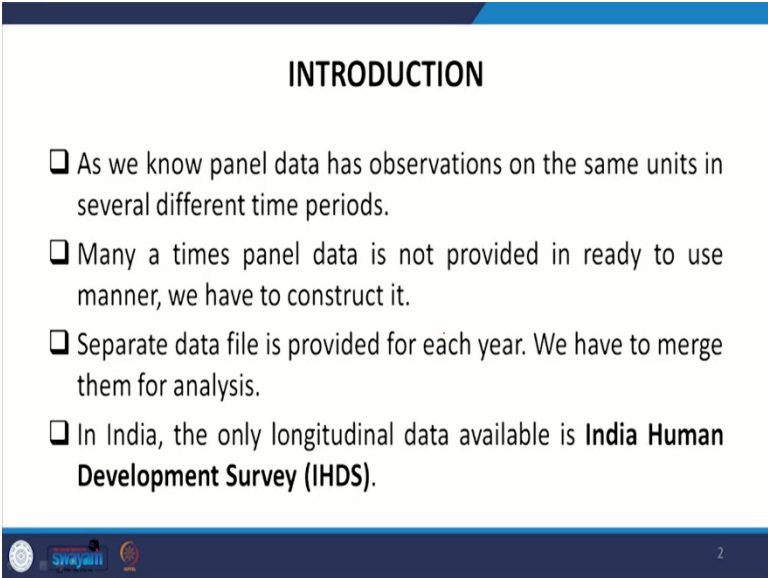


Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 39
Construction of Panel Data

Welcome once again to the NPTEL module on handling large scale data, using STATA. We are now explaining the panel data and their construction, how we can able to construct the panel data correctly.

(Refer Slide Time: 00:39)



INTRODUCTION

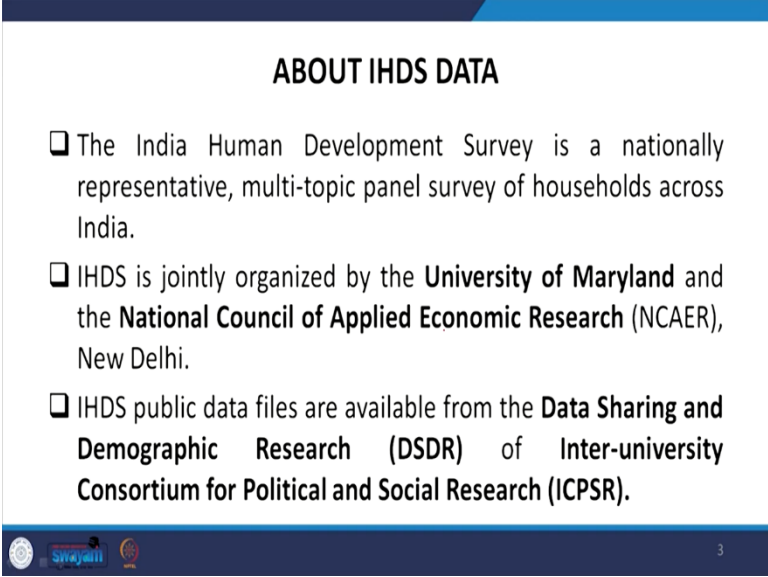
- ❑ As we know panel data has observations on the same units in several different time periods.
- ❑ Many a times panel data is not provided in ready to use manner, we have to construct it.
- ❑ Separate data file is provided for each year. We have to merge them for analysis.
- ❑ In India, the only longitudinal data available is **India Human Development Survey (IHDS)**.

2

And that will help us to go for panel model and the analysis very systematically and correctly. Coming to the understanding of panel data. Once again, we just wanted to emphasize that panel data has observation on the same units in several time periods. Many times the panel data is not provided in ready to use manner, and we have to construct them in a panel format.

The separate data file that is provided for each year is important. We have to merge them for our analysis. In India particularly, the only longitudinal data in the panel format is through India human development survey datasets. So, IHDS is very important that is unique in this particular direction. And we are just clarifying IHDS data once again though we have done it earlier, but in a panel setup we did not mention earlier.

(Refer Slide Time: 02:01)



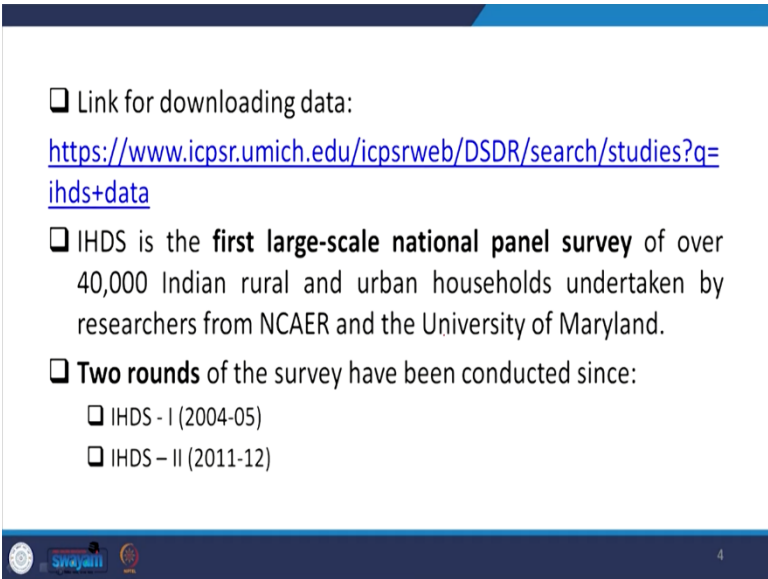
ABOUT IHDS DATA

- ❑ The India Human Development Survey is a nationally representative, multi-topic panel survey of households across India.
- ❑ IHDS is jointly organized by the **University of Maryland** and the **National Council of Applied Economic Research (NCAER)**, New Delhi.
- ❑ IHDS public data files are available from the **Data Sharing and Demographic Research (DSDR)** of **Inter-university Consortium for Political and Social Research (ICPSR)**.

3

I just wanted to mention once again that IHDS is a national representative data with multi panel survey of households across India. IHDS is jointly organized by University of Maryland. And the NCAER that is the National Council of Applied Economic Research based in New Delhi. IHDS public data files are available from the Data sharing and Demographic Research of inter University Consortium for Political and Social Research. That is ICPSR, I will show you all those links.

(Refer Slide Time: 02:37)



- ❑ Link for downloading data:
<https://www.icpsr.umich.edu/icpsrweb/DSDR/search/studies?q=ihds+data>
- ❑ IHDS is the **first large-scale national panel survey** of over 40,000 Indian rural and urban households undertaken by researchers from NCAER and the University of Maryland.
- ❑ **Two rounds** of the survey have been conducted since:
 - ❑ IHDS - I (2004-05)
 - ❑ IHDS – II (2011-12)

4

Here is the link ICPSR link once you click on this link, you will be directed to the exact website where the data is available. So, IHDS is the first large scale national panel survey of over 40 thousand Indian rural as well as urban households undertaken by the researchers from NCAER and University of Maryland. There are in fact two rounds of the survey and those have been conducted since 2004-05 till 2011-12. So, 2004-05 is famously known as IHDS one and second that is 2011-12 is called IHDS 2.

(Refer Slide Time: 03:18)

- ❑ IHDS-I surveyed **41,554** households consisting 26,734 rural and 14,820 urban households.
- ❑ Total number of household surveyed in **IHDS-II** is **42,152** households that consists of 27,579 rural and 14,573 urban.
- ❑ IHDS-II re-interviewed **83 per cent** of original households of IHDS-I as well as split households residing within the village and an **additional sample of 2,134 households**.

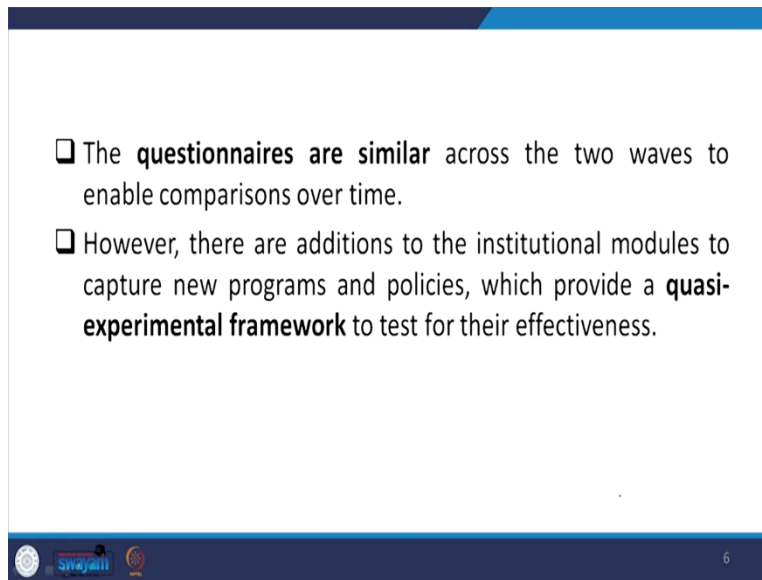
Households in IHDS	41,554
Households in IHDS-II	42,152
Households surveyed in both IHDS and IHDS-II	40,018
IHDS households lost to recontact for IHDS-II	6,911
IHDS-II households not included in IHDS	2,134

IHDS-1, surveyed over 41554 households. And these numbers are very important. While understanding panel, we need to remember these numbers because then only you can able to understand how many are dropped and how many are exactly matching. This is very very essential while constructing panel data. So, out of 41554 households, 26734 are from rural and 14820 are from urban households.

Total number of household, surveyed in IHDS-2 is of 42152 and those consist of again 27579 rural and 14573 urban households. IHDS-2 re-interviewed that is important to mention re-interviewed. We are going to clarify 83 percent of the original households. So, in the second wave, that is IHDS-2 considered 83 percent of the original that is IHDS-1 data as well as split households residing within the household and additional sample of 2134 households' additional samples.

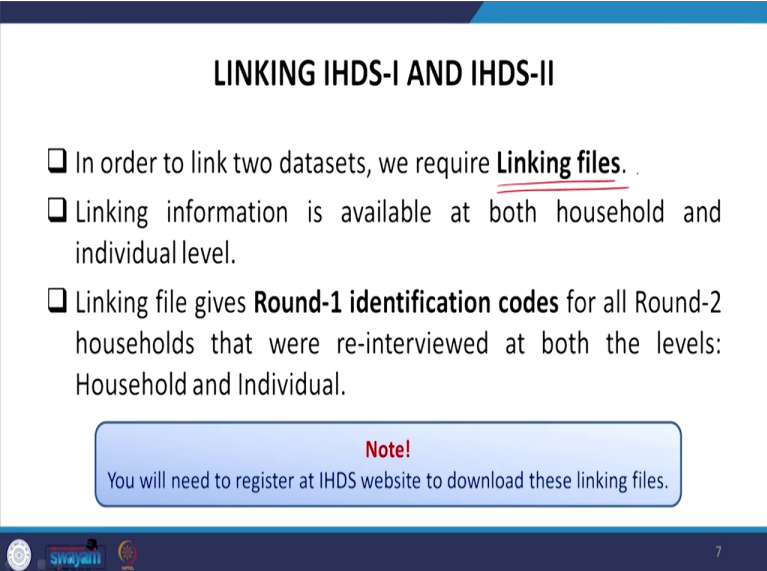
We are going to clarify everything. This is the number we mentioned and IHDS-2. It is of 42152 households surveying both round that those common is of 40018 and the household lost re-contacted in second wave and they are of 6911 and the households that are not included in the second round of the surveys of 2134. And why these are the cases we are going to clarify right now.

(Refer Slide Time: 05:32)



The questionnaires are in fact similar across the two waves to enable comparison over time. And it has been very clearly mentioned that two rounds can be compared. Why comparison important? Those who are interested in doing some experimental framework for analysis, in comparison some effectiveness of a particular policy over that two-time period, if it is somewhere suitable is going to be very useful.

(Refer Slide Time: 06:03)



LINKING IHDS-I AND IHDS-II

- ❑ In order to link two datasets, we require Linking files.
- ❑ Linking information is available at both household and individual level.
- ❑ Linking file gives **Round-1 identification codes** for all Round-2 households that were re-interviewed at both the levels: Household and Individual.

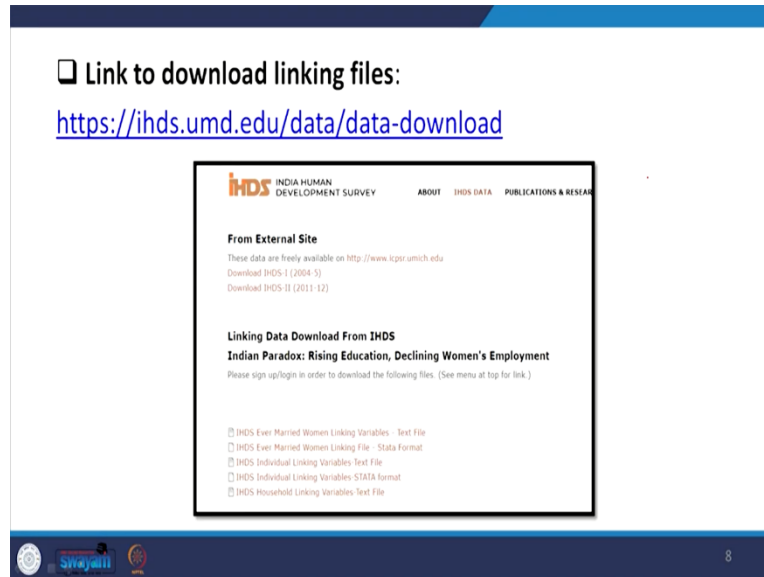
Note!
You will need to register at IHDS website to download these linking files.

7

Understanding the linking of the both the rounds of the data in order to link two datasets, we require a linking file that is important I am underlying it so, just to emphasize that whenever two rounds of data to be merged to make a panel which one are common by the exact unit, the exact respondent or the household is the unit of the survey. Those linking files must have been available.

So, linking information is available at both households and individual level in the IHDS website. So, linking files gives round one identification codes for all the round two households that were interviewed at both the levels, that is for individual level as well as household level. You will need to register at IHDS Website to download this linking data.

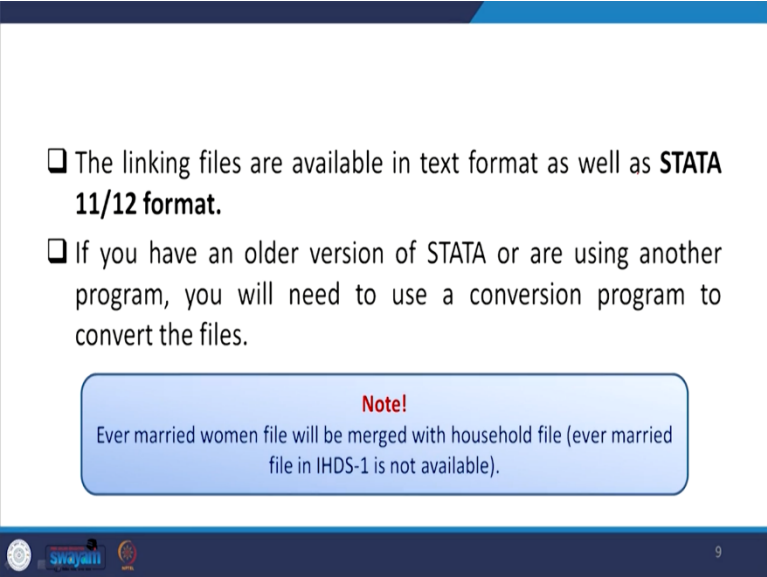
(Refer Slide Time: 07:06)



So, just registration. If you do it, it is freely available. You need to give minimum information for registration. You will be given this page to download the data. It gives information such as understanding ever married women. They are in also text file, in STATA file also, individual linking variables are available. Then household linking variables are also available.

Linking data is essential sometimes, as they also mentioned, that understanding rising level of education, declining women empowerment, some paradoxical changes in two rounds are also important.

(Refer Slide Time: 07:53)



- ❑ The linking files are available in text format as well as **STATA 11/12 format**.
- ❑ If you have an older version of STATA or are using another program, you will need to use a conversion program to convert the files.

Note!

Ever married women file will be merged with household file (ever married file in IHDS-1 is not available).



9

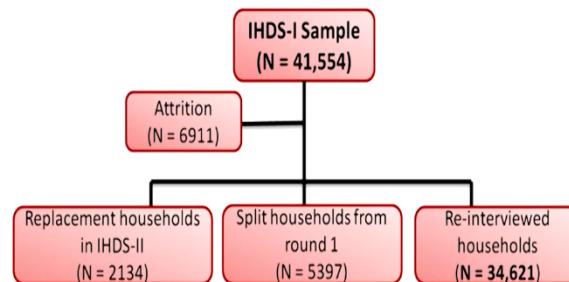
Coming to the linking files that are available in text format as well as in STATA format of 2011-12 format. If you have earlier version, you can convert it to the other version of the data also. A note here that ever married women file will be merged with the household file. it is important to mention, ever married women file will be merged with Household File since IHDS-1 does not give information specifically on ever married.

So, in that case, you have to merge with the household file and household file will be merged with certain limited information related to ever married women. Basically it is 15 to 49 age of women. Further clarifications on understanding the relationship between these two rounds for merging to make a panel data.

(Refer Slide Time: 09:05)

Understanding Relationship Between IHDS-I and IHDS-II

- ❑ Let us understand sampling structure of IHDS-II dataset first to get an idea about how many repeated households were surveyed from IHDS-I.



We first try to understand the structure of the data in both the rounds. In IHDS-1 sample, if we consider it in total, it is of 41554, and there were attrition number of observations. Those have been dropped is of 6911 may not be those observations are available. And there might have been some migration or that might have been some lapses. And there are various reasons of not covering the same individual household from the same location there might even change sifted to another place. So there is an attrition that will be subtracted from the original data.

How many replacements are there in that place in the new round that is of 2134 will clarify in our next slides in detail. There are some split households also like in the household gets splited to a number of observations like in different lead to different households, so those all have been included. That is of number 5397. Some re-interviewed households are there. This is the re-interviewed means whatever they are, out of 41000 straight away 31621 have been directly re-interviewed in the next round.

(Refer Slide Time: 10:40)

- ❑ **Attrition** is the process of drop out of individuals from the panel. Attrition here means IHDS-I households lost to re-contact in IHDS-II.
- ❑ Some of the original IHDS households had to be replaced in some urban areas where interviewers were unable to locate the former households.
- ❑ A replacement household was randomly selected in the same neighborhood to refresh the sample. This has led to 2,134 new households being included in the IHDS-II sample.
- ❑ Replacement households are identified by a **9** in HHSPLITID.



11

We will clarify this replacement and split household once again. Also, the attrition households are the one who are in the process get dropped out from the questioning round or the question round of the next level. Attrition it means that IHDS-1 household lost to re-contact in IHDS-2. Some of the original IHDS households had to be replaced in some urban areas where interviewers were unable to locate the former households,

So, those have been shifted or like replaced in some urban areas, very difficult to locate. They are also part of the attrition level. But a clarity is also given in the dataset related to the replacement household like a replacement household was randomly selected in the same neighbourhood to refresh the sample. This has led to 2134 new households being included in the IHDS-2 sample. So, that has been contacted to the neighbourhood and more or less the same nature of sample has been replaced. That is of 2134 new households. These are called new replacement household. So, replacement households are identified by a 9 in the split file 9 is the code entered for the replacement household.

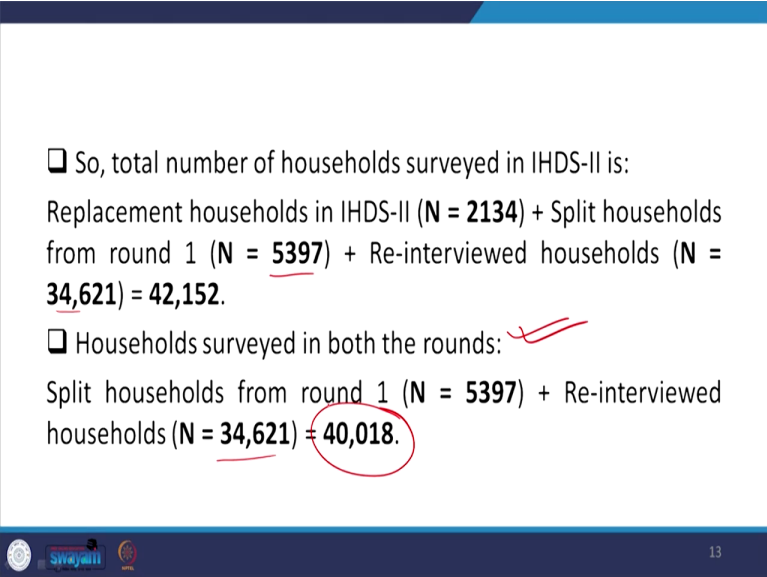
(Refer Slide Time: 12:20)

- ❑ It means there are some households which are not interviewed in IHDS-I (2134).
- ❑ Households from IHDS-I may have **split** into two or more households in IHDS-II. If the split households resided in the same village or urban block, all splits were interviewed.
- ❑ Because of this split, households in IHDS-I match with more than one household in IHDS-II.

It means there are some households which are not interviewed in IHDS-1. Basically, since they are replaced in a new survey. That is very clear that number that is of 2134 have not been included or were not including IHDS-1. So, households from IHDS-1 may have split into two or more households in IHDS-2. Like one household between IHDS-1 and IHDS-2 the number of years, from 2005 till 2011-12 around 5-6 years in between, there are high possibility of some split households.

One household gets split to two or more households. And they are all included in IHDS-2. If the split household resided in the same village or urban block. All splits were included for survey. Because of the split households in IHDS-1 match with more than one households in IHDS-2. Because of the split. So, the total number of household surveyed in IHDS-2 is what exactly that is.

(Refer Slide Time: 13:42)



□ So, total number of households surveyed in IHDS-II is:
Replacement households in IHDS-II (**N = 2134**) + Split households
from round 1 (**N = 5397**) + Re-interviewed households (**N = 34,621**) = **42,152**.

□ Households surveyed in both the rounds: ✓
Split households from round 1 (**N = 5397**) + Re-interviewed
households (**N = 34,621**) = **40,018**.

The slide features a blue header and footer. The footer contains logos for Swayam and other institutions, along with the number 13. Handwritten red annotations include a checkmark next to the second bullet point and a circle around the final result 40,018.

N equal to 2134, plus 2134. We have said that the new replacement households, and the split household that would of 5397 plus re-interviewed households. So, they are of 34621 in total it is of 42152. So, the total observation we get in IHDS-2 but they may not be entirely of longitudinal type, entirely of final types. The exact number we are going to discuss right now.

Households surveyed in both the rounds, this is more important to be noted for our analysis that is split household from around 5397 that we have already mentioned, plus the re-interviewed that is of this. So, 34621 if you add these two. So, both the rounds and their observations are now similar and that is of 40018 number similar in both rounds.

(Refer Slide Time: 14:54)

What we need for linking two datasets?

- ☐ IHDS-I household file (N = 41,554)
- ☐ IHDS-II household file (N = 42,152)
- ☐ Household linking file (N = 42,152)

Linking file will provide IHDS-I identification codes for all IHDS-II households and missing codes for replacement households (2134) that are new.

- ☒ **Linking Variables:** STATEID, DISTID, PSUID, HHID, HHSPLITID, PERSONID. (note: PERSONID is required for linking individual files).

The slide features a blue header and footer. The footer contains logos for the Government of India, the Ministry of Health and Family Welfare, and the National Institute of Population Research and Training (NIPORT), along with the number 14.

And this number is important for panel analysis. Coming to the linking of two datasets once again to mention that IHDS-1 household in total were of 41554 and in two it is 42152 household link file contains household linking, file contains the entire 42152 households so linking file will provide IHDS-1 identification codes for all IHDS-2 households and missing codes for replacement households that we have already mentioned, as 9 is the code with the number 2134, And those are very new.

Coming to the linking variables, mark it very carefully linking variables. They are specifically state ID, district ID, psu ID, HHID, household basically the ID label. Then this is of primary sample unit they have defined then with an id this is the split id and this is person ID. So, to be noted here that the person ID is required, especially this particular ID is required when we are linking for the individual files, not for the household files.

(Refer Slide Time: 16:18)

Steps to Link IHDS-I and IHDS-II:

Step 1: launch stata ->

Step 2: use "IHDS1_HH.dta", clear

Step 3: rename all the variables in IHDS-I

rename *x*

Note! There is a space between first asterisk symbol (*) and X.

This command adds a prefix x to all variables. Renaming is important because when we merge two files, it will overwrite the variables. So if you want to keep variables from wave 1 and wave 2 separate, you may need to rename all the variables of wave 1.

Coming to the steps of linking IHDS and IHDS-2, we need to launch the data now onwards, for another 15 to 20, around 13-14 slides will be experimenting with the original data for linking its one and two to make a panel data. With 40000, the number of which already 40018 is the number and wanted to find out by linking. So, let us use the dataset that is IHDS-1 household data. At this moment we are linking and we will open.

But I wanted to mention here that in order to get the link very clearly it might be the case that both the datasets that is IHDS-1 household and IHDS-2 household datasets have the same name, If the same name is there but we are linking from Masters to the new both the variable name might be merged with the same name and the information with the new variable that will be identified in the mod database will carry the master's variable name.

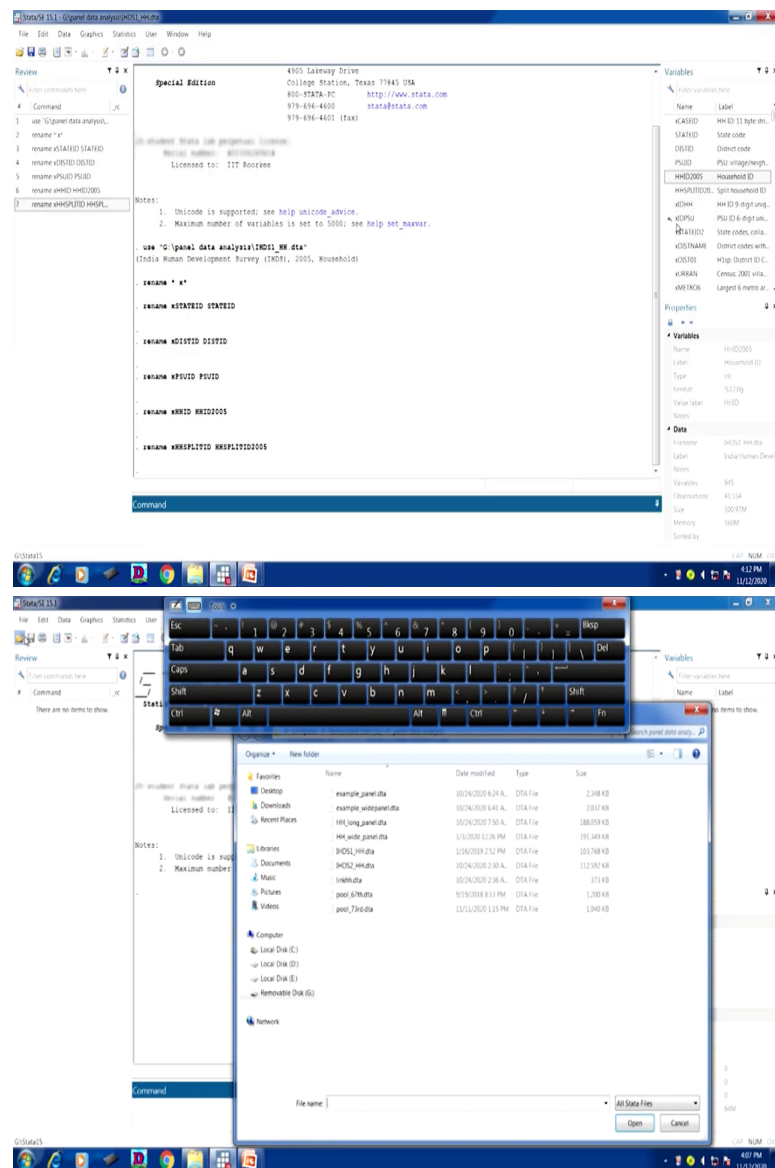
So, what does this mean basically when it has two variables separately and to both the variables, give different information. But, we are carrying the same name that will be misleading. Alright. Since both the variables are different in different rounds and they give different information, so in order to differentiate the variables, we are renaming the variable. first of all, we rename the one dataset that is IHDS-1.

So, you need to rename the entire set of variables. Those are given in IHDS-1 household data. I suggest that you please likewise do for renaming of the variable name of a particular variable, then its new name, it gets renamed. But since we want to differentiate by all the variables with all

the variables in another round, for simplicity we take x is another indicator addition to the name of the variable.

So, the exact suggestion is that there is space mark here between the asterisks to asterisks mark. There is a space here. So, we are highlighting here there is space between first asterisks symbol and X, in between this space mark, if you will just do that, all the variables will be added with x. We are going to show it, with the IHDS-1 household data. So, it is here.

(Refer Slide Time: 19:16)



Steps to Link IHDS-I and IHDS-II:

Step 1: launch stata ->

Step 2: use "IHDS1_HH.dta", clear

Step 3: rename all the variables in IHDS-I

rename *_x*

Note!

There is a space between first asterisk symbol (*) and X.

This command adds a prefix x to all variables. Renaming is important because when we merge two files, it will overwrite the variables. So if you want to keep variables from wave 1 and wave 2 separate, you may need to rename all the variables of wave 1.

So, we are going to open it. It is here. IHDS-1 household data this is the one, isn't it? And so we will rename. Then asterisks say space is there then x and with asterisk enter. You just mark all the variables having added with x, x was not there earlier. All the variables are added with x before the variable name you just mark very carefully. X is there in every variable alright. Coming to the same interpretation.

This is what we have clarified, this command adds a prefix of x to all variables. Renaming is important because when we merge two files, it will overwrite the variables. So, if you want to keep variables from wave one, that is IHDS-1 and we have to separate. You want to just keep all the variables separate. You may need to rename all the variables and we have already renamed.

(Refer Slide Time: 20:34)

Step 4: rename the id variables as original one.

The previous command renames all the variables in the data. But for merging two files we need id variables with same name that uniquely identifies the data in both the files.

```
rename xSTATEID STATEID
rename xDISTID DISTID
rename xPSUID PSUID
rename xHHID HHID2005
rename xHHSPLITID HHSPLITID2005
```

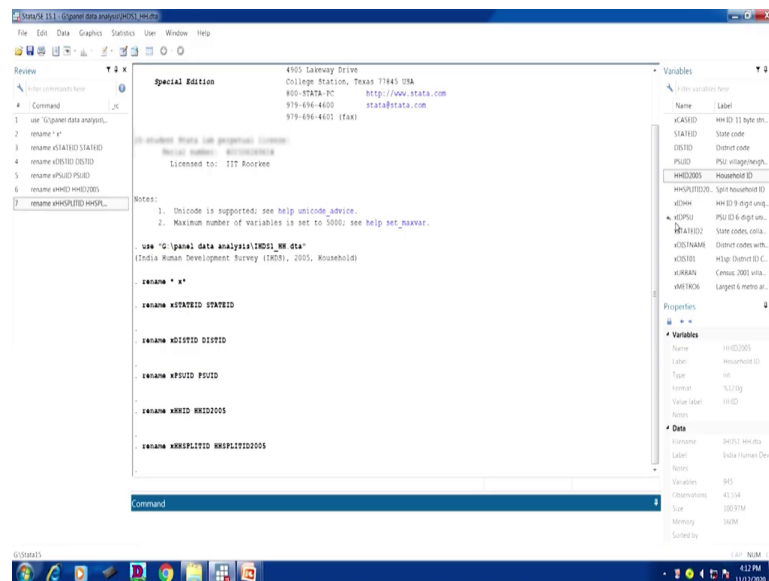
HHID and HHSPLITID is not same in both the waves. In linking file given with name HHID2005 and HHSPLITID2005, so renamed it as per requirement.

So, rename the id variables as original. One id variable to be original then only based on the id variable. Such as the similar id variable which are based on that. The same id will be carried to another one and that same information against the individual id will be added with different variables, different information. So, ID variables should be the same. What we will do? We will change the ID variable. We have now changed all ID variables to be added prefix with x, x is there. We will change to only the same name that is state ID.

Similarly, district ID, PSUID but another information to be very important that the household ID. Since we have changed the variable name of Particularly 2005. Our household ID is added with x but we wanted to mark that it is of 2005 households because in split ID we are going to the split folder, split file the variable name with HHID 2005 is given also a split household 2005 is given.

So, we have to make the renamed the household file to be HHID household ID with 2005. So, if you do that, simply renaming this we will get those name, which is going to be very useful. So, HHID and household split ID is not same in both the waves. Which I just said now. In linking file given with name HHID 2005. When HHID split ID2005. So, renamed it as per requirement. Since it is essential we will rename it. So, you can follow this step on your own and it is very easy to do alright. So, like we can do it so rename we will do it here. Alright. We will get to that clearly, so it is here, we will simply rename all those files. Alright.

(Refer Slide Time: 23:00)



Simply, if you have renamed those ID, specifically the common ID which we wanted to discuss that state ID, district ID, PSUID has been removed with x and others are also removed with x and we added 2005 against to this HHID 2005 and HHID split ID 2005 rest variables we have kept it with x as the prefix. That is one of the guidance, have done it.

We have very clearly marked the variables. Those are common, we will sort those as for merging because there are two rounds. In any case, we have to merge with their common ID. So, this common ID which we have already defined and it has been already instructed in IHDS rounds also which are the common ID for both the periods and for individual periods.

So, since these are the common ID, we will sort first. As per the merging guidance we have already given in earlier lectures. So, we will sort it. We will simply sort those five IDs state ID these, this, this and this, and we will save and replace. Alright. So, with a name, I think it has already been saved.

(Refer Slide Time: 24:46)

Step 5: sort these ID variables and save this file with different name.

```
sort STATEID DISTID PSUID HHID2005 HHSPLITID2005
```

```
save hh_ihds1
```

Step 6: open IHDS-II household data

```
use IHDS2_HH.dta, clear
```

Step 7: sort all id variables (note: if you want to check whether these identifiers uniquely identifies the data or not you can use `isid` command)

```
sort STATEID DISTID PSUID HHID HHSPLITID
```

17

So, it is with IHDS-1 alright.

(Refer Slide Time: 24:47)

The screenshot shows the Stata 15.1 interface. The Command window contains the following commands:

```
1. use "G:\panel data analysis\IHDS1_HH.dta"
2. rename "x"
3. rename xSTATEID STATEID
4. rename xDISTID DISTID
5. rename xPSUID PSUID
6. rename xHHID2005 HHID2005
7. rename xHHSPLITID2005 HHSPLITID2005
8. sort STATEID DISTID PSUID HHID2005 HHSPLITID2005
```

The Command window also shows the command `save hh_1h` at the bottom.

The Variables window on the right lists the following variables:

Name	Label
STATEID	State code
DISTID	District code
PSUID	PSU village/hamlet
HHID2005	Household ID
HHSPLITID2005	Split household ID
AGE	Age in years
AGE2	Age in years squared
AGE3	Age in years cubed
AGE4	Age in years to the fourth power
AGE5	Age in years to the fifth power
AGE6	Age in years to the sixth power
AGE7	Age in years to the seventh power
AGE8	Age in years to the eighth power
AGE9	Age in years to the ninth power
AGE10	Age in years to the tenth power
AGE11	Age in years to the eleventh power
AGE12	Age in years to the twelfth power
AGE13	Age in years to the thirteenth power
AGE14	Age in years to the fourteenth power
AGE15	Age in years to the fifteenth power
AGE16	Age in years to the sixteenth power
AGE17	Age in years to the seventeenth power
AGE18	Age in years to the eighteenth power
AGE19	Age in years to the nineteenth power
AGE20	Age in years to the twentieth power

The Properties window on the right shows the following information:

Variable	Label	Format	Value label
STATEID	State code	%10.0g	
DISTID	District code	%10.0g	
PSUID	PSU village/hamlet	%10.0g	
HHID2005	Household ID	%10.0g	
HHSPLITID2005	Split household ID	%10.0g	

So, you have already saved. So, we come to the guidance once again.

(Refer Slide Time: 24:54)

Step 5: sort these ID variables and save this file with different name.

```
sort STATEID DISTID PSUID HHID2005 HHSPLITID2005
```

```
save hh_ihds1
```

Step 6: open IHDS-II household data

```
use IHDS2_HH.dta, clear
```

Step 7: sort all id variables (note: if you want to check whether these identifiers uniquely identifies the data or not you can use `isid` command)

```
sort STATEID DISTID PSUID HHID HHSPLITID
```

Coming to another step, the first step related to sorting, the important common IDs we have made. Now will process it for the IHDS-2 household. Similarly, will open IHDS-2 household data because now we are merging the household data. So, in this case, again, let me open that first. So, let me clear it here. So, where is clear.

(Refer Slide Time: 25:23)

The screenshot displays the Stata 15.1 command window with the following commands entered:

```
1. use "G:\panel data analysis\IHDS1_HH.dta"
2. rename "x"
3. rename STATEID STATEID
4. rename DISTID DISTID
5. rename PSUID PSUID
6. rename HHID HHID2005
7. rename HHSPLITID HHSPLITID2005
8. sort STATEID DISTID PSUID HHID2005 HHSPLITID2005
9. save hh_ihds1
10. clear
11. use "G:\panel data analysis\IHDS2_HH.dta"
```

The right-hand pane shows the variable list for the current dataset, including:

- STATEID: HHSplit household
- DISTID: HHSplit household
- PSUID: HHSplit household
- HHID: HHSplit household
- HHSPLITID: HHSplit household
- EDUC: PSID of unique 6 d...
- EDUC: Household id un...
- WT: Weights in IHDS-II
- INCAT: Weights in IHDS-II
- INCAT: Integer weights in...
- INCAT: Integer weights in...
- DISTID: District ID correct...

The bottom status bar indicates the current command is `sort STATEID DISTID PSUID HHID HHSPLITID`.

Step 5: sort these ID variables and save this file with different name.

```
sort STATEID DISTID PSUID HHID2005 HHSPLITID2005
```

```
save hh_ihds1
```

Step 6: open IHDS-II household data

```
use IHDS2_HH.dta, clear
```

Step 7: sort all id variables (note: if you want to check whether these identifiers uniquely identifies the data or not you can use `isid` command)

```
sort STATEID DISTID PSUID HHID HHSPLITID
```



17

So, we will open the household two. Alright. This one, so this has been opened. It is in front of you, we will sort the important variables, those common IDs. Alright. So, this has already been sort. So, this step for you have now guided to you so sort all those indicator or important identifier, unique variables. We also check, check the IHDS ID as well. Just a minute, we have already sorted it out.

(Refer Slide Time: 26:10)

```
isid STATEID DISTID PSUID HHID HHSPLITID
```

Step 8: save the sorted file

```
save, replace
```

Step 9: open linking file

```
use linkhh.dta, clear
```

Step 10: sort linking IDs in linking file

```
sort STATEID DISTID PSUID HHID HHSPLITID
```

Step 11: merge IHDS-II household file to household linking file

```
merge 1:1 STATEID DISTID PSUID HHID HHSPLITID using  
"IHDS2_HH", gen (_merger2link)
```



18

We can check their IHDS also, whether they are uniquely identified or not.

(Refer Slide Time: 26:13)

```

1. use "G:\panel data analysis\IND02_MH.dta"
2. rename STATEID DISTID
3. rename ADISTID DISTID
4. rename AFPUID PSUID
5. rename AHUID HHID
6. rename HHID2005 HHID2005L
7. rename HHID2005L HHID2005R
8. sort STATEID DISTID PSUID
9. save hh_1dta1
10. clear
11. use "G:\panel data analysis\IND02_MH.dta"
12. sort STATEID DISTID PSUID
13. merge STATEID DISTID PSUID hh_1dta1
14. sort STATEID DISTID PSUID
15. save STATEID DISTID PSUID

```

Simply will change a two IHDS. So, variable. These are not uniquely identifying the observations. Alright, State ID then district ID, PSUID then household, ID then split ID, 1, 2, 3, 4, 4 were only mentioned earlier. 4 to be 5, So, that is why there are some error mark. But it has been corrected. After understanding their unique identifiers, we have also saved and we have also sorted those file we will open the linking file, we will save it then save and replace then we will find the linking files.

(Refer Slide Time: 27:26)

```

1. use "G:\panel data analysis\IND02_MH.dta"
2. rename STATEID DISTID
3. rename ADISTID DISTID
4. rename AFPUID PSUID
5. rename AHUID HHID
6. rename HHID2005 HHID2005L
7. rename HHID2005L HHID2005R
8. sort STATEID DISTID PSUID
9. save hh_1dta1
10. clear
11. use "G:\panel data analysis\IND02_MH.dta"
12. sort STATEID DISTID PSUID
13. merge STATEID DISTID PSUID hh_1dta1
14. sort STATEID DISTID PSUID
15. save STATEID DISTID PSUID
16. use replace
17. clear
18. use "G:\panel data analysis\IND02_MH.dta"
19. sort STATEID DISTID PSUID

```

Linking files are here. Let me just open, It is here. Then we will go for understanding the linking files. It is linked HHID. Alright. So, the linking file is already given in IHDS website for our clarification. I just wanted to give some information that it gives the unique identifiers the state ID, district ID, PSUID. , interestingly, it gives information about HHID.

And split ID of that without the name of 2012, but only given HHID. This means it is of 2012 household ID and split ID, households split ID that mean this is of 2012 and these are these last two are of household ID of 2005 and household split ID of 2005. So, while we merge the two dataset we have to take very carefully about which linking files we are supposed to connect.

(Refer Slide Time: 28:47)

```
isid STATEID DISTID PSUID HHID HHSPLITID
```

Step 8: save the sorted file

```
save, replace
```

Step 9: open linking file

```
use linkhh.dta, clear
```

Step 10: sort linking IDs in linking file

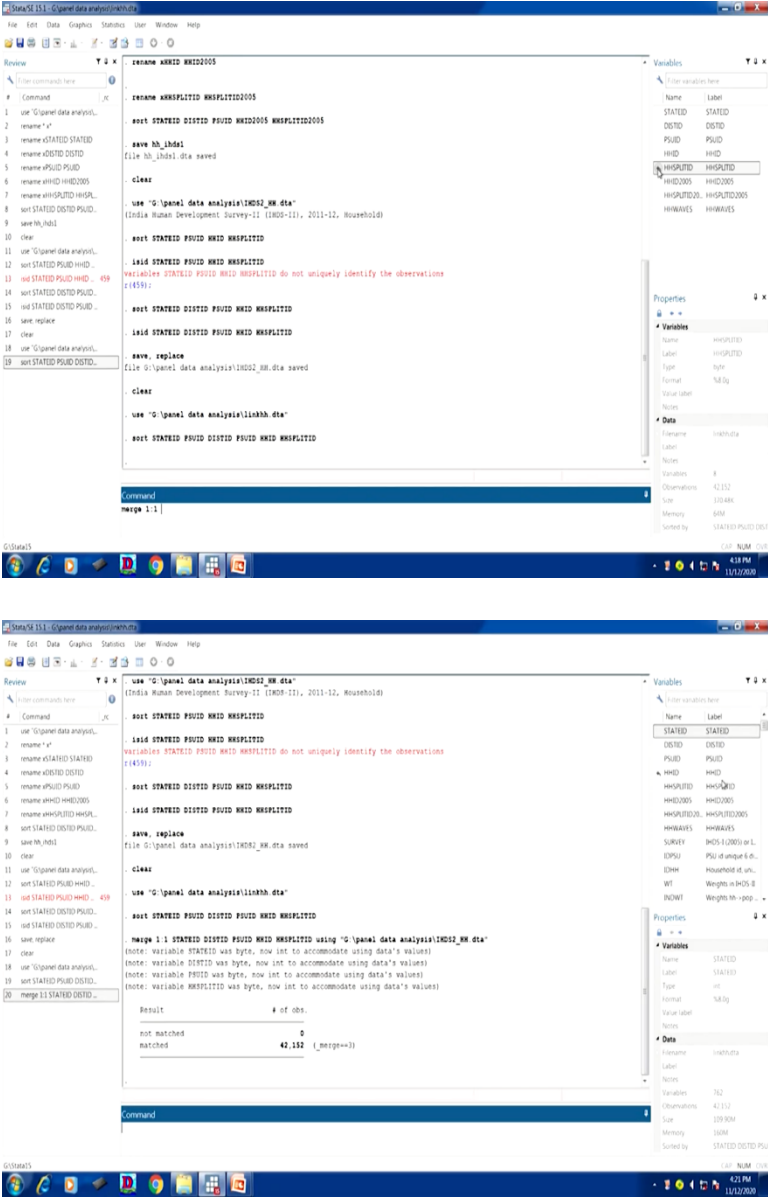
```
sort STATEID DISTID PSUID HHID HHSPLITID
```

Step 11: merge IHDS-II household file to household linking file

```
merge 1:1 STATEID DISTID PSUID HHID HHSPLITID using  
"IHDS2_HH", gen (_mergeR2link)
```

Our important linking files for merging two datasets like linking file with IHDS-2. So, we have to sort those 5 ID. So, as per 2011-12 data, these three, four, five. Alright. So, we have already sorted out. We can also check the IHDS ID as well. But since it has already been defined so no need to check. We will simply merge, merge it here,

Journal of Management Education 36(7) 809-824



So, it is basically 1 to 1 merging and 3. So, this is 5 IDs using our linking file. So, using our IHDS-2 here. You can close. So, this we have already sorted out all its unique identifier, alright. It is already opened. So, it is here then once again we will open. Alright. So, IHDS-2 we have used now there is a merging, we have gone through the merging. The merging has already been completed at this moment.

You can mark, which we have guided from the beginning, that 42152 are the total observation in IHDS-2 and the linking file also the entire information were also available for and that to we followed one to one merging. So, that is very clearly understood and our merging has been

(Refer Slide Time: 31:22)

isid STATEID DISTID PSUID HHID HHSPLITID

Step 8: save the sorted file

save, replace

Step 9: open linking file

use linkhh.dta, clear

Step 10: sort linking IDs in linking file

sort STATEID DISTID PSUID HHID HHSPLITID

Step 11: merge IHDS-II household file to household linking file

merge 1:1 STATEID DISTID PSUID HHID HHSPLITID using
"IHDS2_HH", gen (_mergeR2link)

So far since we are trying to link and merge the round 2 dataset, we can generate file with this name. So, already generated.

(Refer Slide Time: 31:40)

File Edit Data Graphics Statistics User Window Help

Review

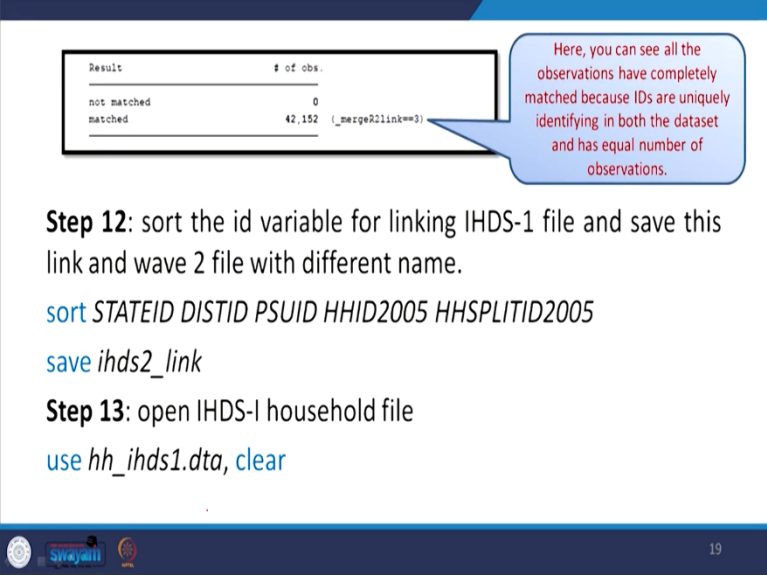
```

1 use "G:\panel data analysis\IHS02_RK.dta"
2 (India Human Development Survey-II (IHS02-II), 2011-12, Household)

3
4 Command
5
6 use "G:\panel data analysis\IHS02_RK.dta"
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1
```

I can show you that is already there at the end. So, this can be renamed with the merge. Since merge either merge is to right link? So, we have created a variable with this particular name. Alright. So, we will come back to this further clarification.

(Refer Slide Time: 32:31)



The slide displays a table with the following data:

Result	# of obs.
not matched	0
matched	42,152 (_mergeR2link==3)

A callout box points to the 'matched' row with the text: "Here, you can see all the observations have completely matched because IDs are uniquely identifying in both the dataset and has equal number of observations."

Step 12: sort the id variable for linking IHDS-1 file and save this link and wave 2 file with different name.

```
sort STATEID DISTID PSUID HHID2005 HHSPLITID2005
```

Step 13: open IHDS-1 household file

```
use hh_ihds1.dta, clear
```

At the bottom of the slide, there are logos for IIT Bombay and Synovate, and the number 19.

So, this is what we have guided. And here you can see all the observations that have already been completely merged because IDs are uniquely identifying both the datasets and has equal number of observations. Coming to the next step, that sort the ID variables of IHDS-1 and the IHDS-2 and the linking file, we have already merged. We have a combined dataset of that.

We need to get the IHDS-1 into it. So, IHDS-1 we already saved by sorting their unique identifiers. We will save that then we have two file with different name. There will also save it accordingly. What is suggested here that you need to sort identifier of IHDS-1, that we have already done it.

So, we saved it with this name as well. But this is the file we have created basically the merge file we have created. But we will try to merge with IHDS-1. IHDS-1 contains the unique file and information like this and this. Since this file has not been sorted with this information. First duty is to sort with these names, these variables, then only we can able to merge correctly. So, let us sort these five indicators.

(Refer Slide Time: 34:11)

```

clear
use "G:\panel data analysis\IHDS1.dta"
use "G:\panel data analysis\IHDS2.dta"
merge 1:1 STATEID DISTID PSUID HHID2005 HHSPLITID2005
using "ihds2_link.dta", gen(_mergeR1R2)

```

Result

	# of obs.
not matched	0
matched	42,152 (_mergeR1R2)

So, sort just a minute. It is here, so 1, 2, 3, then 5. So, we have sorted out. So, this we are naming it, save with IHDS-2 with link IHDS-2 and link file has already been saved. So, what is the next important aspect we will open the IHDS-1 household file that those IDs we have already sorted in IHDS-1. So, we need not sort further.

(Refer Slide Time: 35:00)

Step 14: merge ihds2_link file with IHDS-I file

```

merge 1:m STATEID DISTID PSUID HHID2005 HHSPLITID2005
using "ihds2_link.dta", gen(_mergeR1R2)

```

Result	# of obs.
not matched	9,045
from master	6,911 (_mergeR1R2=1)
from using	2,134 (_mergeR1R2=2)
matched	40,018 (_mergeR1R2=3)

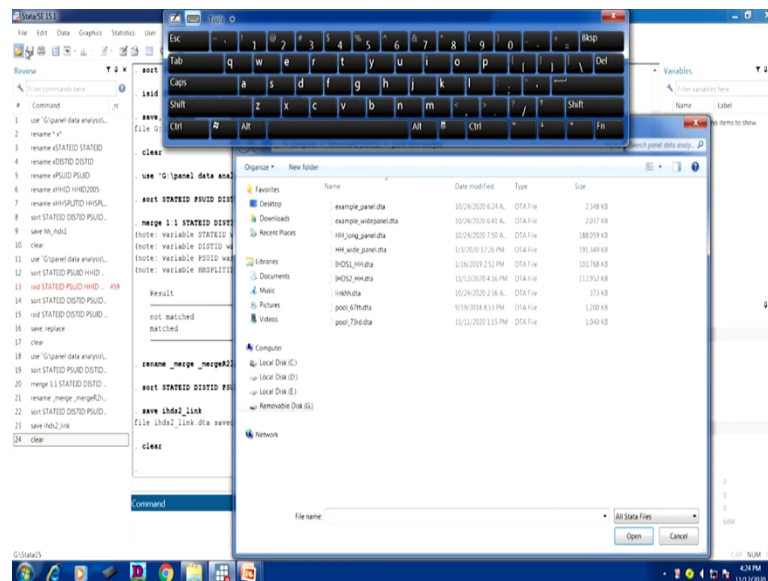
Number of households surveyed only in IHDS-I. attrition in case of panel (6911)

Households surveyed only in IHDS-II. Nothing but replacement households in this case

Here, you can see the total number of household interviewed in both the rounds are 40,018 are matched.

We will open that first then we will merge with the IHDS-2 link file. So, let me first clear this out. So, we will clear now will open IHDS-1 Household file.

(Refer Slide Time: 35:19)



So, IHDS-1 and that we have already sorted has been opened. Now, what will do? will merge.

(Refer Slide Time: 35:43)

Result	# of obs.
not matched	0
matched	42,152 (_mergeR2link==3)

Here, you can see all the observations have completely matched because IDs are uniquely identifying in both the dataset and has equal number of observations.

Step 12: sort the id variable for linking IHDS-1 file and save this link and wave 2 file with different name.

```
sort STATEID DISTID PSUID HHID2005 HHSPLITID2005
```

save ihds2_link

Step 13: open IHDS-I household file

```
use hh_ihds1.dta, clear
```

So, we have used this file, isn't it? And I think we have already sorted it.

(Refer Slide Time: 35:49)

Step 14: merge ihds2_link file with IHDS-I file

```
merge 1:m STATEID DISTID PSUID HHID2005 HHSPLITID2005  
using "ihds2_link.dta", gen (_mergeR1R2)
```

Result	# of obs.	
not matched	9,045	
from master	6,911	(_mergeR1R2==1)
from using	2,134	(_mergeR1R2==2)
matched	40,018	(_mergeR1R2==3)

Number of households surveyed only in IHDS-I. attrition in case of panel (6911)

Households surveyed only in IHDS-II. Nothing but replacement households in this case

Here, you can see the total number of household interviewed in both the rounds are 40,018 are matched.

All those indicators. So, we are supposed to go for merging. So, we will merge it with this information, we will straight away take this command, straight away copied. Why copied because we will carry with the same name generate the way we defined earlier will go by that only. And this is here.

(Refer Slide Time: 36:20)

The screenshot shows the Stata 15.1 Command Window with the following commands and output:

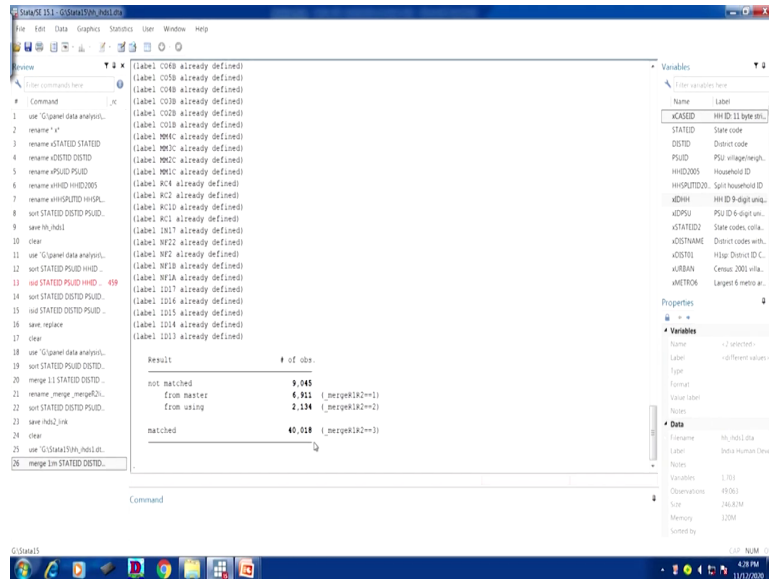
```
use "G:\panel data analysis\IHDS2_08.dta" saved  
clear  
use "G:\panel data analysis\linkhh.dta"  
sort STATEID PSUID DISTID HHID HHSPLITID  
merge 1:1 STATEID DISTID PSUID HHID HHSPLITID using "G:\panel data analysis\IHDS2_08.dta"  
(note: variable STATEID was byte, now int to accommodate using data's values)  
(note: variable DISTID was byte, now int to accommodate using data's values)  
(note: variable PSUID was byte, now int to accommodate using data's values)  
(note: variable HHSPLITID was byte, now int to accommodate using data's values)  
Result # of obs.  
not matched 0  
matched 42,152 (_merge==3)  
clear  
rename _merge _mergeR1R2  
sort STATEID DISTID PSUID HHID2005 HHSPLITID2005  
save ihds2_link  
file ihds2_link.dta saved  
clear  
use "G:\State\15\hh_1hds.dta"  
(India Human Development Survey (IHDS), 2005, Household)  
clear  
merge 1:m STATEID DISTID PSUID HHID2005 HHSPLITID2005 using "ihds2_link.dta", gen (_mergeR1R2)
```

The Variables window on the right shows the following variables:

- NAME: HHID 11 type int
- STATEID: State code
- DISTID: District code
- PSUID: PSU village/hamlet
- HHID2005: Household ID
- HHSPLITID2005: Split Household ID
- HHID9: HHID 9-digit uniq.
- PSUID9: PSU ID 9-digit uniq.
- STATEID2: State codes, coll.
- DISTNAME: District codes with
- VILLAGE: Village District ID
- URBAN: Census 2001 villa
- WETROK: Largest 6 metro ar...

The Command window shows the following command:

```
merge ihds2_link file with IHDS-I file  
merge 1:m STATEID DISTID PSUID HHID2005 HHSPLITID2005 using "ihds2_link.dta", gen (_mergeR1R2)
```



I think merge know the first line to be. So, it is the link to be given. So, that is basically here using file that has to be first deleted. Then now this is not there it is here. The link we have created, alright. Basically, after using the path name correctly, we observed that we have successfully matched the correct number that is of 40018. And from the Masters, those who are not included.

That is of 6911. And from the using file, that is basically the merge file, the replacement numbers that is up to 2134 we guided from the beginning. Those are not included. So, the net number to be defined as the merge is the panel information is of 40018.

(Refer Slide Time: 38:01)

Step 14: merge ihds2_link file with IHDS-I file

```
merge 1:m STATEID DISTID PSUID HHID2005 HHSPLITID2005  
using "ihds2_link.dta", gen (_mergeR1R2)
```

Result	# of obs.	
not matched	9,045	
from master	6,911	(_mergeR1R2==1)
from using	2,134	(_mergeR1R2==2)
matched	40,018	(_mergeR1R2==3)

Number of households surveyed only in IHDS-I. attrition in case of panel (6911)

Households surveyed only in IHDS-II. Nothing but replacement households in this case

Here, you can see the total number of household interviewed in both the rounds are 40,018 are matched.

20

And we will clarify further. So, basically when we have three indicators, that one code one, 1, 2 and 3 are there. So, what we are interested, only for 3, because three are the common IDs, common information in both round. So, we can draw for one and two because these are no longer important for the analysis. So far as panel is concerned, we can draw and can proceed accordingly.

(Refer Slide Time: 38:34)

Note!

You must have noticed in case of first merging (Link + IHDS-II) we have used 1:1 merging while in second merging (IHDS1 + IHDS2_Link), we have used 1:m merging. In first case information about all the households were given in linking file. In second case its multiple merging because some of the IHDS-I households are split into two or more but they share same household ID from IHDS-I.

❑ Total number of households would be:
 $6911 + 2134 + 40,018 = 49,063$ households.

❑ For panel data analysis keep only those households that are interviewed in both the rounds (40,018).

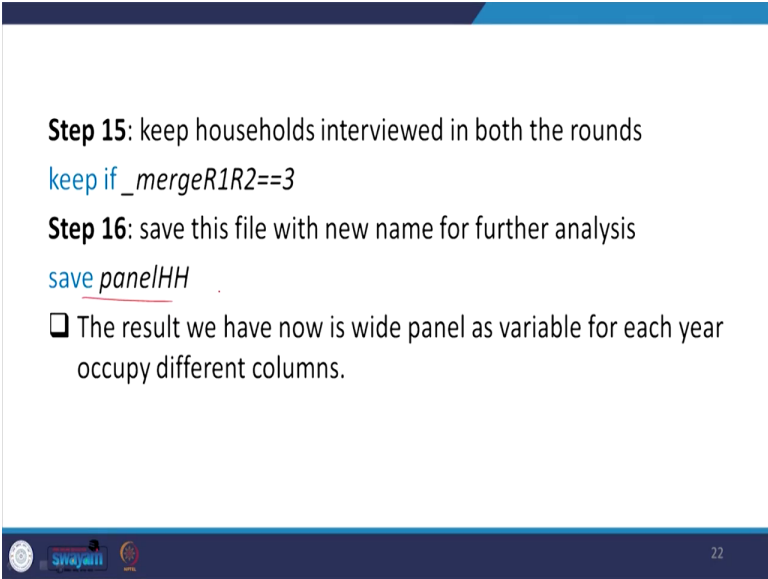
21

So, you must have noticed in the case of first merging, that is the link with IHDS-2. We have used with one is to one merging dividing while in the second one, we merge the IHDS-1 with the

combined link file. So, we applied one to one multiple file. That is merging technique. In first case, information about all the households were given in linking file. In the second case, it's multiple merging because we know some of the IHDS-1 households are split into two or more. And they share same household ID from IHDS-1.

Since there are more numbers and because of split we applied one to m as the merging technique. so the 2 common the total number of household would be in this case is 6911, 2134, 40018 that is of 49063 households so far as the total number is concerned. But out of that 40018 is the final number and they are interpreting both the rounds. Alright.

(Refer Slide Time: 39:44)



Step 15: keep households interviewed in both the rounds
`keep if _mergeR1R2==3`

Step 16: save this file with new name for further analysis
`save panelHH`

☐ The result we have now is wide panel as variable for each year occupy different columns.

So, we can keep that 3. You can do this experiment on your own. We are not repeating. And next, you can simply save with the panel HHID, and that has been already guided and will show while we are using that data for further analysis. So, coming to the dataset which we have already shown for your analysis to make a panel version. You can also click on this link to get for those datasets directly.

(Refer Slide Time: 40:17)

- ❑ Recently, the panel version of IHDS data has been populated on DSDR website.

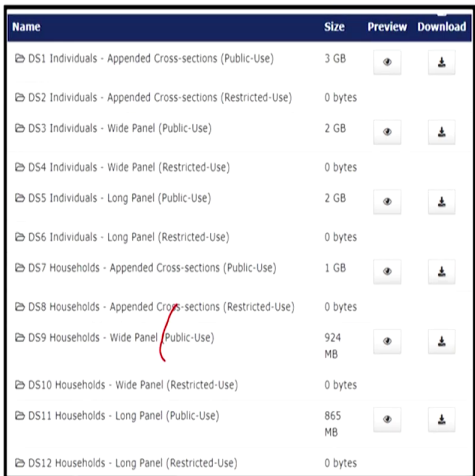
Web-link:

<https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/37382>

- ❑ The data is available in wide panel, long panel and pooled cross section or appended form.
- ❑ You can choose your version as per your requirement.

And readily available provided by the ICPSR and they are providing in wide panel, long panel, pooled cross section and appended form as well.

(Refer Slide Time: 40:31)



Name	Size	Preview	Download
DS1 Individuals - Appended Cross-sections (Public-Use)	3 GB		
DS2 Individuals - Appended Cross-sections (Restricted-Use)	0 bytes		
DS3 Individuals - Wide Panel (Public-Use)	2 GB		
DS4 Individuals - Wide Panel (Restricted-Use)	0 bytes		
DS5 Individuals - Long Panel (Public-Use)	2 GB		
DS6 Individuals - Long Panel (Restricted-Use)	0 bytes		
DS7 Households - Appended Cross-sections (Public-Use)	1 GB		
DS8 Households - Appended Cross-sections (Restricted-Use)	0 bytes		
DS9 Households - Wide Panel (Public-Use)	924 MB		
DS10 Households - Wide Panel (Restricted-Use)	0 bytes		
DS11 Households - Long Panel (Public-Use)	865 MB		
DS12 Households - Long Panel (Restricted-Use)	0 bytes		

For whichever the requirement you have you can download and start using it. So, these are all information given for clarity. You can use it for better explanation further alright. So, I think we need to proceed for the use of all those data that we have already converted, we have made panel and in the next class we will be suggesting you the exact technique of deriving result based on

the panel. So, that is all for today's lecture. Next lecture will be purely understanding and analysis of the panel data. Thank you so much.