Handling Large Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 36
Introduction to Panel Data

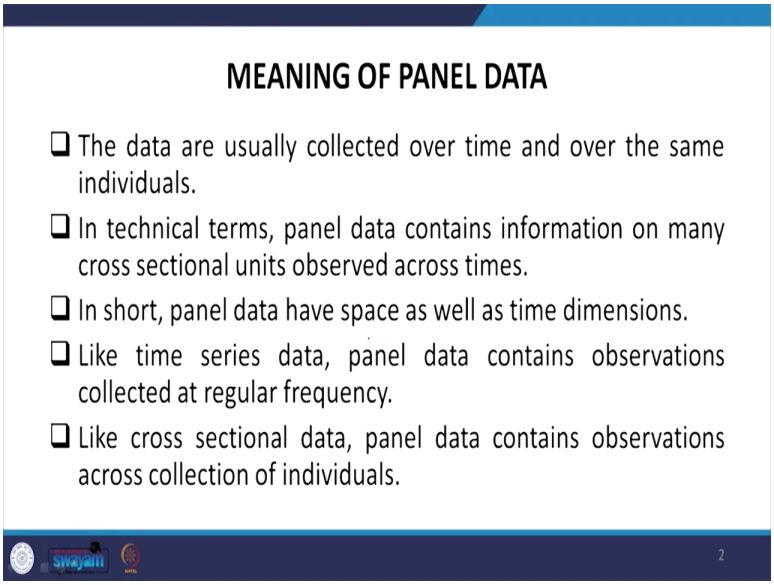
Welcome once again to the NPTEL MOOC module on Handling Large Scale Unit Level Data with STATA. We are here at the verge of the penultimate section of understanding Stata operations for the unit level data. So, this is purely dedicated to panel data understanding and estimation through Stata. And the previous sections are largely on understanding Stata through cross-sectional approaches also on qualitative variable models.

And even from the beginning of the lectures if you recaptulate that was largely a discussion on the understanding of unit level data, then sampling design to some extent where we can connect to the unit level data and Stata. And then we started with cross-sectional approaches of understanding Stata, then qualitative dependent variable and independent variable models. The last week is dedicated for the panel models.

Since panel is very important, especially for the scholars, for the researchers who deal with the information for over a time period or over invariant or variant observations, or in variables, in that case making panel or observing from panel data is very important. Why I am saying researchers do need a panel data especially for publications in the topmost channel, because this has contents which can be connected to a larger cohort, a larger sections and because of the fact that it either through appending or through merging, we can able to deal with both the sections, cross-sectional issues as well as over time issues.

And when over time aspects are added there are possibility of drawing better inferences, especially for policy purposes. So, panel is very very important. In this lecture, I am going to discuss what is panel data, why panel data is necessary, what are the different types of panel data, and how to understand whether the data is panel or not? And if it is panel then how we can handle with our Stata package for understanding or interpretation of the data.

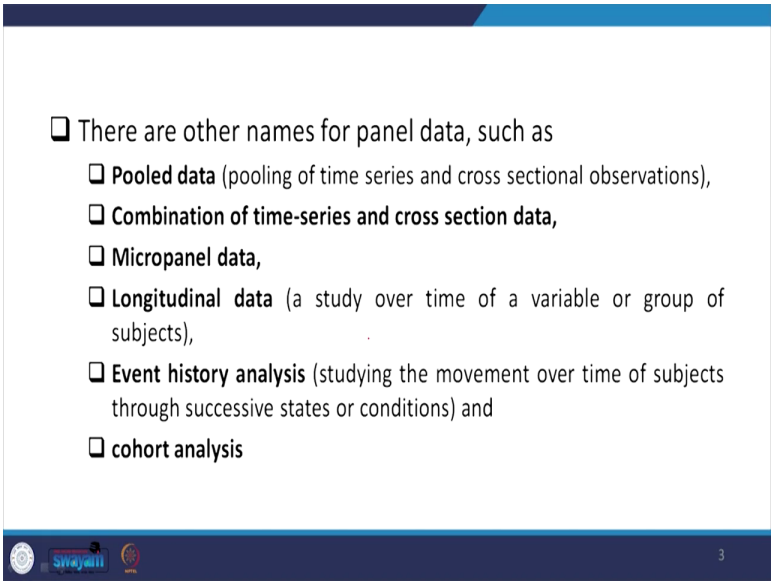
(Refer Slide Time: 03:33)



So, coming to the meaning of panel data. Panel data usually collected over time and over the same individuals. The individuals which is represented over different time and their responses over time is discussed in the panel data approach. In technical terms, panel data contains information on many cross-sectional units observed across time.

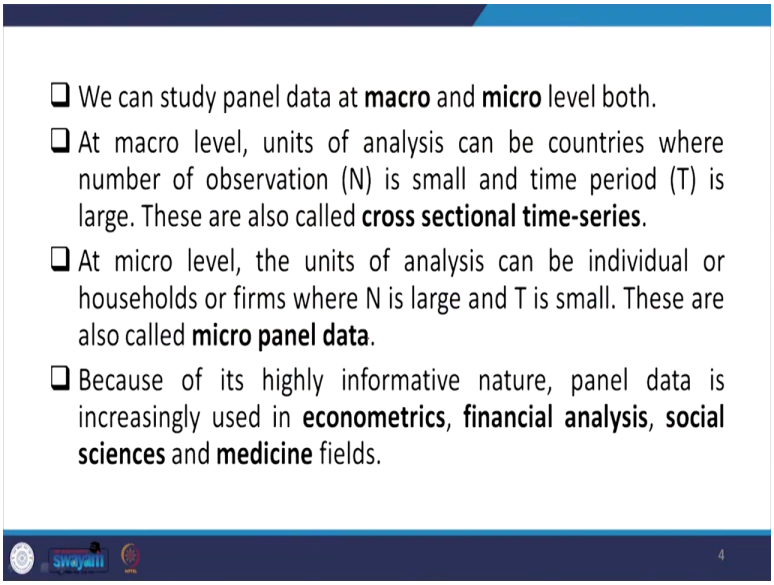
So, in short, panel data have space as well as time dimensions. So, space we mean the cross-sectional aspects as well as over time. So, like time series data, panel data contains observations collected at regular frequency or regular timing. So, like cross-sectional data, panel data contains observations across collection of individuals as well. So, it is an overlapping of cross-sectional and time series.

(Refer Slide Time: 04:29)



There are other names of panel data as well. What are those? Those are called pooled data. Pooled, basically pooling of time series and cross-sectional observations and we will clarify all those in our successive slides. Then this is also called combination of time series and cross-sectional data. Micropanel data, we are going to clarify, whether micro or macro, longitudinal data. So, generally, the time aspect is very variable in case of longitudinal data. And another one is called event history analysis, especially when we have some study on the movement over time of subjects through successive states or conditions is called event history. And the last part of the panel data analysis or names of panel data analysis is called cohort analysis as well.

(Refer Slide Time: 05:24)

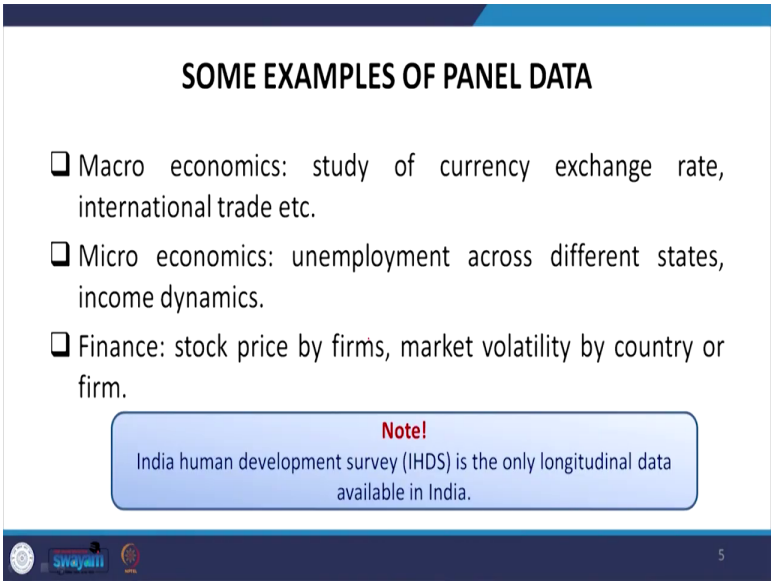


So, we can study panel data at macro as and well as micro level units. At macro level what is generally in aggregate sense like units of analysis can be countries where number of observations that is N is small and time period is large. That is called macro panel data. And so in case of macro, T, the time period, basically the countries, usually number of countries are limited. And if the time for each country over, maybe 15 years, 20 years and within the year there are quarters or within quarter there are months, and then within months also there are different time period. So, large time, if time exceeds than that of the N, number of observations, if that exceeds then that type of panel is called cross-sectional time series data. That is simply called cross-sectional time series data.

At micro level, the units of analysis could be individual or households or firm where N is large, especially the number of observation is large, whereas the T is small. When it is very micro-based study the number of cases to be observed is more important than that of the time. So, it is not just N that matters here. Had it been only N, we could have mentioned that is called cross-sectional data. But a T is added, but T varies very less or the number of time period that is T is very small that type of data is called micro panel data.

So, I am clarifying another aspects like those things I am clarifying through example as well. Because of its highly informative nature panel data is increasingly used in econometrics, financial analysis, social sciences and medicine field etc.

(Refer Slide Time: 07:23)



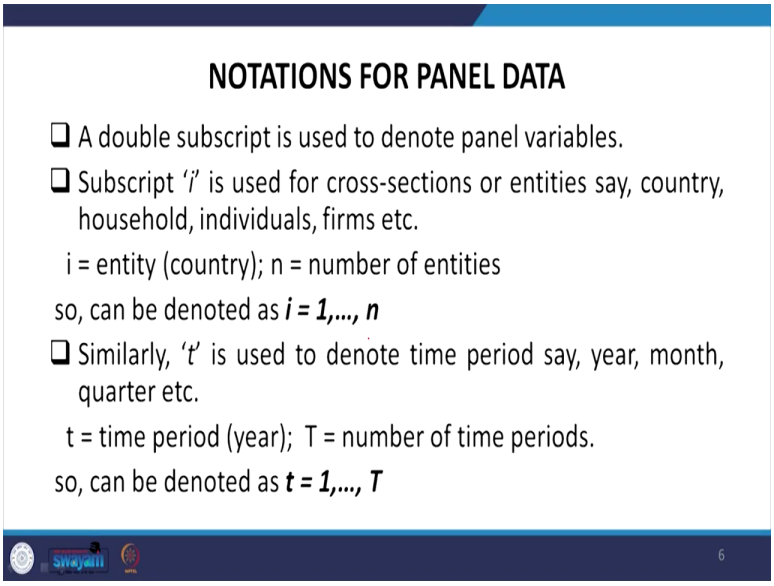
In macroeconomics we are citing some examples for it. In macroeconomics like for stock market, currency exchange rate, international trade or monetary units any aspects we generally give those in aggregate are called macro-based panel data. Whereas microeconomics we generally deal with so far on employment, like we have already cited for PLFS data, Periodic Labor Force Survey data. Those come up with panel very recently, usually not in the early period, but in the recent from 2017-18 to 2018-19 and to period even within one field of survey also panel contents are there. We will be clarifying which kind of panel it contains through our operations.

Also we have panel data of microeconomics type is through IHDS, India Human Development Survey also. If you just wanted to have other panel from microeconomic context if you look at young life data, young life survey. That contains data of four countries in the world. So, that is also based on purely a panel model. Similarly, even if the data is not panel we have some techniques to make it panel, how to make it and not, in our successive lectures somehow will guide you.

Coming to finance aspects, example stock price by firms, market volatility by country or firm etc. are example of panel data. So, we have given in the note IHDS is the only longitudinal data available in India. So far as a clear cut cross-sectional coverage of different aspects of human development is concerned is very much available in the IHDS data and that is also a panel.

Question again occurs which kind of panel, balanced or unbalanced we are going to clarify very shortly.

(Refer Slide Time: 09:41)

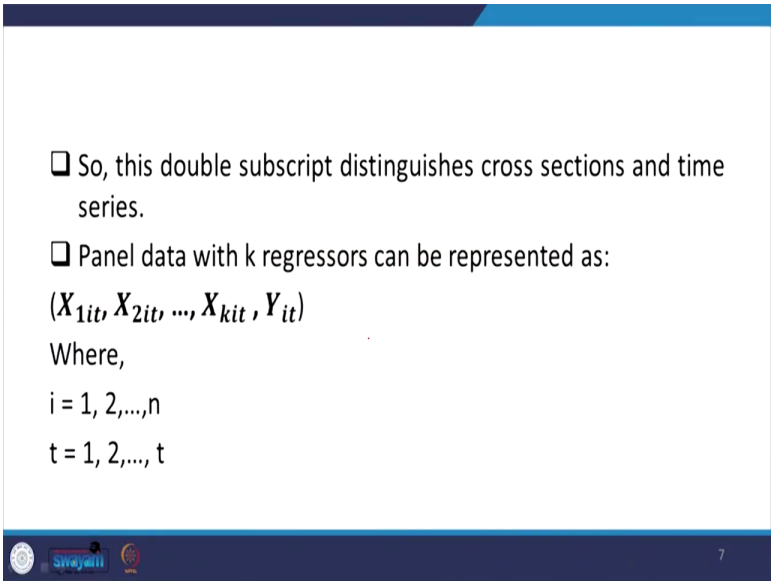


Coming to the notations for panel data. Like just understanding panel is not enough. We need to focus on what is the crux behind panel, what is the rationale behind panel, what are the model behind panel, why we are concerned about panel? We need to understand through some theoretical or equation-based approach in order to have a clear start for the panel model.

So, we are clarifying one by one. A double subscript is used in case of cross-sectional since time was not there, this subscript is only i, jth component is not there or t component is not there. So, here double subscript is one for the cross-sectional observation and another is for the time component. So, subscript i is used for cross-sections or entities, say maybe country, maybe household, may be individual or firm, whereas in our example we will discuss about country or firm. And n in this case is number of entities.

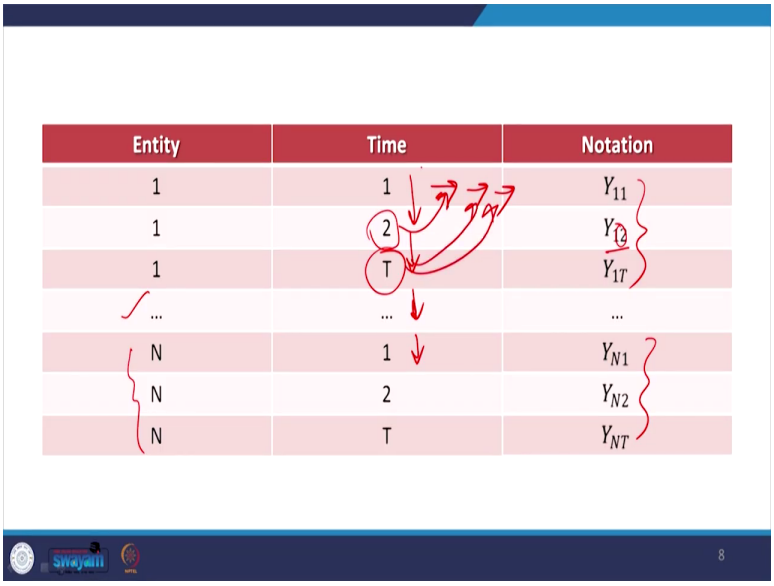
So, this can be noted as i varies from 1 to n if we are citing number of countries to be n. Similarly, t denotes time period like year, month, quarter, whatever, but t, small t we are counting for the time period specifically and total T, capital T is identifying the number of time periods in total. So, it spans from 1 from capital T.

(Refer Slide Time: 11:19)



Coming to the discussion of double subscripts once again. This distinguishes cross-section and time series data. It is amalgamation usually. Panel data with k regressors, we are emphasizing the regressors, can be represented here as first X1 of ith individual for t is the time period, then this is the second time period, second regressor till kth regression if you are interested to estimate. Then our impact on the variable we can find out through Y on X.

(Refer Slide Time: 12:03)

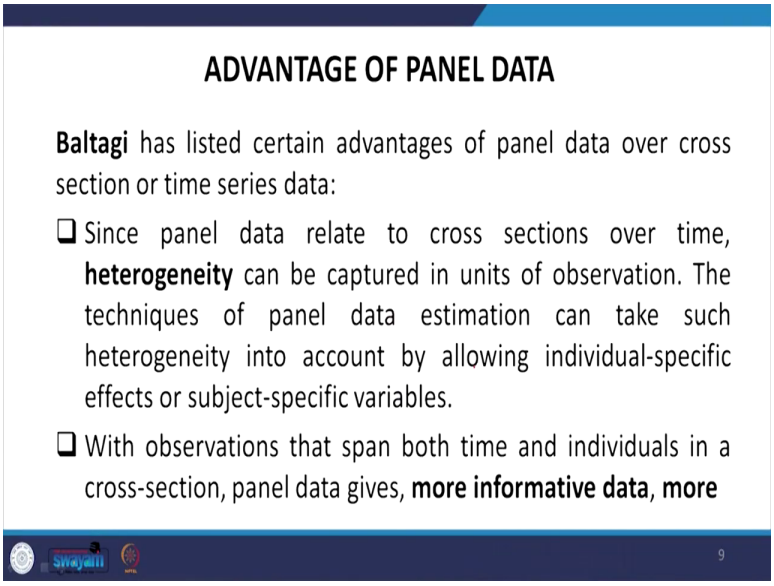


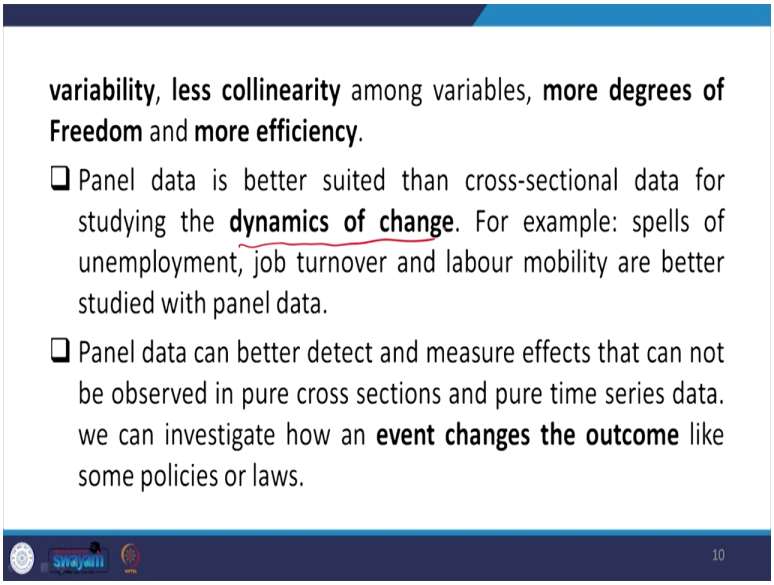
So, we have already clarified subscripts, we are presenting here a lucid picture of panel, like entities or N we wanted to mention through here in our first column that represent the N. N, suppose the first individual, maybe individual, maybe firm, first or may be country, if the data is presented, here we are mentioning time. Time first period, then here it is second period then till Tth period. The notation here for the dependent variable or for the regressor is as Y_{11} because for the first entity and its impact and this is for the second time period. 12 stands for the time period. Then I will discuss it here.

So, till Tth period if Tth period data is given for the first category individual then it is interpreted like this. Similarly for others can be interpreted for Nth observations could be interpreted like this. Question here occurs either we can interpret like this vertically or we can interpret horizontal also. We are going to discuss that in a short while.

So, when we interpret this T here, another T is here, another T is here, horizontally or in a wide format that is called wide panel data. When we are putting it vertically that is called long panel. We are going to discuss this in a short while what is the necessity of long panel and what is the necessity of wide panel.

(Refer Slide Time: 14:06)



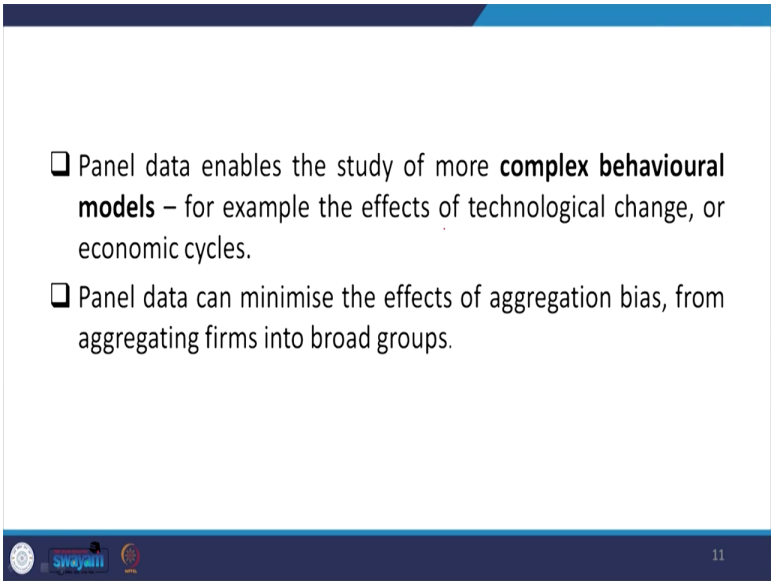


Before going to discuss all those conceptual details, advantages of panel data through Baltagi suggested some aspects of panel over cross-section and time series data. Since panel data relates to cross-sections over time heterogeneity can be captured in units of observation. The heterogeneity is very important in the modeling. The techniques of panel data estimation can take such heterogeneity into account by allowing individual specific effects or subject specific variables with observation that span both time and individuals in its cross-section.

Panel data gives more informative data and more variability, less collinearity among variables, more degrees of freedom and more efficiency. These are specific advantages of panel data. And panel data is better suited than cross-sectional data for studying the dynamic of change. Cross-sectional dynamics of change is not captured, because we are not, only at maximum over cross-section variable, over another variable we can check, but not over time.

So, there are many examples possible for it. Panel data can better detect and measure effects that cannot be observed in pure cross-sections and pure time series data. We can investigate how an event changes the outcome like some policies or law. So, that is why I said from the beginning that this data is highly useful for policy interpretation, for policy prescription or some intervention approach as well. If your technique or model is very appropriate then through the panel data you can find out the impact of the interventions.

(Refer Slide Time: 15:54)

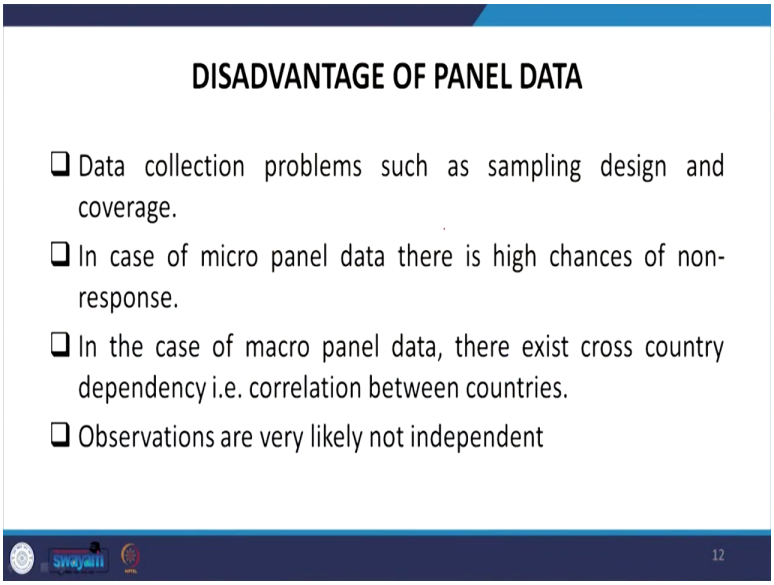


So, panel data enables the study of more complex behavioral models because some forms of interventions through the behavioral change can be captured, like the effects of technological change or economical cycles or even in the health cases like in COVID-19 we are now at the verge of conclusion of the classes for NPTEL. We are now at the 36th lecture. This is the very crucial time in the world so far as COVID is concerned. And some studies have identified placebo effect of COVID.

So, placebo effect without giving any physical medicine some impact has been observed as if the medicine has been invented, though medicine in reality does not exist. But still patients feel that I have been given medicine and I am getting cured. So, that means the government or the particular institution is doing fantastic. Some observation of such variety and the policy implications could be observed during the COVID time also. And we are recording over the very crucial pandemic time.

Anyway, let us come to the discretion of the panel and its advantages once again. Panel data minimize the effects of aggregation bias from aggregating firms into broad groups. So, like just getting the broad groups in cross-section is very difficult. But here since a time component is there, it may help making a group better and that may not be biased.

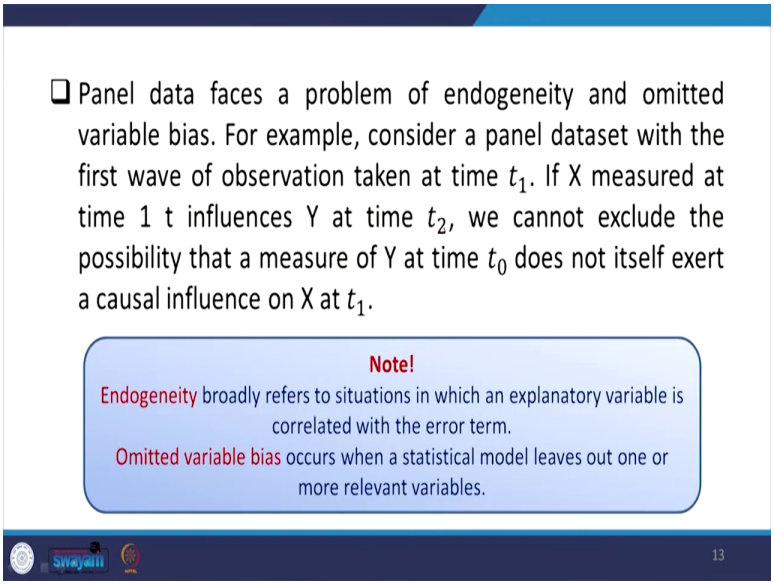
(Refer Slide Time: 18:43)



Coming to the disadvantages of panel data. There are some problems such as sampling design and coverage is very crucial. And in case of micro panel data there is high chances of non-response. Yes, that is the reason we have already said micro panel where the number of observation is higher than that of T, the time period. But all the observation may not have responses, because since it is highly qualitative observations, usually in cross-sectional approaches are qualitative and cross-sectional, so there are number of behavioral responses attached. Responses may not be there every time. So, it is very challenging to convert them into a panel format.

Similarly, in the case of macro panel data there exist cross country dependency. So, correlation between countries are also there. So, cross country dependency how those can be reduced, like entries, one country, second country, third country if you do it, they might be also dependent to each other. So, that how to tackle that is a very important challenge and very difficult in this case. Observations are very likely not independent. That is another problem. May be independent through cross-section, but over time might be dependent. So, there are some disadvantages of the panel data.

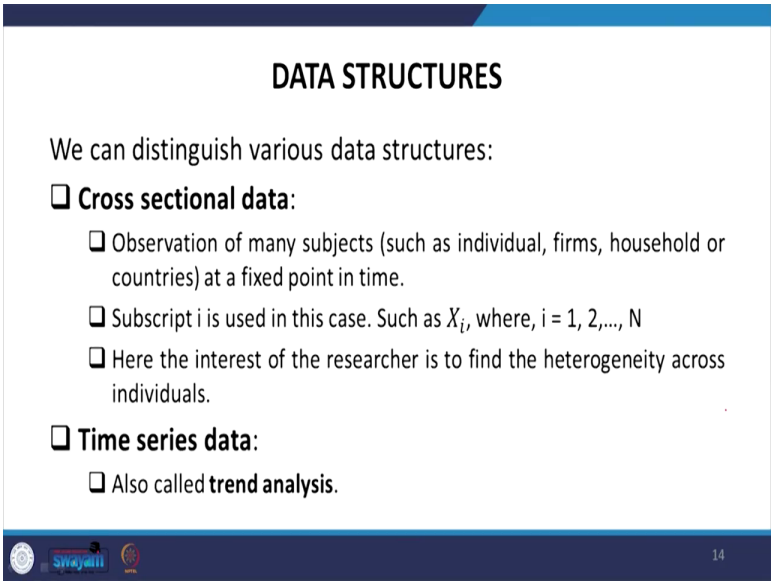
(Refer Slide Time: 19:24)



Panel data faces a problem of endogeneity. Basically the error term is correlated with the explanatory variables and the omitted variable bias. So, variable if gets omitted by, because of larger cohort we are taking, maybe very problematic. In case of omitted variable bias that occurs when we leave out or omit or drop by any reason by any unnoticed approach or steps if that leaves out, that is creating omitted variable bias.

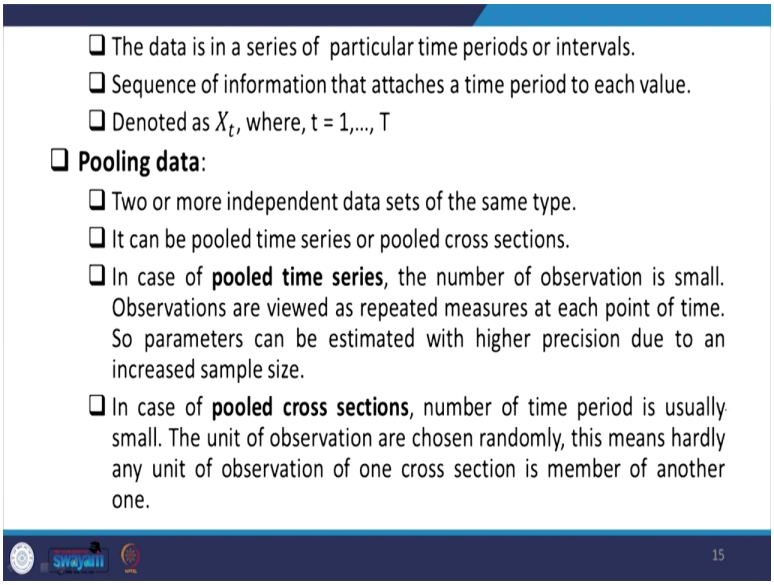
So, coming to some of the example like consider a panel data set with the first wave of observation at time t_1, if X measured at time 1 and t influences Y at time t_2, we can exclude the possibility that a measure of Y at time t_0 does not itself exert a causal influence on X at t_1. Basically, over the time period how the influence is observed is very difficult to capture. And how it is captured because of the problem of endogeneity and omitted variable bias is a matter of concern. And those are the challenges attached with panel data to tackle.

(Refer Slide Time: 20:49)



Coming to the data structure, we can distinguish various data structure like cross-section, time series, panel, within panel different types of panel. So, cross-section data already said many times is basically the responses at a fixed point in time. We usually denote with X varies from 1 to N and here the interest of the researcher is to find the heterogeneity across individuals, how individuals are varying across different heterogeneous groups or variables. Coming to the time series data usually most focus is on understanding stationarity or understanding the trend of the particular variable over time.

(Refer Slide Time: 21:34)



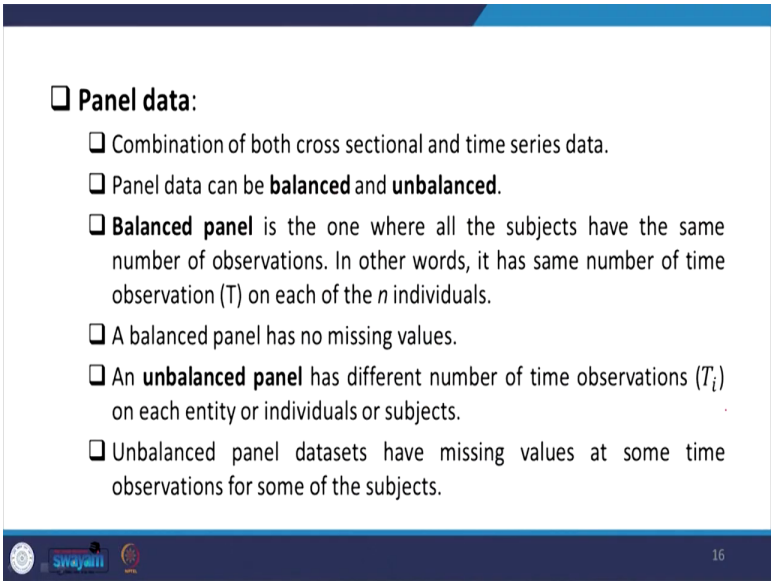
Data is usually is in a series of a particular time period or intervals. Sequence of information that attaches at a time period to each value is important. It is denoted as X_t, where t varies from 1 to T that is time period.

Another data usually discussed called pooling of the data, like two or more independent sets of the same type could be pooled. So, it can be pooled time series or pooled cross-section. In case of pooled time series data could be pooled and cross-section data could be pooled depending upon the independent data with the same type of independent series of data.

Coming to the case of pooled time series the number of observation is small. Observation is small, but T is large. We have already mentioned. So, observation are viewed as repeated measures at each point of time. So, parameters can be estimated with high precision due to an increased sample size.

In case of pooled cross-section, number of time is usually very small, isn't it? The unit observations are chosen randomly. This means hardly any unit of observation of one cross-section is member of another one, so because of cross-sectional approaches.

(Refer Slide Time: 23:04)

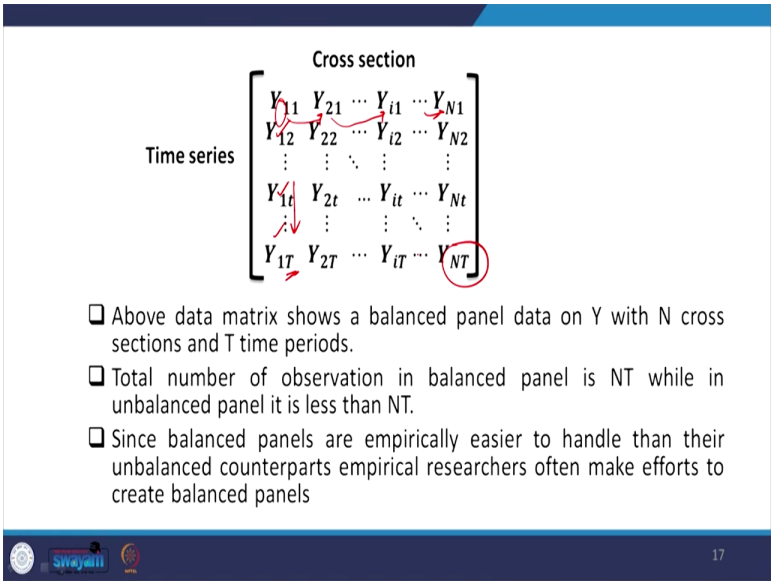


The last aspect of the data is important that is we are covering now called panel data. Panel data, already clarified the basics of it, discussing once again that after understanding cross-section and panel when we combine both are called panel. Panel data could be balanced or could be unbalanced.

What do you mean by balanced then? Balanced where the same number of observations are repeated. Every time it is same. In other words, it has same number of time observations on each of the n individuals. Every time the same time and same number of individuals continued in data called balanced panel data.

Balanced panel data has no missing values. That is more important. That is why we are saying it has to be 100 percent same. Missing value should not be there. Whereas in case of unbalanced panel, it has different number of time observations on each entity or individuals or subjects. Unbalanced panel data set have missing values at some time observations for some of the subjects.

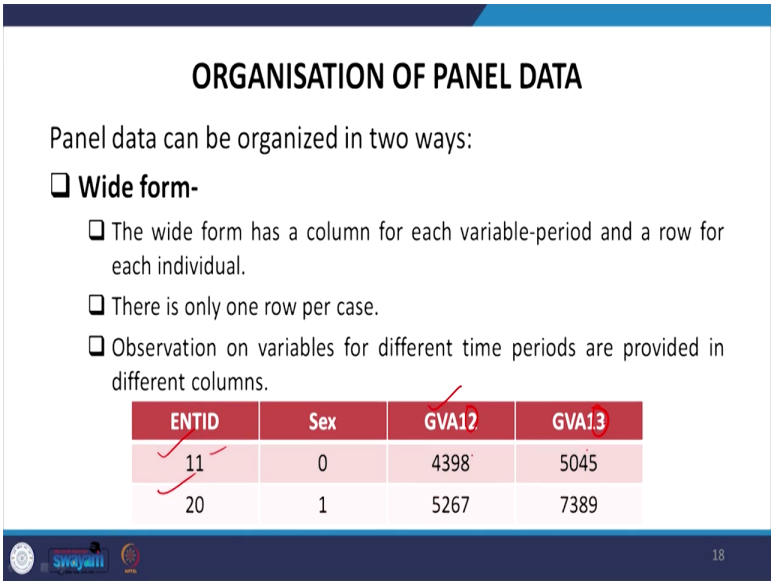
(Refer Slide Time: 24:04)



This is the time component we are mentioning here like you look at T here. T is mentioned over time for the individual 1, individual 1 varies over time. The individual 1 is changing to 2 then to i then to N. So, in total, we are going to discuss N times T total number of observations in the series of the panel.

If each of the entries are very valid then those type of data without any missing value called balanced panel data. Total number of observation in this case is NT. Since balanced panels are empirically easier to handle than that of unbalanced counterparts the researcher would do empirical estimation, often make efforts to create balanced panel. So, balance panel is easy to operate.

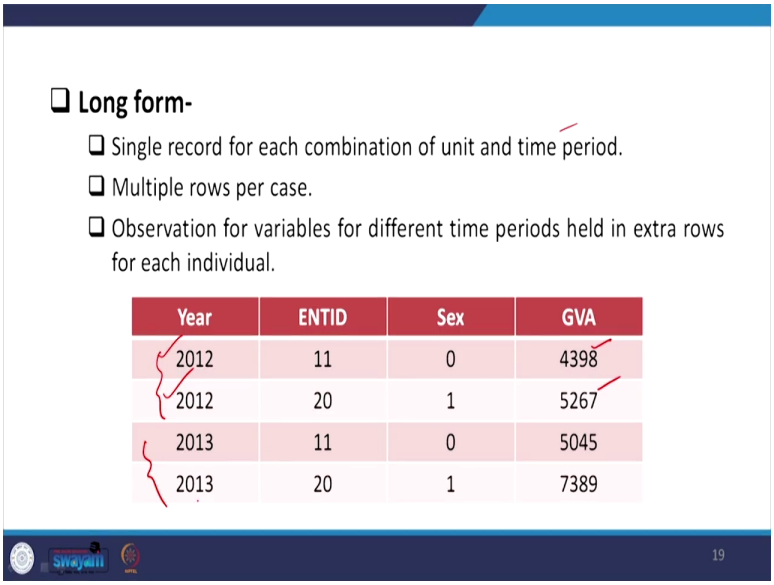
(Refer Slide Time: 25:17)



As I pointed earlier, understanding the organization of the panel data, usually it is of two ways that has been organized. One is called wide panel or wide form panel data and long form panel data. How the panel data could be organized, could be entered. The wide form has a column for each variable period and a row for each individual. Each row, but your time for each period we have like GVA, gross value added is another period GVA is mentioned. But in case of long it is entered vertically. We will show it.

Here is the case, the first case, this is the second case. One time invariant variable is given, sex where it is given for this individual. GVA for this individual is given as 2 times. In period 2 and period 3 it is given here.

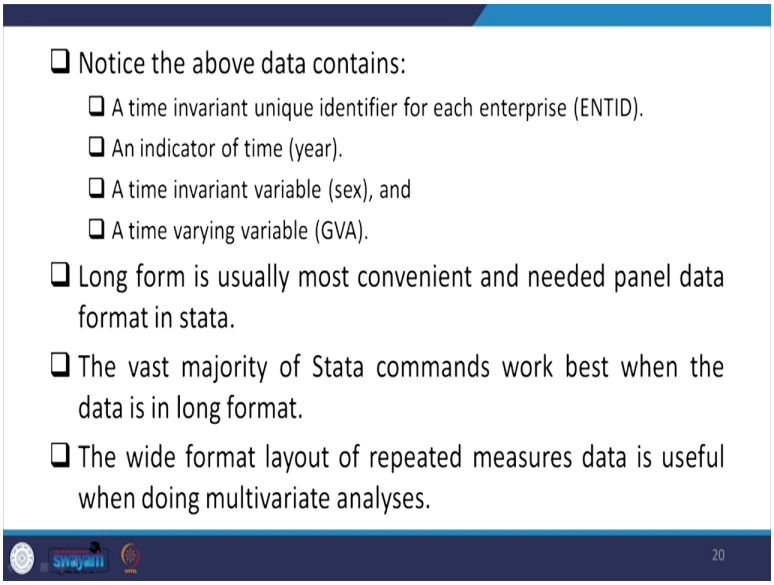
(Refer Slide Time: 26:11)



So, but in case of long panel the entries are kept vertically. A single record for each combination of unit and time period, multiple rows per case. Per case there should be multiple rows, like year is our unit of the panel.

Here 2012, 2012, two times is entered for the same that is for our per case it is 2012. There are double entry. For GVA, look at this. So, earlier in our case we entered as GVA in different variable in the column. But here in another row we have entered. Similarly, 2013 also.

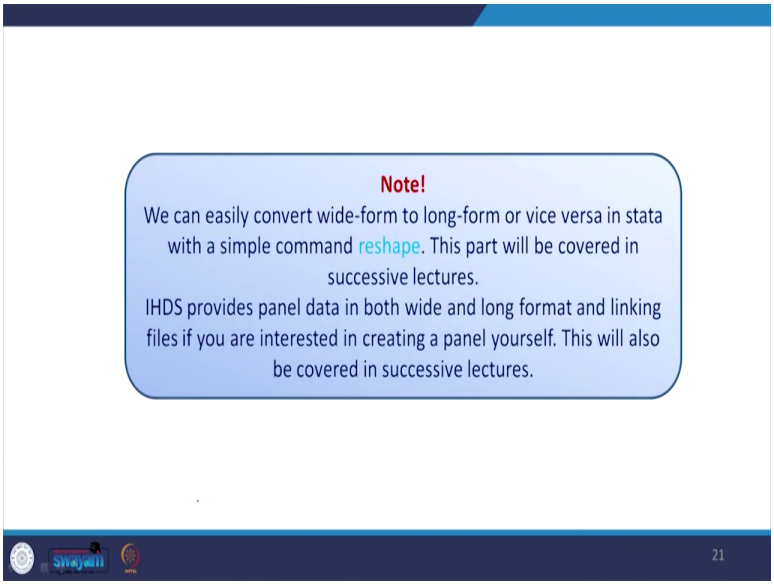
(Refer Slide Time: 27:05)



So, coming to some other features of this data are like time invariant unique identifier for each enterprise that is ENTID we used to discuss in our 73rd NSS data. Basically what it contains? What the data contains? It contains time invariant unique identifier variable that is time invariant. Then indicator of time T has to be mentioned. The time period has to be mentioned. Then the time invariant variable that is sex in our case, that is invariant over time. That is fixed. Then a time varying variable that is GVA over time that gets varies, we mentioned.

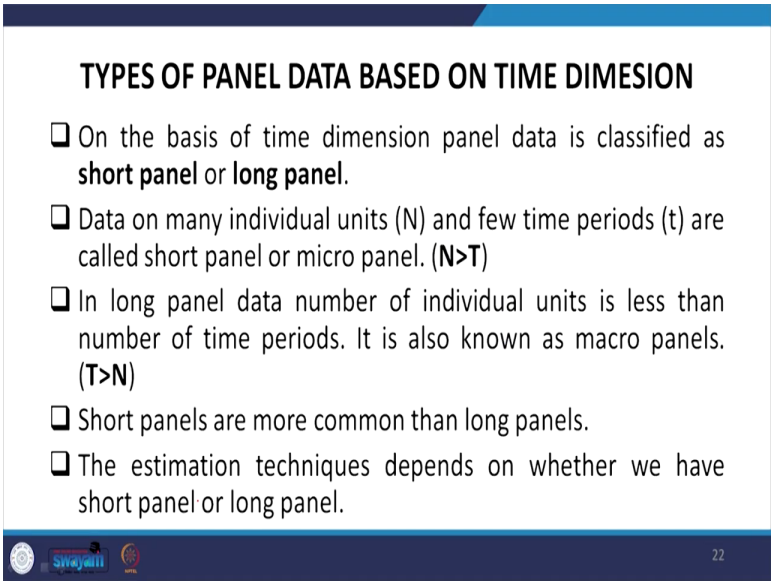
The long form usually most convenient. The long form basically we are arranging vertically. Number of rows increases in that case. Long form is usually very convenient and needed panel data format in Stata. The vast majority of Stata command work best when the data is in long format. But the wide format layout of repeated measures data is useful when doing multivariate analysis. In case of multivariate analysis that is where you use the wide format.

(Refer Slide Time: 28:25)



We need to take a note here that we can easily convert the wide form to long form through the command called reshape. And reshape command in Stata that can be easily converted. In IHDS data, and how to convert those data we will do in our other successive lectures. IHDS provides panel in both wide and long format. If you go to their site you will get those details separately. Linking files also where you can link net panel through the linking variables. Those linking files if you are interested is also available for creating a panel on your own. This will also be covered in our successive lectures.

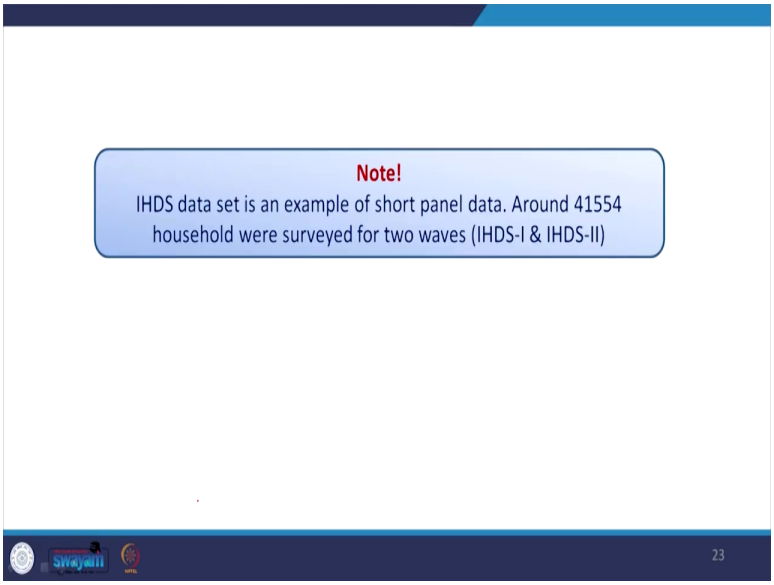
(Refer Slide Time: 29:05)



The types of panel data based on time dimension is important. We have already mentioned micro and macro panel. Based on the observations and the time, again another clarification I wanted to do it here called short panel and long panel. Like the time dimension panel data is classified as short or long whereas when the individual units, that is we have already mentioned, when the individual units that is N is greater than that of the T, time period that is called, in fact, that is called short panel.

Whereas usually in our micro study or in our cross-sectional panel data, we usually go for the short panel. In case of long panel the reverse is true. The time period is higher than that of the number of observations. So, short panels are very common than the long panels. The estimation technique depends on whether we have short or long panel.

(Refer Slide Time: 30:17)



IHDS data is called a short panel data. Around 41554 households were surveyed for two time period, two waves that is IHDS-I and II. So, those are the details we explained in this lecture, basically the conceptual framework of the very basic concepts of panel data, and in the next class we will also give further details of conceptual framework on panel data to start the application with Stata. With this, let me stop here. Thank you.