Exploring Survey Data on Health Care Prof. Pratap C. Mohanty Department of Humanities and Social Sciences Indian Institute of Technology, Roorkee

Lecture - 31 Balanced and Unbalanced Panel Data in Healthcare

Welcome friends, to this NPTL MOOC program on exploring healthcare survey data. We are in the 7th week, trying to explore the healthcare data and their panel information; I am Pratap Mohanty, at present teaching in IIT Roorkee. This particular lecture is targeted to deal with the panel data of balanced or unbalanced kind; how balanced panel and unbalanced panel is understood in healthcare data.

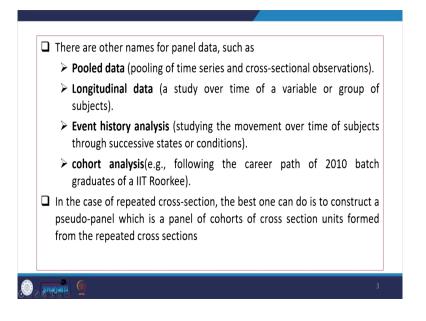
We will give you the practical handouts to understand the panel data in healthcare. Now, I am clarifying the meaning of panel data; panel data is basically a mixture of cross section and time series. As we have already explained cross section then as compared to the time series, this is basically repeated cross section data and it is also called same cross section units followed over time. Panel data set could be at the individual level and also could be at the household level.

(Refer Slide Time: 01:45)

Panel Dataset - same cross section units (e.g. individuals, households) are followed over time.
Normal data have seen a well as the advectory
panel data have space as well as time dimensions.
Like time series data, panel data contains observatio collected at regular frequency.
Like cross sectional data, panel data contains observatio across collection of entity(individuals, households).

Panel data have space as well as time dimensions. Like time series data, panel data also contains observations collected at regular frequency. Like cross section data, panel contents observation across collections of entity such as individuals and households.

(Refer Slide Time: 02:06)



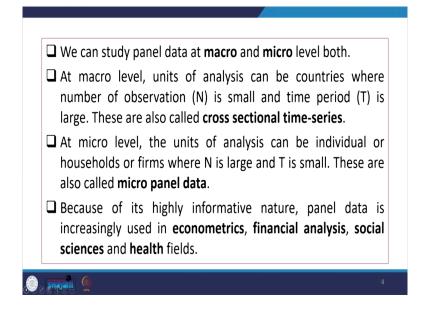
There are other names of panel data, such as pooled data; like pooling of time series and cross-sectional observations, where the same unit of observation may not be repeated over time, let us say simply pooled. Whereas, in case a longitudinal data, a study over time of a variable or a group of subjects. Another one is called even history analysis; we are studying the movement over time of subjects through successive states or conditions.

Another aspect or type of panel data is called cohort analysis; Suppose in this case, we follow the career path of 2010 graduates of IIT Roorkee for example. So, that batch- how it is reflected in different successive periods; when we are catching our analysis for a particular batch, particular cohort, that type of analysis is called cohort analysis.

In the case of repeated cross section, the best one can do is to construct a pseudo panel, which is a panel of cohorts of cross section units formed from the repeated cross sections.

We will have also a dedicate session called pseudo panel; we will be discussing a bit on how to form it, we will also give some example dataset for your better understanding.

(Refer Slide Time: 03:40)

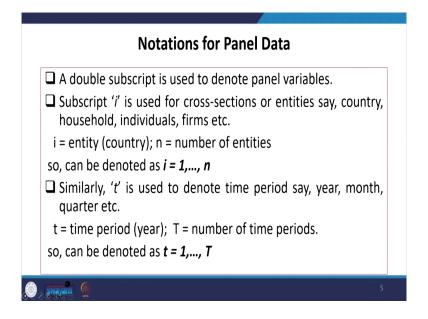


We can study panel data at macro as well as micro level. At macro level, units of analysis can be countries, states etc. The number of observations is usually less and time period is large in case of macro panel. Over the time period, number of observations are available; but the time period horizon is quite large in macro context. These are also called cross sectional time series data.

At micro level, the units of analysis could be of individuals or households or firms, where N is large and T is small, T refers to time. These are called micro panel data.

So, now you can differentiate between macro panel and micro panel; because of this property, it is highly informative in nature, moreover panel data is increasingly used in econometrics, financial analysis, social sciences and health fields these days.

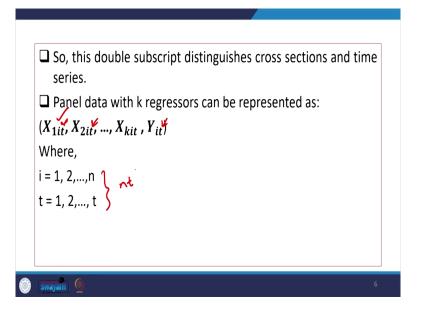
(Refer Slide Time: 04:48)



There are different notations used for panel data, which is distinctive than that of the earlier conventional regression techniques or cross-sectional data.

In case of panel, a double subscript is used to denote panel variables. So, it is not just one subscript, it is 2 subscripts; if a model has used those, those are actually denoting panel variables. Subscript like i is used for cross sections or entities; let it be country or household, individuals etc., they are called cross-sectional subscript. i is your entity and it matters with how many entities are there, that is called n. So, it can be denoted as i, where i stands for 1 to n. So, maybe i varies from 1 to n, and that variation we will get it in panel. Similarly, another subscript that is used is called t; t stands for time, time period over whether the data is available over month or over year or over quarters, accordingly the t dimension is also given along with i dimension. So, t may also vary from 1 to capital T, i.e., the Tth period we have mentioned.

(Refer Slide Time: 06:15)



So, this double subscript distinguishes panel from cross sections and time series. In cross section we have only i; whereas in time series we have only t, but in panel, we have both i and t. Therefore, the dimension of panel in total is of i into t or since i varies from i to n across times; so it should be n times capital T.

So, n T dimensions are actually included in panel data set. Panel data set with k regressors can be represented like the following, as we have presented here. So, like in cross sections, we used to have X 1, X 2, X 3,Xk and here we have X1it to upto Xkit and at the end we have Yit, Y variable denoted as dependent variable. Here we are mentioning that it is not just X1i, it is X1it; t component is mentioned, this is your t in addition to that.

Similarly, in the Y variable we have also added with the subscript called t; says we have already pointed out that i varies from 1 to n, then t varies from 1 to t. So, total dimensions number of dimension will be n times t.

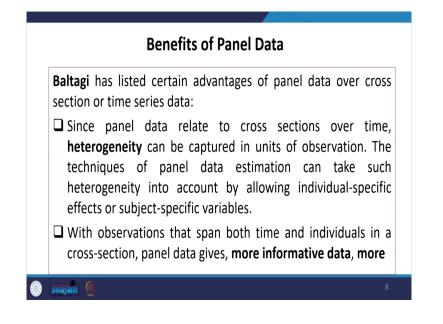
(Refer Slide Time: 07:40)

Entity	Time	Notation
1		
1	1 2-	Y ₁₂
1	Т	Y ₁₁ Y ₁₂ Y ₁₇
N <mark>/</mark>	1	Y_{N1}
N	2	<i>Y</i> _{<i>N</i>2}
N	Т	Y _{NT}

Now, just look at an example on the screen for your clarity; here entity is given; time dimensions are given, the notations that is used is given.

Entity 1, it is still repeated here, then it is N entities. And then notations are given as if it is time period 2. It is available for the same entity and the same entities is also available till the T time period. Then the notations are r is equal to Y 1 1, since 1 is here and 1 is there (in first row); then this is 1 and 2 is here in second row, then 1 and T in third row. Similarly, if N is the entity, then N is represented in T time period; so, similarly N 1, N 2 and till N t is mentioned.

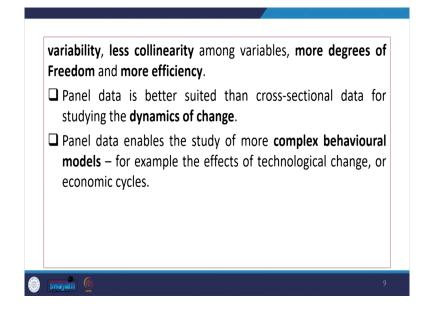
(Refer Slide Time: 08:40)



Now, we are discussing about benefits of panel data, you will get better idea from the Baltagi's book. This has listed certain advantages of panel data over cross sections or time series data. The first important advantage is that, panel data relates to cross sections over time.

The heterogeneity can be also captured in units of observations, that is most important. The techniques of panel data estimation can take such heterogeneity into account by allowing individual specific effects or subject specific variables. Then the second advantage is that, with observations that span over time and individuals in a cross section, the panel data keeps more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency.

(Refer Slide Time: 09:31)



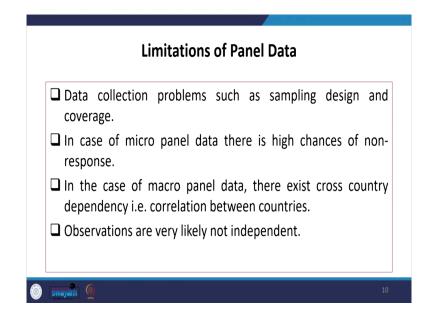
So, one explanation behind all those reasons is that, we are including more dimensions, more frequencies; since your t times component is multiplied with the n number of observations.

Or to the ith, "i" which varies from 1 to nth observations i.e., again multiplied with t, so dimensions actually increase; therefore we will get better results with efficiency. Panel data is better suited than cross section data for studying the dynamics of change. Panel data enables the study of more complex behavioural models; for an example- the effects of technological change or economic cycle. Any sort of effects even in evaluation can also be dealt better in panel data.

Since it also reduces the individual heterogeneity as these are normalized over period of time and is usually present in cross sectional data.

Now, we are also identifying certain limitations of panel data; i.e., these data create some problems, such as framing your sample, then getting it is coverage.

(Refer Slide Time: 11:00)



In case of micro panel data, there is high chances of non-response in the next round; since you are covering more observations, non-responses would be there. We will also show one example data set from India's human development survey in comparison to the previous one and they have given a panel content.

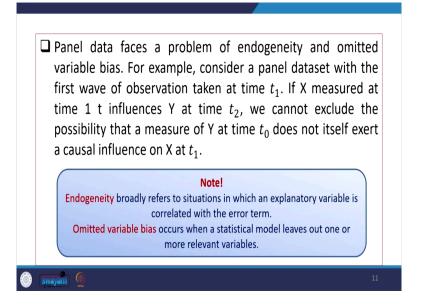
We will discuss that how many observations are actually missing from the previous to the new. And accordingly, the final panel is made and whether that is going to be balanced or unbalanced, we will discuss.

In the case of macro panel data, there exist cross country dependency such as correlation between countries. Since, cross country dependencies are always the case in international trade context; so, correlation between this country might exist, then that may create some problem in the panel explanation.

Observations are very likely not independent so that is another issue; though our assumption is of independent-ness. But actually, they are not independent, so over time they are not necessarily going to be independent.

Now, we are clarifying certain conceptual things like endogeneity issues and omitted variable bias as the limitations of the panel data. Panel data faces a problem of endogeneity and OVB that is omitted variable bias.

(Refer Slide Time: 12:33)



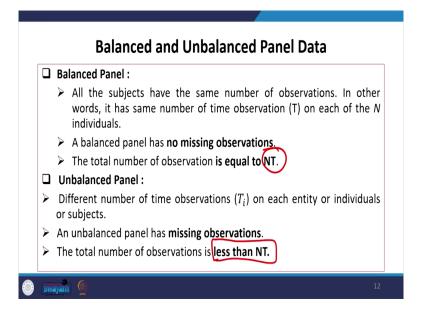
Let us consider a panel dataset with the first wave of observations taken at time period t_1 . If X measured at time 1t influences Y at time t_2 ; we cannot exclude the possibility that a measure of Y at t_0 does not itself exert a casual influence on X at t_1 . So, t_0 and t_1 whatever is the response is over 2 time period, we cannot just conclude that there is no casual influence.

So, here the problem of a and endogeneity arises; because of the fact that the explanatory term might be correlated with the error term or the explanatory variable might be also explaining the model because of some correlation exist.

So, endogeneity bias or endogeneity broadly refers to a situation in which an explanatory variable is correlated with the error term. Like we said Y is a function of X and X is a function of Y. So, in that case your error term is in fact correlated with your explanatory variable i.e., X. And this happens because of the repeated information over time and there might be correlation in the error term.

Another problem that we have in panel data is we are supposed to leave certain important variables or observations, as we cannot take or are unable to take all variables or observations, may be due to unavailability; so, in that way there might be a huge possibility of omitting certain important variables. So, that is in sort called- omitted variable bias. So, there are some techniques to control this biasness.

(Refer Slide Time: 14:35)



Now, we are going to clarify you about the balanced versus unbalanced panel data. I am just now going to give you the conceptual background about it or you can say conceptual information about it; then at the end we will be experimenting with the data and I will show you that how we can clarify our concepts practically.

First of all, what do you mean by balance panel? Balance panel are the panel where all the subjects have the same number of observations, all the subjects or the entity have the same number of observations that is entered. In other words, it has same number of time observations i.e., T should be same for each of the N individuals. So, N and T is expected to be perfect and there are no missing observations. So, in that case total number of observations is equal to N T; N times T, T since no missing observations are available or no missing observations are reported.

Then second one is called unbalanced panel; here obviously another side is true, there must be missing observations and that too all N observations by the time are not available. So, that means different number of time observations on each entity or individual or subject are not available. And unbalanced panel has missing observations for sure; the total number of observations is of course less than that of "N times T" i.e., the total dimensions of the panel.

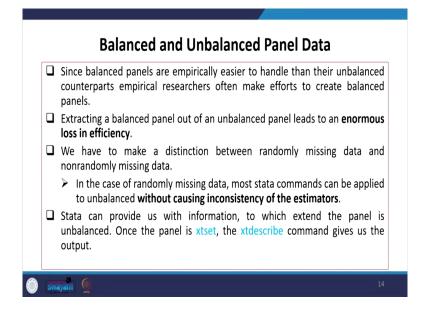
(Refer Slide Time: 16:17)

	Health expenditure	Nature of treatment	N	Ţ	Health expenditure	Nature of treatment
2019	5000	Modern	0	2019	5000	Modern
2020	6000	Modern		2020	6000	Modern
2021	l 7000	Modern	N.			
2019	2500	Traditional	S^2	S ²⁰¹⁹	2500	Traditional
2020	2300	Traditional	7 ²	2020	2300	Traditional
2021	L 2600	Traditional	62	2021	2600	Traditional
	alanced Panel		•		alanced Pan	el

Here is an example on the screen for you to clarify the balanced panel and unbalanced panel. Here your N, T, healthcare expenditure and nature of treatment is given. The first N observation i.e., entity 1 is repeated over different time periods like 2019, 20 and then 21. This is "N times T" dimension we are discussing here since the second observations is also repeated with the same time period. And all the entries are available, that is why we are saying this is a balanced panel.

Second table is called unbalanced panel because you can just see that here entity 1 is available in 2019 and 20. So, 2021 is missing here, whereas 2nd entity of N is available in all the 3 periods. So, 2021 for the first observation is in fact missing and providing missing information; therefore, this is called an unbalanced panel.

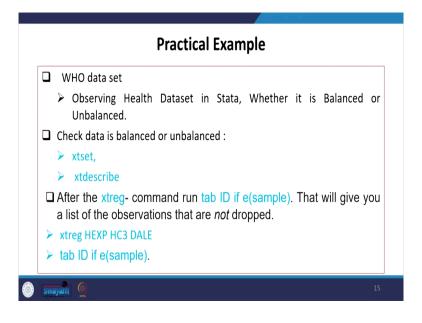
(Refer Slide Time: 17:24)



So, further information about balanced and unbalanced panel are mentioned here that- since, balanced panels are empirically easier to handle than their unbalanced counterparts, the empirical researchers often make efforts to create balanced panels. But, extracting a balanced panel out of an unbalance panel leads to an enormous loss in efficiency.

We have to make a distinction between randomly missing data and non-randomly missing data. So, some data points might have been randomly missed and some might be non-randomly or are actually deliberately missed, so these non-randomly missed data points are actually a case of concern. If it is randomly missed, still the data set can be considered as balanced panel. In the case of randomly missing data, most data commands can be applied to unbalanced without causing inconsistency of the estimator.

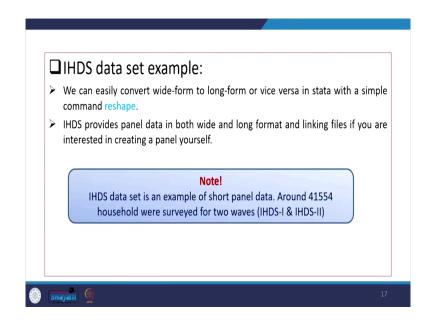
So, once Stata finds that missing entries are there, Stata simply reads it as unbalanced panel. But it is not creating inconsistencies in the result or the estimator is not inconsistent. Stata can provide us with information the extent to which the panel is unbalanced. Once the panel is given the command i.e., xtset and the xtdescribe command; then we can understand whether the data is unbalanced or not and to what extent it is unbalanced. (Refer Slide Time: 19:05)



(Refer Slide Time: 19:08)

	Types of Panel Data based on time dimension
	On the basis of time dimension panel data is classified as short panel or long panel.
	Data on many individual units (N) and few time periods (t) are called short panel or micro panel. (N>T)
	In long panel data number of individual units is less than number of time periods. It is also known as macro panels. (T>N)
	Short panels are more common than long panels.
	The estimation techniques depends on whether we have short panel or long panel.
0	Swapani 💮 16

(Refer Slide Time: 19:09)



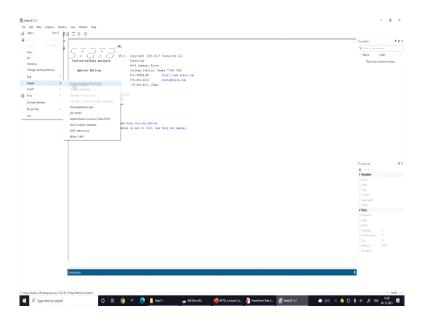
Now, we will have a practical data set; the practical data set we have taken is from WHO sample data set; this information will also be provided to you as an additional document for your own practice. And we are going to explain all those things, I am just referring what are the data sets we are considering here.

We are observing health dataset in data whether it is balance or unbalance. We need to check this through these 2 commands (xtset and xtdescribe) to identify whether the dataset is balanced or unbalanced.

After that in the next lecture, we can think of running regression command through the xtreg command. And we can accordingly find out whether that is more suitable or not, that will give you a list of observations that are not dropped.

So, I will try to give the detailed explanation of xtset command in the next lass; but at this moment I am going to talk about some practical examples, so let us go for it, then we will also look to other aspects.

(Refer Slide Time: 20:21)



(Refer Slide Time: 20:33)

7 # ×						A	Variables	
ommands have 0							A litervasabled	
mand ,rc		Copyright 1905-2017 5t	stsCorp LLC				Nove	Label
e are no items to show.	Statistics/Data Analysis	StateCorp					There are no	items to sho
	Special Edition	4905 Lakeway Drive College Station, Texas	334.17 B43					
	Special Edition		tp://www.stata.com					
		979-696-4600 at	ata@atata.com					
		79-696-4601 (fax)						
			🗐 Import Excel	×				
	PROGRAMMENT FOR MONTAGE		9					
			Exad Re					
	Notes:			trasse				
	 Unicode is supported; see 2 Haximum number of variables 		Worldheet	Cellrange:				
	2. NALIGN DISOTT OF TREESON	18 Sec 10 2000, See 10	· · · · · · · · · · · · · · · · · · ·	10				
			Import Bist-row as variable names	Variable case:				
			Droport al data es stringe	Rentos				
			Preview					
			There are no items to	-her-			Properties	
			11000 (10 100 100 100 100 100 100 100 10	(CON)			8	
							* Variables	
							Label	
							5.94	
							Format	
							Value label	
				OK Canod			Notes + Data	
					6		Liename	
							Label	
							Notes	
							Variables	
							Observations Size	
							Memory	54M
							Sorted by	
	Command			and the second second second			1	

(Refer Slide Time: 20:35)

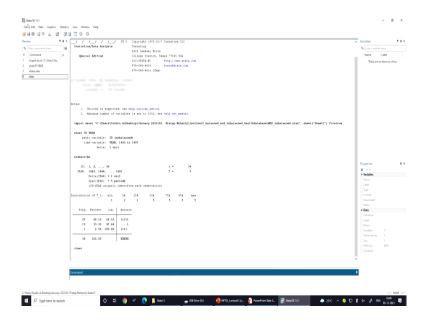
🚑 SatarSE 15-1											- 0 ×
File Edit Data Graphics Statistics											
S 1 8 1 3 1 4 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5 1 5	3 🗇 O · O										
Review T # X									×	Variables	т # х
A liter commands have										A liter catables	
# Command _rc		Copyright 1985-2017 Sta	teCorp LLC								Label
There are no items to show.	Statistics/Data Analysis	StateCorp 4905 Lakeway Drive								There are no	e items to show.
	Special Edition	College Station, Texas SOO-STATA-PC htt	77045 USA p://www.stata.com								
		979-696-4600 814	taĝatata.com								
		79-696-4601 (fax)									
			🔲 Import Excel			×					
			🗓 Open						×		
	Notes:			- Dr. Pra	tap Mohanty >: Lecture31,balanced,and,ur	nbelanced healthdatabasa	~ (5 P Septhia	ture31, belowed,		
	1. Unicode is supported; see b										
	Maximum number of variables	i is set to 5000; see he	Organise • Nev						• • •		
			CrieDrive		Name	Date modified		Sax			
			💷 This PC		who bainced	02-12-2021 19:35	Mensell Loof W., Monsell Loof W.,	43 153			
			3 10 Objects		0 WHO_unbalanced	02-12-2021 1941	Morrielt Ecol W.,	410			
			Cesktop								8 ×
			Documents	1.1						Properties	4 ×
			Downloads							* Variables	
			Music							Name	
			Videos							5,94	
			Local Disk (C.)							Format Value label	
			USB Drive (D.)	÷.,						Notes	
			E Local Disk (E.)							+ Data	
			Local Disk (F.)							Label	
			= US8 Drive (D.)	×						Noter	
				Filename	1 WHD_unbalanced			 Excel Files (*.als) 	c*ah) ~	Variables Observations	
								alin	Cancel	524	
									-	Memory Sorted by	6M
	Command									i	
C1Uses/Studio #/Desktog/January 2022/Di	Pratap Mohanty/Stata15										COP NM OR
JP Type here to search	o 🛱 🌍	🧈 👩 📘 Salati		Drive (D)	🕼 NFILL Lecture 21, b. 👔	PowerPoint Side S.	Roto/SE 15.1	23°C	^ 🧿 🖸 🛙	4 JA DAG	1638
						_					00-12-2021

Now, here is our data, I am just opening it for your reference. Now, I am going to the file, then importing the data; this database we have taken is from the WHO and this is in excel base format.

(Refer Slide Time: 20:39)

T # X				 Variables
er commands hara 🛛 🕕				🔦 filter væsables here
ommand _rc		Copyright 1985-2017 St.	staCorp LLC	Name Label
here are no items to show.	Statistics/Data Analysis	StateCorp		There are no items to show.
		4905 Lakeway Drive		
	Apecial Edition	College Station, Texas SOO-STATA-PC htt	77943 00A gi//www.stata.com	
			itajatata.com	
		79-696-4601 (fax)		
	Annual and an annual state			
			Import Excel X	
			bod for	
			C:Laertjäude #DesitepUaruary 2022Dr. Protep RohantyLecture Browse	
	Notes: 1. Unicode is supported; see 1	ale unicode admice	For stars a formula stars and the stars of the stars and stars and the stars and the	
	 Maximum number of variables 		Violaheet Celinage SheetLA3,42143 V A12,42143	
			PERIALATE TANKING	
			Depart fint row ex variable names Variable Gases	
			Compart al data as strings	
			Prevent (Revery rave 1-50 of 140)	
			A 0 C 0 E F G 10 Covery COMP YEAR HDP HC3 DALE	Properties
			2 1 Cesk Idands /5.46/54 1997 344.01/5 84/3607 41.43486	A
			2 Andona 90.98447 1997 1213.999 7.12806 72.49335000000	* Variables
			4 3 Alghanistan 52.12939 1997 28.28551 1.158772 37.62949	Name
			5 4 Angola 52,29271 1997 47.17409 2,248598 37.52818	Label
			6 5 Albania 78.70827 1997 63.5321 4.89903 60.17552	3pe Lorent
			716 Doted Auto E., 81.818 1903 811,2003 4,00002 45,04400	Value label
				Notes
			OX Canol	+ Data
				Filename
				Label Notes
				Variables D
				Sat 0
				Memory 64M
				Sorted by
	Command			
	Contraine			N N

(Refer Slide Time: 20:53)



Now, I have opened the unbalanced data. But we need to specify that which row is considered to be the variables, Stata has automatically taken the first row as variables. We have opted for it and now we are going to take it as it is. Then I am going to open the command, going to run the command to understand you whether the data is in fact unbalanced or not.

We have opened though an unbalance panel, I am just going to specify it. So, I am giving you 2 dimensions; one is the ID that is the observation dimension, second one is time dimension. So, now after entering command of check, with the enter we can find out the result.

The xtset command is defined. It is ID variable and year. Now, you can see, it has clearly identified that the data is unbalanced. And the data is available from 1993 to1997. This has obtained a delta method to find out whether there is any missing observation or missing time period.

So, now we will check it to describe details with xtdescribe command; that will clearly highlight where exactly the data is missing. So, data is defined as unbalanced. Now, you can see that out of 39 observations, the 25 observations where the pattern is explaining at the end box; you can see in the pattern here it is written as 5 times 1.

So, 5 times 1 means, these 25 observations are actually repeated in all 5-time period as provided by the WHO data. There are 13 such observations, where only the last time period information are the latest time period entries.

In only one case, one observation where first 4 time period are available, but the last time period it is in fact missing. Therefore, the data has earmarked as unbalanced. So, next we are going to open the data set to highlight or to understand the balanced panel.

v Tax		opyright 1905-2017 StataCorp LLC	 Variables
iter commands have 0		HataCorp 905 Lakeway Drive	1 iter variables here
Command ,rc		ollege Station, Texas 77045 USA	Name Label
import excel "C/User/Sta		00-STATA-PC http://www.stata.com	There are no items to show.
visiet ID YEAR		79-696-4000 stataBstata.com	
ndexribe		79-686-6601 (fax)	
lear			
	Anne aggree to construct a		
	and the second s		
		Trace first X	
	Notes:	Import Excel X	
	 Unicode is supported; see hell 		
	Maximum number of variables i	a set to 5000; see hel ClibertDado AlbedingUseuwy 2022Dr. Protop PatrantyLettre Browse.	
	import excel "C:\Deers\Studio A\Deskt	op\Jannary 2022\Dr. P multhest Delongs: Dase\WED unbalanced.xlsx", ebeet("Sheet1") firstrow	
		Spect ALA/366 V ALA/266	
	. xtset ID YEAR		
	panel variable: ID (unbalanced) time variable: YEAR, 1993 to 1		
	delta: 1 unit	Dirport of data as strings Preserves V	
		Preview: (phowing raws 2-51 of 166)	
	. xtdestribe	10 Country COMP 11AR HOP HC3 DNIE SM	
	10: 1, 2,, 39	2 Under Aub Ever 815/1 1961 8112901 4/6682 650468	Properties
	1EAR: 1993, 1994,, 1997	3 1 United Acab (mir: 81.84541 1996 804.0387 4.77231 65.28066	B 11.1
	Delta(IEAR) = 1 mnit	4 United Acub Envir 82.0729 1995 833.5784 4.832752 65.50619	* Variables
	Span(YEAR) = 5 periods	5 1 United Arab Emir 42.30042 1996 3404166 4.898788 6575187	Name
	(ID'YEAR uniquely identifies		1,pr
	Distribution of T_1: min 54	7 2 Augentina 80,7625 1901 750,9965 6,460517 6553031 254 504 7 8 2 Augentina 80,9625 1904 701,4254 6,50506 61,5524	1991 .
	1 1	1 5	Value label
			Notes
	Freq. Percent Cun. Pattern	OK Canod	4 Data
	25 64.10 64.10 11111		Label
	13 33.33 97.441		Laber Noter
	1 2.56 100.00 1111.		Variables 0
			Observations 0
	39 100.00 XXXXX		5ex 0
	. clear		Memory 64M
			Sorted by
			v
	and the second se		
	Command		•

(Refer Slide Time: 23:11)

So, so here I am going to open a kind of panel data, which is a balance one; we have again taken from the same source i.e., WHO. So, WHO balanced panel data it is on healthcare issues. We will also share this with you, we have identified here the variables names as in the first row.

(Refer Slide Time: 23:31)

iev T#X	10: 1, 2,, 39 5 - 39 TEAR 1995, 1994,, 1997 7 - 5	A Variables
Eiter commands have Command //re import excel "C/Aber/Sha	The 100, 100,, 1007 T = 5 Description = 1 most (pertilia) = 5 pertons (PTUE upper information and discrementant)	Name Label Dere are no items to show.
vtuet ID YEAR vtdescribe clear import excel "C-/Diery/Sta	Dintribution of T_11 min 54 224 504 504 504 max 1 1 3 5 5 5	
vouet ID TILAR indexceller dear	Trep. Review Con. Price 29 64.01 61.01 30 35.01 111 1 2.56 01.06 10 100.05 1003	
	- clear	
	. import excel "C:\Deers\Studio AlDesktep\Jennery 2022\Dr. Frstep Nobasty\Lecture3]_balanced_and_unbalanced_bealthdatabase\who_balanced_ster", elsest("Deet1") firstrow	
	 rear ID Wak ponel versionalise ID derivenity balanced time versionalise 1 WBA, 1030 26 10307 derivate 1 work 	
	. xtdescribe	Properties B = =
	The 1, 1, 2,, 23 0 0 The 1, 1, 2,, 23 7 0 The 1, 1, 2,, 24 7 0 Specification 2, and the comparison of the com	Variables Nome Listel Type Formut Nome Nome
	Distribution of T_1: min 54 234 554 554 854 max 5 5 5 5 5 5 5 5 5	Note Outa
	Treg. Percent Cun. Pettern	Label Noter
	33 166.46 11111 33 166.46 XXXX	Variables D Observations D Size D Memory 6MM
	- clear	Sorted by
	Connund	

Now, with the same commands, like xtset, we will confirm whether this data is balanced or unbalanced with the help of ID variable and the time variable that are defined here.

Now, you can see results of the xtset command, xt is the standard command in panel; though xtset is conforming us that your data from third 1993 to 1997 is a strongly balance panel. You can also check it using the description as well, through the xtdescribe command.

These do files will also be shared with you. Now, you can see all those 33 observations are actually repeated in all 5 time periods. So, 1 is mentioned 5 times. So, all 5 time periods, all 33 observations available. So, I think this is not problematic and I am sure that you can easily understand it in any database, whether it is a panel or not a panel data.

So, this is how I have explained. Let us move to the understanding of all the datasets; I will also clarify some other concept related to panel, like short panel and long panel. On the basis of time dimension, panel data is classified as short panel or long panel. Data on many individual units and few time periods like where N is higher than that of the T. In that case it is called short panel or that is called micro panel. Just the reverse of this is observed in the long panel or in the macro panel. So, long panel are also called macro panels. The long is the word refers to mostly emphasizing the time; short panels are more common than long panels, the estimation techniques depend on whether we have short panel or long panel.

We are now exploring in front of you the data of IHDS i.e., India's human development survey. This keeps panel observation and this is in fact a micro panel data, where N is higher than that of T, T here is only for 2 periods. So, we are going to open it and here number of observations are of 41554.

Households were surveyed for 2 periods in IHDS I and IHDS II. We have already discussed something about the IHDS. You can also search in any website about India's human development survey and you will get to know more about it. So, IHDS dataset is an example of short panel data.

We can easily convert wide form to long form or vice versa in Stata with a simple command called "reshape". So, like a wide panel, where number of cross sections are much higher and our time dimensions are much higher; if it is so, then we can actually flip it through reshape command.

IHDS provides panel data in both wide and long format and linking files. So, if you are interested in creating a panel yourself, this is going to help you better. Now, I am going to show it here about this IHDS data for your better understanding.

(Refer Slide Time: 27:11)

	Pratap Mohanty + Lecture11_balanced	and_unbalanced_healthdatabas		▼_00 is	arch Lecture 21, balanced,	-	 Variables 	T
unise • New folde	,				E • 🖬 🛛		1 the variables he	
Dektop / 1	Name	Date modified	3pe	See			Name	bel
Downloads /	37382-0011-Duta.dta	68-01-2020 12:36	DGA File	1,87,742.88			There are no its	work of arra
Decuments /	panel,do.do	01-12-2021 16:17	DO File	1.0				
Pictures /								
Dr. B.R.Gurjer								
Dr. V. C. Sriveste								
Lecture27_Qualit								
Lecture31_balan								
OneDrive								
OveDrive								
This PC						ped and unbalanced healthdatabase\who balnced.xiss", sheet("Sheat1") firstrow		
10 Objects				D.		pred and underlanded nearthdatabase(who_Bainoed.xist", sheet("Sheet1") firstrow		
Dektoo				ч				
fie ro	17382-0011-Outa-dta			~ All State	files v			
				Op	en Cancel			
	. Ktosstripe					4	Properties	
	10: 1, 2,				33		* Variables	
	YEAR: 1993, 1	994,, 1997		÷ -	5		Name	
		EAR) = 1 unit AR) = 5 periods					Lev	
		R uniquely identified	each observe	(105)			Format	
	Distribution of T	11 min 54	254 501	754 9	54 mex		Value label Noter	
		5 5	, ,	5	5 5		4 Data	
	Treq. Percer	c Cum. Pattern					Label	
							Noter	
	33 100.0	0 100.00 11111					Variables Observations	
		0 XXXXX					548	
	33 100.0						Memory	6IM
	33 100.0						Constant Inc.	
							Sorted by	
							Sorted by	

(Refer Slide Time: 27:16)

	x	 Variables 	т а :
l iter commands have) 39 100.00 XXX	\$ 0.04	
		Nave	Label
Command re- report loced "C1/User/Stu	, clear	HHBASE	Label Survey mark + co
ingen eicer Cristien stall.	. import excel "C: (Berra\Stadio AlDesktop)/Janary 2022)Gr. Fratap Nohasty\Lecture31 balanced and ubblanced healthdatabase\who balnced.stas", sheet("Sheet1") firstrow	SURVEY	H051 (2005) er IH.
Mexile	· des care constructions interesting and a second	HHWAIES	which surveys hh
las.	. steet TD YEAR	PRIMAIES	which surveys PS
mport excel "C-/Usen/Stu	panel variable: ID (strongly balanced)		
Over 10 YEAR	time variable: YBAR, 1993 to 1997 delta: 1 unit		
tdescibe .	071.5.4 400.5		
lear	. stdesribe		
use "C1Uses/Studio A/Des			
egenID = group(STATED D	10: 1, 2,, 33 n = 33		
ord ID SURVEY	YEAR: 1999, 1994,, 1997 T = 5 Delta(YEAR) = 1 moit		
vouet ID SURVEY Br	UPERATURAN = 1 mit Span(TEAE) = 5 periods		
B4	(ID'YTAR uniquely identifies each observation)		
	Distribution of T_1: min 54 254 504 754 515 max		
	5 5 5 5 5 5 5		
	Freq. Fercent Cun. Fattern		
	33 100.00 100.00 11111	Properties	
		Properties	9
	33 166.46 186.06 11111 33 166.46 20033		,
		8.44	a suiver
	33 166.40 XXXXX	· Variables	SURVEY IHDS1 (2005) or II
	Image: State of the state None clear	A Variables Name Label Spr	SURVEY INDS1 (2023) or it ret
	33 100 40 00000 cline	Warkables Name Label Type Format	908/vEV HIDST (2025) or HI HIT %7.0F
	31 166 30000 "data	Varlables Name Label Torret Volue label	9,085(1) 14051 (2005) or 14 146
	Image: State of the state None clear	Autobles Name Label Spet Format Vulue label Name	908/vEV HIDST (2025) or HI HIT %7.0F
	31 166 30000 "data	Varlables Name Label Torret Volue label	SURVEY HIDST (2025) or HI HIT SURVEY SURVEY
	Image: State State XXXX - Class XXXX XXXXX - State XXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	Warkbles Name Label Type Formut Whote label Notes Outs	SURVEY INDS1 (2003) or IH int SURVEY 1780-0011-Entrol
	31 106 300000 clase	Vorlables Norma Label Jose Foresut Wohne label Hotee Data Elimeanne	SURVEY INDS1 (2003) or IH int SURVEY 1780-0011-Entrol
	International Description 10 100.00 XXXXX XXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	A Variables News News Label Notes Description Description Description Notes Label Notes	SURVEY INDOT (2005) or IN INT OF SURVEY 21702-8011-8046 A India Human Devi 580
	Image: State State NOTE 104 of State S	A Vaciables Norme Norme Latel Gos Format Watch label Norme Constant Label Norme Label Norme Label Norme Collamona Distantion Distantion	SURVEY HIGS 10250 pr H H H SURVEY 1782-4011-8544 India Human Devi S0 8018
	International Description 10 100.00 XXXXX XXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	A Vachdar Norwe Later Tore Tore	SURVEY INDEX (2021) or II wet %7.07 SURVEY 17102-4011-4049. India Human Dev 500 IID 194 133.49M
	Image: State State NOTE 104 of State S	A Vaciables Norme Norme Latel Gos Format Watch label Norme Constant Label Norme Label Norme Label Norme Collamona Distantion Distantion	SURVEY HIGS 10250 pr H H H SURVEY 1782-4011-8544 India Human Devi S0 8018

So, I am going to clear it i.e., clear the earlier data. Now, we are going to open the IHDS data in front of you. First of all, we need to understand which are the common ID or the ID variables and those would be uniquely identified in the panel data.

So, we are generating an ID variable. So, "egen" is the command that is used to generate ID variable as per the instruction given in IHDS, you can search in Google in their webpage. You are making an ID variable through this command and there are 5 important variables i.e., starting from state ID, district ID, psu ID, HH ID, and household split ID.

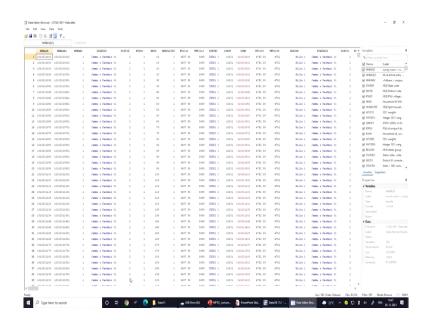
Now, you can see the state ID is mandatory, then districts, then psu, then primary sampling unit; then you come to the household level, then how household is getting split over time period, who are the final households. So, your split ID is also very important.

So, now we have defined an ID variable. So, ID variable with the name ID; now we can check whether these are actually uniquely identified or not. So, "isid" is the command that is used for ID, ID variable we have defined. Now, at the end it is there. And the time dimension has to be given as well to survey a period as per the data.

So, I know that these 2 are in fact uniquely identified, since it does not throw any error. Now, we can check whether this is a balanced or an unbalanced data. The xtset command is the one through which we can check it. So, it suggests that it is a strongly balanced data.

Strongly balanced data as the time period is from 1 to 2 as I already mentioned. We can also check it from the description command i.e., "xtdescribe" in case if you want and in the way that we have already guided. Also, we can check it through the browsing window; once open the browse window and through the display we are able to get it.

(Refer Slide Time: 29:48)



(Refer Slide Time: 29:58)



So, these are all the information in front of you and you can check own your own. And since, there is no missing points, so data is balanced and even if it is there but by the time and the cross sections, it is already covered. Though missing dots are there, but Stata reads it, as there are certain entries, so it is considered as a balance panel and strongly balanced panel.

Now, this helps you to work better with the panel data. I hope I have given you all a correct direction at this moment. Further, we will be explaining about pooled, fixed effect, random effect etc. panel models and it is applications along with their interpretation in the next class.

With this I look forward to your participation,.

Thank you.