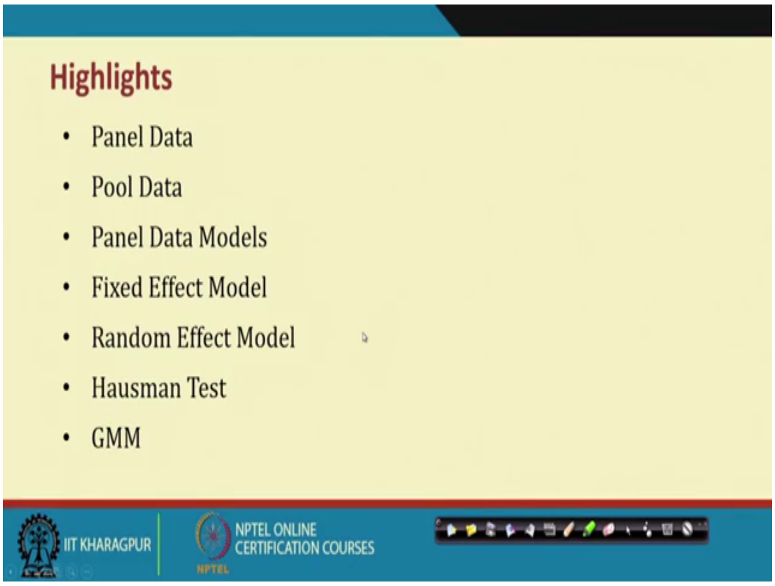
Knowing the difference between pool data modelling and panel data modelling.

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

So, first structure is that you know to know the data structure that too with respect to pool data and panel data. And, the objective is a to you know solve some of the engineering problems by using the time series data as well as cross sectional data. Since we are pooling time series data with a cross sectional data by default it is a kind of you know different structure, so, by default some kind of you know difference will be there in the modelling setup. So, we like to check what are this difference and how we have to bring the kind of you know you know concept. So, that we can analyse the engineering problems more accurately and that too as per the requirement of a kind of you know a situation where you know initially we may have some kind of you know scarcity of you know samples because, of you know non availability of data that too in a time series setup or in a kind of you know cross sectional setup.

So, the advantage is that you know when will pooling time and cross then by default sample size will be high and then we can address the problem more accurately. So, let us first understand what is exactly the concept and then we will go ahead with the kind of you know discussion.

(Refer Slide Time: 05:16)



Highlights

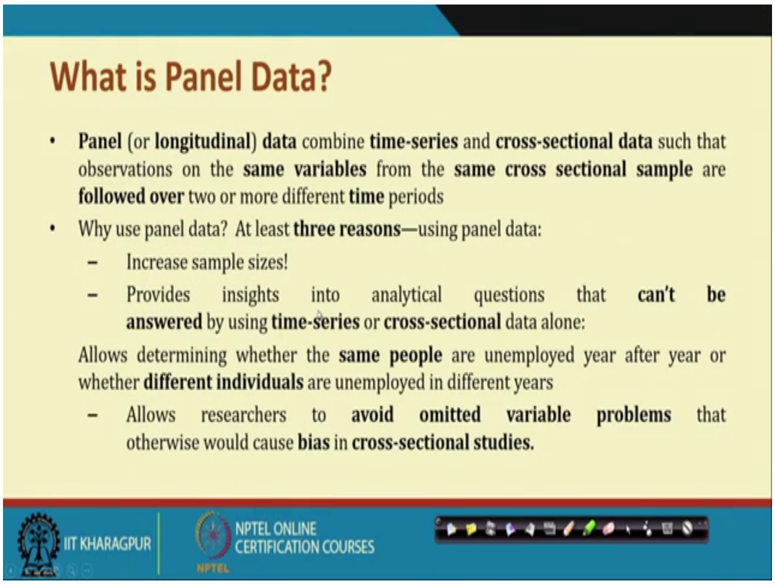
- Panel Data
- Pool Data
- Panel Data Models
- Fixed Effect Model
- Random Effect Model
- Hausman Test
- GMM

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, in this particular you know lecture we are exclusively going to discuss some of the concept like you know panel data, then pool data various panel data models and that too we specifically cover fixed effect model, random effect model and generalized methods of you know moments.

(Refer Slide Time: 05:45)



What is Panel Data?

- **Panel** (or **longitudinal**) **data** combine **time-series** and **cross-sectional data** such that observations on the **same variables** from the **same cross sectional sample** are **followed over** two or more different **time** periods
- Why use panel data? At least **three reasons**—using panel data:
  - Increase sample sizes!
  - Provides insights into analytical questions that **can't be answered** by using **time-series** or **cross-sectional** data alone:

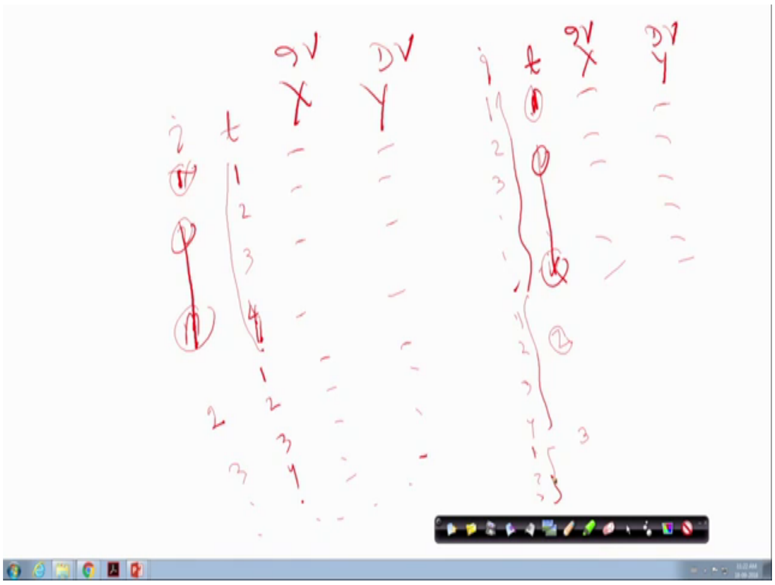
  Allows determining whether the **same people** are unemployed year after year or whether **different individuals** are unemployed in different years
  - Allows researchers to **avoid omitted variable problems** that otherwise would cause **bias** in **cross-sectional studies.**

So, the starting is that you know to understand the panel data first. So, panel data combine a simply time series with cross-sectional such that you know the observation on the same variables from the same cross-sectional sample are followed over two or more different time periods. Or the this structure maybe is you know someway different with you know you know same time series data with two or more different cross-sectional you know situations.

So, that is what you know either way. So, that means, what we like to do here; so, we have cross sectional structure and time series structure symbolically we can put you know i as a you know representative to cross sectional setup and t as a representative to time series setup. So, that means, technically for every variables when we report this sample if you know keep constant t and allow i vary that that simple presentation is called as you know cross sectional you know type. And again keeping i remain constant for a particular cross sectional you know situation then allow t to vary then that particular structure is called as a time series you know setup.

And, again i will vary and t will vary means they go to you know they vary together then that particular structure is called as you know pool data and panel data. So, now so obviously, this structure is a simply like this. So, let me bring the kind of you know situation here. So, the situation will be here.

(Refer Slide Time: 07:30)



So, let us say X is a variables then Y is a another variable. So, this is what called as a dependent variable, this is called as you know independent variable for any engineering problems. So, what will you do? So, we the first hand requirement of engineering econometrics, you know as per you know any models or you know as per any kind of you know techniques. So, we have a specification about i and t. So, what will you do here, so, let us i is a constant here 1 and t allowed to vary 2, 3, 4 like this, ok. So, then obviously, there will be some kind of you know entry to both X and Y; obviously, the entry should be uniform as per the you know cross you know the regression analysis requirement.

So, here keeping i equal to 1 and allow t to vary and not specifically i exactly equal to 1 either it is fixed at 1 or if not it is fixed at 2 then it they the t will vary or it will be fixed at you know say 3 or 4 and 5 something like that. So, let us they fixed at a point of you know m then allow t to vary. The other type of you know structure is here again putting actually i and t and then same X and Y. So, this is independent variables and this is dependent variable what we supposed to do here. We supposed to put you know i you know t is constant allow i to vary.

So, that means, let us take t constant then i allowed to vary like this. So, in this case better to put you know t left hand side and i in the right hand side. So, the variation can be more appropriate. So, again either this will be constant or t this will be constant or you know it will be up to you know k you know kind of you know specification. So, here you know we have k time period particular time period and then the sample will vary like this and obviously, the entry will be uniform entry will be you know with respect to X and Y.

So, this what actually the simple understanding about the you know panel data. So, that means, you know if you allow i constant and t vary that structure is called as a time series structure and keeping t constant i will vary that is called as you know cross sectional structure. And, what will you do here just you know let us say this is a remove this once then allow i equal to 1 here then t up to let us say 5 and again allow i equal to 2 then again we can start with 1, 2, 3, 4 and by default the entry will be like this say again 3 it will continue like this, ok.

So, similarly here 1, 2, 3, 4 like this and allow this you know 1 will be remain constant then time period t again there are couple of samples and time period 3 then again a couple of sample periods like this we have to just you know bring into a particular you know structure, where by defaults the sample size will be you know means it will be you know in any kind of you know increasing trend.

So, what is happening if you follow this ones or this ones then this particular structure simply called as you know pool data structure and then we like to study the impact of i and t by default that represents the panel data structure. So, first you bring the you know data and then we try to you know study the specification with respect to i and t of course, the basic setup will be with respect to dependent variable and independent variable. But, in between the i specification and t specification can be actually a you know huge and then we try to explore the particular impact of i and t that too the that too in the relationship between a you know independent variable and dependent variable.

So, this is what the a simple understanding of you know panel data structure. So, let us come to this you know situation. So, here whatever I have already mentioned so, the first understanding is the pooling the time series data and cross sectional data and create a new data structure that is either called pool data or panel data. So, far as a data is concerned there is not much difference between pool and panel. So, both are in the first instance same, but in the case of you know panel data it is something more than you know whatever you know we have already you know pooled. So, that means, technically on the top of pool data we try to explore the impact of you know cross sectional unit and time series unit while linking dependent variable with independent variable, that is what the basic you know signal about the pool data panel data.

So, now, the first question is a why we need panel data of course, I have already highlighted couple of things let us you know brief about this particular you know requirement. So, in the first instance there are at least you know three specific reasons why we need actually panel data. First point that too the requirement of you know increasing sample size, and second it can provide you know better insights into analytical questions that cannot be answered by using either time series data or cross sectional data, because some variations with respect to cross sectional type and some variation with respect to time series type is highly required to you know address the particular you know engineering problem that too while linking the cause and effect relationship you know between a some of the engineering problems.

And, in fact, you know it can allows it can allowed to determine whether this same people are you know unemployed year after year or whatever different individuals are unemployed in different years; that means, technically these are the deals you know we are supposed to actually address. And, then explore you know their impact and the kind of you know influence through the help of you know pooling both time series data and cross sectional data.

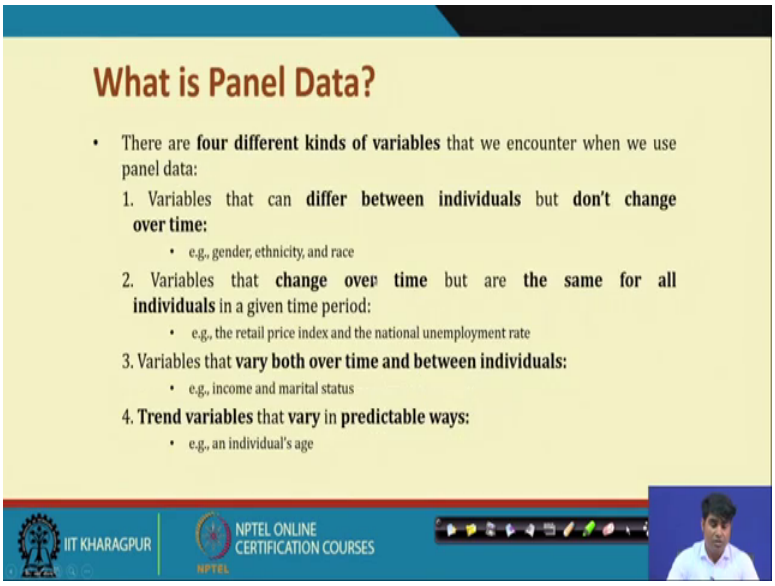
So, it is not you know something you know only to increase the sample size, but it will give you much kind of you know you know you know importance while addressing the problem because there is a two types of you know variations which we are observing here while you know addressing some of the engineering problems. So, it can allow researchers to avoid omitted variable problems that otherwise would create you know bias in the cross sectional setup.

Sometimes you know while linking time series data. So, missing observation will be there and while addressing again cross sectional data some missing observation will be there. So, now, while you know doing the pool and you know using the panel so, we are just you know clubbing both the kind of you know situations. As a result somehow the missing variables or the omitted variable bias will not be actually so you know so, you know so, you know I mean it is kind of you know a problematic while addressing the you know engineering problems.

So, that is what the advantage of this particular you know panel data and that too panel data modelling, that is what the beauty of this particular you know you know you know data structure. So, here you know the issue is that you know while addressing time series type there is any kind of you know missing that may be you know bring some kind of you know issue and again same thing happens while you know addressing the cross sectional data.

So, panel data will remove all this obstacles and reduce this bias in the data structures while addressing some of the engineering problems. By the way so, there are two specific advantage through which we can connect actually pool data or panel data; first the increase sample size, second we can you know address the problem more appropriately by allowing the information to vary with respect to time and you know cross sectional setups. So, that is what the kind of you know deal.

(Refer Slide Time: 16:59)



What is Panel Data?

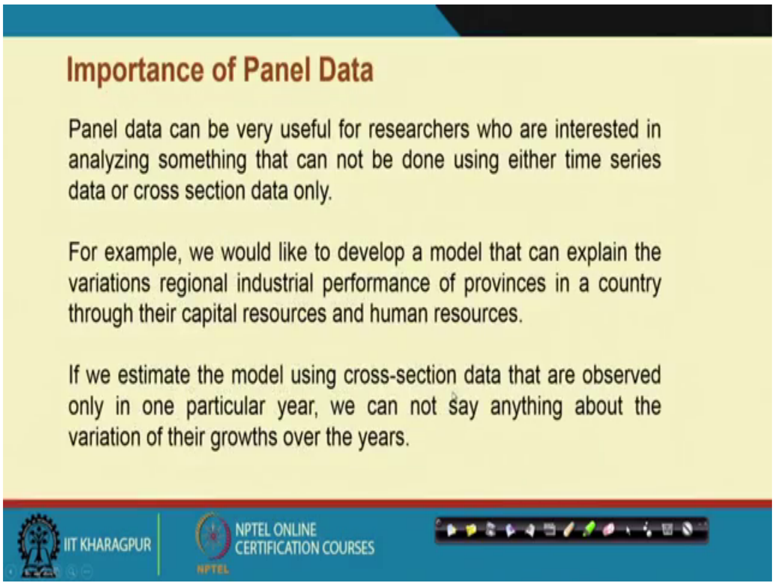
- There are **four different kinds of variables** that we encounter when we use panel data:
  1. Variables that can **differ between individuals** but **don't change over time:**
     - e.g., gender, ethnicity, and race
  2. Variables that **change over time** but are **the same for all individuals** in a given time period:
     - e.g., the retail price index and the national unemployment rate
  3. Variables that **vary both over time and between individuals:**
     - e.g., income and marital status
  4. **Trend variables** that **vary** in **predictable ways:**
     - e.g., an individual's age

And then obviously, question is why exactly you know this panel data. Again, there are four different kinds of you know variable that we usually use in the panel data structure. So, first one variables that can differ between individuals, but do not change over time, for instance you know race, religions something like that. And, second a variables that change over time, just opposite, but they, but they are same for all individuals in a given period of time. That is you know like say retail price index then unemployment rate these are the classic examples which can come under this group.

Then, third one is the variables that vary with over time and you know between cross sectional you know units that is income, marital status something like that and the fourth one is trend variables that vary in predictable ways. For instance individuals age, then individuals salary. So, these are things which will you know vary in a kind of you know predictive you know way.

So, the because of these reasons so, we like to actually use panel data to address some of the engineering problems. So, that means, technically we have we have the idea about the cross sectional data and then the cross sectional modelling then time series data and time series modelling, now we have a different kind of you know flavour by having a pool data and panel data again we have the concept of you know called panel data modelling.

(Refer Slide Time: 18:59)



Importance of Panel Data

Panel data can be very useful for researchers who are interested in analyzing something that can not be done using either time series data or cross section data only.

For example, we would like to develop a model that can explain the variations regional industrial performance of provinces in a country through their capital resources and human resources.

If we estimate the model using cross-section data that are observed only in one particular year, we can not say anything about the variation of their growths over the years.

So, now after knowing the concept of you know panel data, so, we like to just highlight what is the advantage of this you know panel data. So, that means, basically we like to bring some of the importance. Of course, we have already highlighted the importance or still you know we once again you know bring the situation, so that you know we can actually justify that you know panel data is a better kind of you know structure. And, you know better you know kind of you know modelling setup through which you can simplify and solve some of the engineering problem.

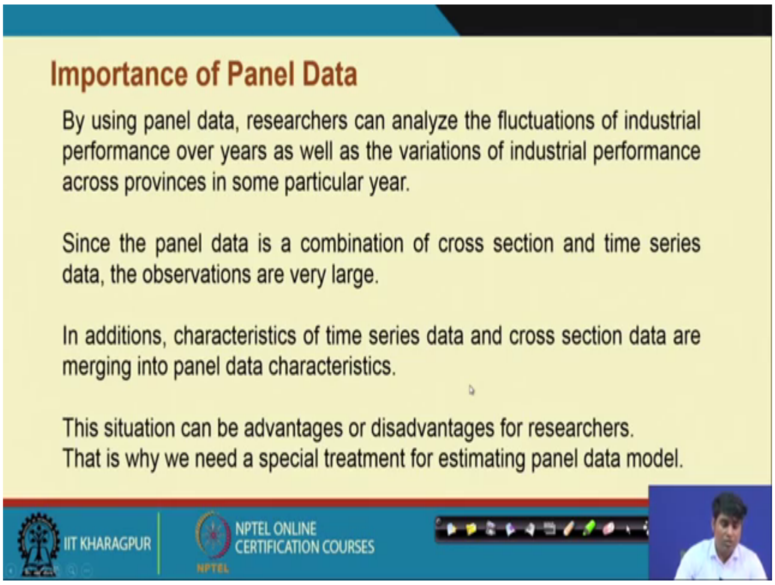
So, panel data can be very useful for researchers who are actually interested in analysing something that cannot be done either by the a by the use of you know time series data or a cross sectional data means only means either through time series data or though cross sectional data. So, that means, you are bound to take both the things simultaneously big you know there, maybe you know two specific reasons for that. First one it is a low sample size and second you know the variation may not be so appropriate if you consider either cross sectional type or you know time series types.

So, you allow both time series to you know vary and then cross sectional to vary. So, that means, you will find much kind of you know randomisations while addressing some of the engineering problems. For instance, we would like to develop a model that can explain the you know variations; that means, regional variations with respect industrial performance of a regions and that too for any country and again that with respect to the variations you know with respect to their you know research base maybe capital resource base or you know human resource base.

In this situation, if you like to estimate the model using cross sectional data that are observed only in one particular year, but in that sense we cannot say anything about the variations of their growth over the years. The other side of the game is also similar. If you use only time series data then you know you can just check the variations you know of these variables over the times and then we are ignoring the cross sectional you know flavour. So, that means, technically if you use time series data then you have to you have to ignore the cross sectional variation and then when you use cross sectional data you are you know ignoring the time series variations.

Now, what is happening if you pool and you know then use pool data and panel data both variations can go simultaneously and that in my opinion that is more appropriate and you know may not be completely, but it is a less biased compared to either you know use of time series data or you know cross sectional data.

(Refer Slide Time: 22:26)



Importance of Panel Data

By using panel data, researchers can analyze the fluctuations of industrial performance over years as well as the variations of industrial performance across provinces in some particular year.

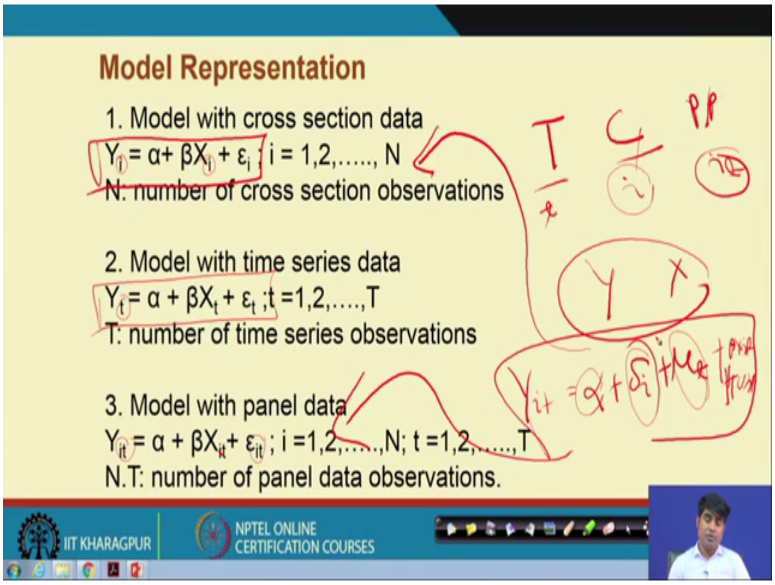
Since the panel data is a combination of cross section and time series data, the observations are very large.

In additions, characteristics of time series data and cross section data are merging into panel data characteristics.

This situation can be advantages or disadvantages for researchers.
That is why we need a special treatment for estimating panel data model.

So, likewise you know there are many different you know reasons you have to find out through which you can actually justify the importance of you know panel data. Of course, there is a some advantage and there is also some disadvantage, but still on the top of the panel data modelling or the use of panel data or pool data is always you know good to handle some of the engineering problems in order to have more sample size and that too in order to have a more variation in the data and that will be more accurate kind of you know scenario where we can get better inference and that too for the decision making requirement.

(Refer Slide Time: 23:16)



Model Representation

1. Model with cross section data
$Y_i = \alpha + \beta X_i + \varepsilon_i$; $i = 1,2,\ldots.., N$
N: number of cross section observations

2. Model with time series data
$Y_t = \alpha + \beta X_t + \varepsilon_t$; $t = 1,2,\ldots.,T$
T: number of time series observations

3. Model with panel data
$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it}$; $i = 1,2,\ldots..,N$; $t = 1,2,\ldots..,T$
N.T: number of panel data observations.

T
t
C
i
PP
it
Y X
$Y_{it} = \alpha + \delta_i + \mu_t + [illegible]$
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

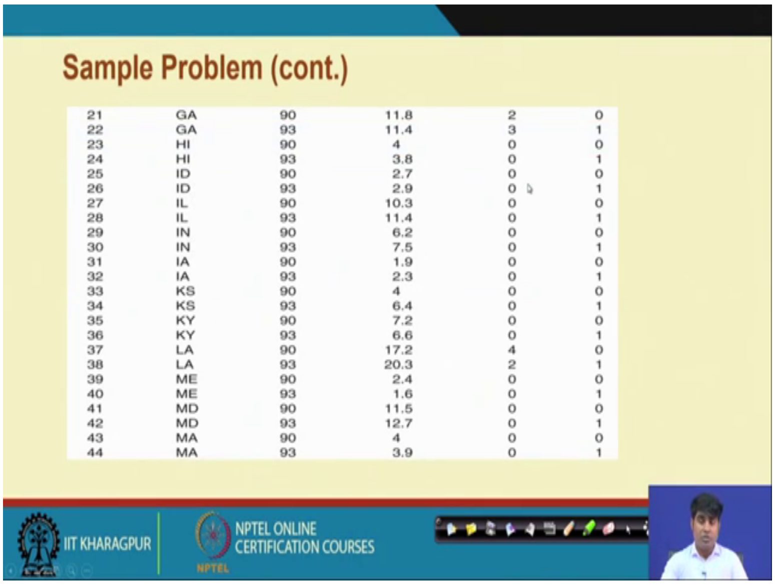
So, in so, coming to the modelling sides, so we have here actually three different setups altogether that too we have the situation called as you know we have the situation called as a you know what we can called as you know the situation what we can called as you know time series cross sectional or pool or you know panels, right.

So, see here so, this is t specification, this is i specification and this is i t specification. So, as a result so, the same the game is with respect to Y and x. So, now, how we can model you know how we can actually differentiate the model with respect to data structure. So, in that case we are linking Y and X then bringing a kind of you know mathematical model like this Y equal to alpha plus beta X and then error term and we are using i i subscript to ensure that you know the model is actually analysed with respect to cross sectional data because we specify here I represent cross sectional data.

And, again in the in the in the second model the same you know mathematical model, but instead of i we are putting actually t subscript so, which represents the use of you know time series data while doing the estimations and the doing the kind of you know analysis. And, in the third models instead of you know i t we use i t together in the subscript both in the case of you know Y and X and again that with respect to the error term. And this particular structure technically called as you know either pool data or panel data. Of course, it is under the pool data again if you add two more you know extra i terms to this models by giving the specification of you know i and t then the particular structure will be called as you know you know panel data.

So, that means, in that case I can write the equation like this Y i t equal to let us say alpha plus delta i plus mu t and then beta X i t and plus u i t. So, that is that could be the one of the structure of the panel data compared to the model you know representing the pool data structure. So, alpha by default will be intercept and delta i and mu t. So, this is actually for cross sectional impact and this will be with respect to time series impact. So, the positioning you know of the samples with respect to cross sectional and time series sometimes you know you know sometimes matter a lot while addressing some of the engineering problems and that too for the decision making requirement.

(Refer Slide Time: 26:42)

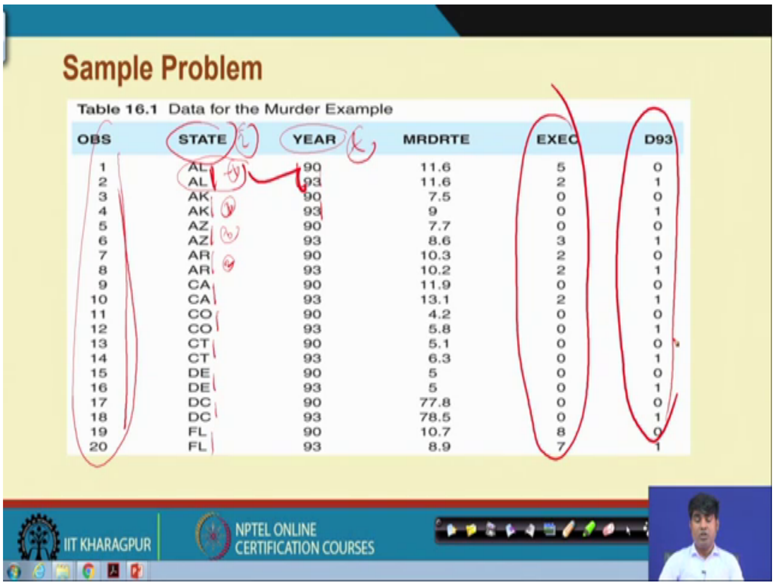


Sample Problem (cont.)

21	GA	90	11.8	2	0
22	GA	93	11.4	3	1
23	HI	90	4	0	0
24	HI	93	3.8	0	1
25	ID	90	2.7	0	0
26	ID	93	2.9	0	1
27	IL	90	10.3	0	0
28	IL	93	11.4	0	1
29	IN	90	6.2	0	0
30	IN	93	7.5	0	1
31	IA	90	1.9	0	0
32	IA	93	2.3	0	1
33	KS	90	4	0	0
34	KS	93	6.4	0	1
35	KY	90	7.2	0	0
36	KY	93	6.6	0	1
37	LA	90	17.2	4	0
38	LA	93	20.3	2	1
39	ME	90	2.4	0	0
40	ME	93	1.6	0	1
41	MD	90	11.5	0	0
42	MD	93	12.7	0	1
43	MA	90	4	0	0
44	MA	93	3.9	0	1

So, obviously, so, this is what the kind of you know structure through which you can understand the modelling about the basic kind of you know you know data structure and the kind of you know panel data modelling.

(Refer Slide Time: 26:46)



Sample Problem

Table 16.1 Data for the Murder Example

OBS	STATE	YEAR	MRDRTE	EXEC	D93
1	AL	90	11.6	5	0
2	AL	93	11.6	2	1
3	AK	90	7.5	0	0
4	AK	93	9	0	1
5	AZ	90	7.7	0	0
6	AZ	93	8.6	3	1
7	AR	90	10.3	2	0
8	AR	93	10.2	2	1
9	CA	90	11.9	0	0
10	CA	93	13.1	2	1
11	CO	90	4.2	0	0
12	CO	93	5.8	0	1
13	CT	90	5.1	0	0
14	CT	93	6.3	0	1
15	DE	90	5	0	0
16	DE	93	5	0	1
17	DC	90	77.8	0	0
18	DC	93	78.5	0	1
19	FL	90	10.7	8	0
20	FL	93	8.9	7	1

And, to understand the you know clear cut view about the panel data or pool data so, I have already highlighted. So, with starting with you know i representation and t representation. So, in this you know in this case we are bringing one you know life problems where you know we have observation. So, this is what the observation; that means, the excels sheet will read like this for the panel data. So, this is what the sample observations so, which is also there in the cross sectional type and the time series type, but still we are bringing here to know what is the exact sample size at means end of the day whether it is cross sectional type or time series type or pool type. Ultimately the number of data points will the matter, rest will follow as per the particular you know model requirements.

Now, here we are putting status the cross sectional you know type and year is with respect to you know time series type. So, we will find the cross sectional types are you know having lots of variations. So, for instance AL AL; so, this is actually one cross sectional unit then AK AK this is another cross sectional unit AZ AZ that is third cross sectional unit so, different states in a country. So, similarly AR AR another you know cross sectional unit CA, CO, CT, DE, DC, FL like this. So, the different cross sectional units and again in the time series sides we have 90 93 again we have 90 93. So, that means, if the cross sectional setup here one cross sectional setup then that is with respect to you know time 90 and 93.

So, that means, here two cross sectional units in a particular you know state and then we have two different data that means, technically. So, for a particular state we have two time series data and again keeping time constant we have a different cross sectional data. So, here you know i this is represent representation of i and this is representation of t. So, i varies with respect to all these you know cross sectional variations and t varies with respect to 90 and 93; that means, we have two different time periods and we have couple of you know cross sectional you know types.

So, then we are you know just bringing together and then make a kind of you know data structure which we call as pool data and then we give some specification and finally, we can call panel data. So, this is how the kind of you know structure. So, this is how the specification about the kind of you know cross sectional and that is with respect to actually a time. So, that means, technically we can create two different dummies in the linking between Y and X that is what we declared earlier delta i and mu j. So, that is called as you know dummy representation only.

If there actually two cross sectional then delta will vary from you know delta 1 to delta 2 or you can put you know delta is the symbol then the sample variation will be 0 1 or 1 2 kind of you know scenario. Similarly, if the time series data will vary so, we like to check how many variations are there. So, keeping time so, we can give specification you know since in this particular data set we have two different times. So, we can start with 0 1 or you can put 1 2. So, 1 represents for the year 90 and 93 represents the kind of you know 2 or you can put you know 0 for 90 and 1 for 93.

So, now after having this kind of you know you know excel spreadsheet. So, we can classify actually two different you know dataset; one with respect to let us set that is with reference to time one with respect to 90 and another with respect to say 93. Then finally, we are pooling this two, then it becomes a called as you know big data and then pool data and the panel data. So, likewise we have here actually couple of samples. So, we this is 20 different observations and that too with both cross sectional and time and if you only consider 90 then by default the sample size technically will reduce to 10. And, again if you allow you know time to vary and you specify only one cross sectional then ultimately two data points are having.

For instance you specify for the state like AL only then only two data points are there which is actually not feasible to you know analyse some problem even if it is a bivariate structure because, two sample size cannot be used for the kind of you know model estimations or the kind of you know model validation. So, we need actually more sample. So, now, with this particular you know data structure if you do not allow pool and panel then the analysis will not be very effective and even if you can get the estimated output. So, the reliability of the output will be question mark and even if it will pass through reliability, but we cannot just generalize because of you know very very low sample size.

So, to generalize any kind of you know output for the future requirements like you know prediction and forecasting, so, the particular you know inference should be available from the large you know dataset and again allowing with it time to vary and cross sectional to vary. Because that will give you less bias and more authentic and more variations to the data set up through which you can analyse the engineering problem more accurately and more appropriately. So, how will it go about this particular you know structure and bring the kind of you know variations and kind of you know estimation we will discuss in the next lecture. So, we will stop here today.

Thank you very much. Have a nice day.