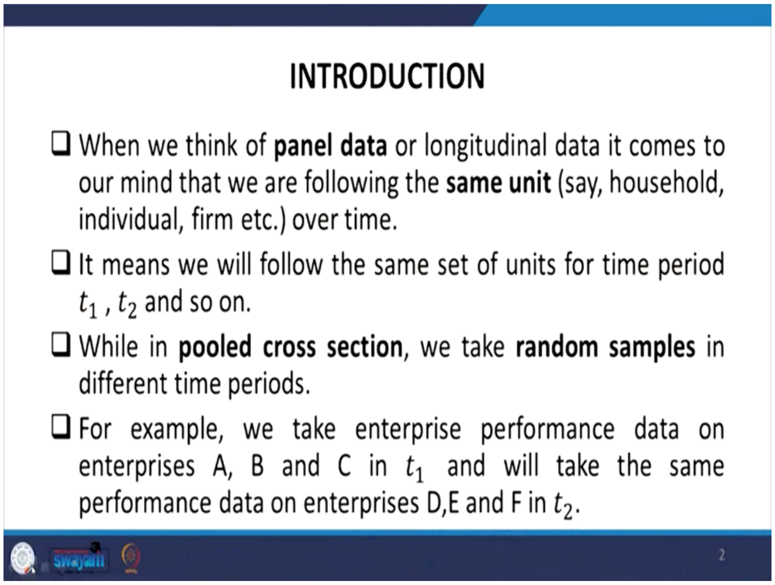
Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 38
Pooled Cross-Sectional Data

Welcome friends once again to the NPTEL MOOC module on handling large scale unit level data using STATA. We are at the verge of understanding panel data. And, you have already been given the background of the panel data in the previous 2 lectures. In the last three lectures of the entire module, will be giving the hands-on experience of dealing with the panel data. So, this particular lecture is dealt with a form of panel, but clearly called as pooled cross-sectional data. So, with the meaning of pooled cross-sectional and panel and there are differences we already given. But some other details we are going to explain to you in this module.

(Refer Slide Time: 1:31)


INTRODUCTION

❑ When we think of **panel data** or longitudinal data it comes to our mind that we are following the **same unit** (say, household, individual, firm etc.) over time.

❑ It means we will follow the same set of units for time period t_1 , t_2 and so on.

❑ While in **pooled cross section**, we take **random samples** in different time periods.

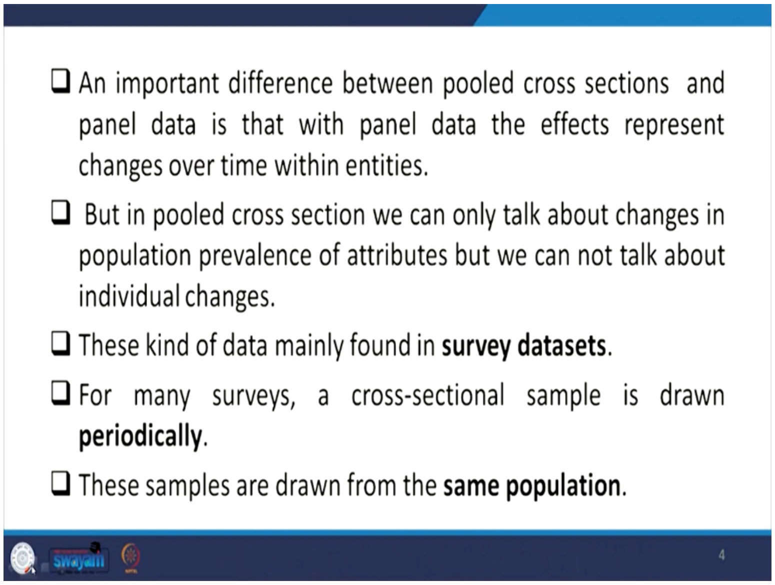
❑ For example, we take enterprise performance data on enterprises A, B and C in t_1 and will take the same performance data on enterprises D,E and F in t_2.

So, when we think of panel or longitudinal data, it comes to our mind that we follow the same unit which gets repeated in different time period. This means that we will follow the same set of units of the observation over time, maybe of t1 period, t2 period and so on. While we discuss the full cross section, we take random samples in different time periods. So, our samples are completely random. It is not repeated. For example, when we take enterprise performance data or we take the performance data on enterprises. That is A, B, C in t1 and we will take the same performance data on enterprises that is D, E and F in t2, then the enterprises we are referring not same. They are different in a different time period.

So, it is not repeated. The data are same, that means in terms of the information they are same, in terms of the variable they are same but the samples are different. Like we have different enterprises in our example and the key difference between pure panel data and pool cross-section is that the unit on which the data are collected is different. Units are to be compared. The fact that the random samples are collected independently of each other suggests that there need not be equal size.

The data can be pretty much analysed like ordinary cross-sectional data, except that we must use dummies, in order to differentiate the time period or the kind of changes in the enterprises over time. So, dummy method is going to be very useful in this setup. An important difference between pooled cross section and panel data is that with panel data, the effects represent changes over time within entities.

(Refer Slide Time: 03:48)



❑ An important difference between pooled cross sections and panel data is that with panel data the effects represent changes over time within entities.

❑ But in pooled cross section we can only talk about changes in population prevalence of attributes but we can not talk about individual changes.

❑ These kind of data mainly found in **survey datasets**.

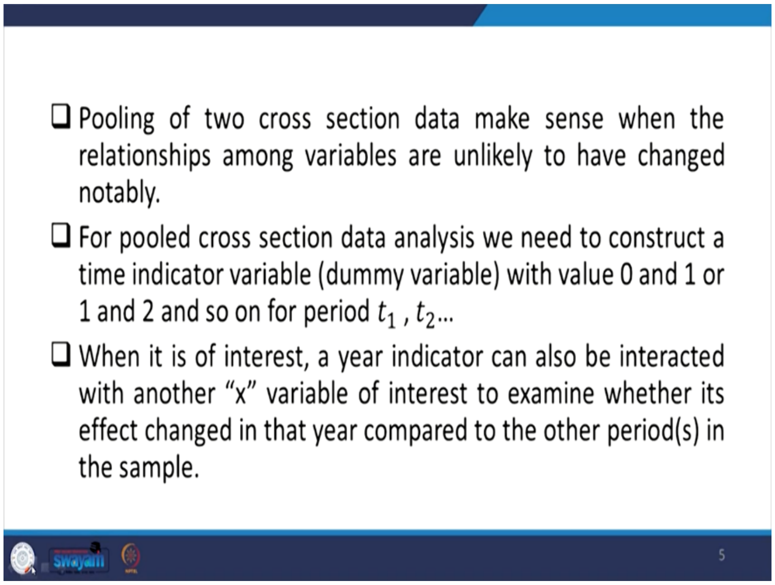
❑ For many surveys, a cross-sectional sample is drawn **periodically**.

❑ These samples are drawn from the **same population**.

4

But in pooled one, we can only talk about changes in population prevalence of attributes, but we cannot talk about individual changes over time. These kind of data mainly found in survey datasets. And for many surveys, a cross-sectional sample is drawn periodically. These samples are drawn from the same population.

(Refer Slide Time: 04:09)



□ Pooling of two cross section data make sense when the relationships among variables are unlikely to have changed notably.

□ For pooled cross section data analysis we need to construct a time indicator variable (dummy variable) with value 0 and 1 or 1 and 2 and so on for period t_1 , t_2...

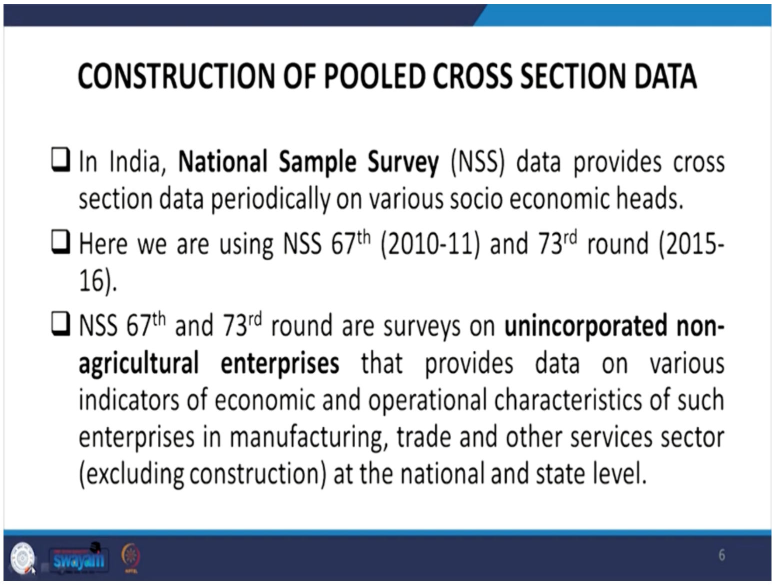
□ When it is of interest, a year indicator can also be interacted with another "x" variable of interest to examine whether its effect changed in that year compared to the other period(s) in the sample.

Pooling of two cross-section data makes sense when the relationship among variables are unlikely to have changed notably. So, if there are no such major changes expected, then pooling is generally preferred for pooling cross-sectional data analysis. We need to construct a time indicator variable that is dummy variable, let it be value 0 and 1 or 1 or 2 and so on by any binary form to define the dummy, to indicate time period 1 and time period 2 can be done.

When it is of interest a year indicator can also be interacted with another explanatory variable that is X of interest to examine whether its effect change in that year compared to other features. For example, in the Enterprises cases when we take another enterprise is taking loan in one period and taking loan in another period.

What is the difference in terms of enterprise performances in that case, the kind of loan or credit with the time factor interacted with the time factor has important interpretations? So, interaction dummy could be considered while making a pooled cross-sectional data over time.

(Refer Slide Time: 05:47)



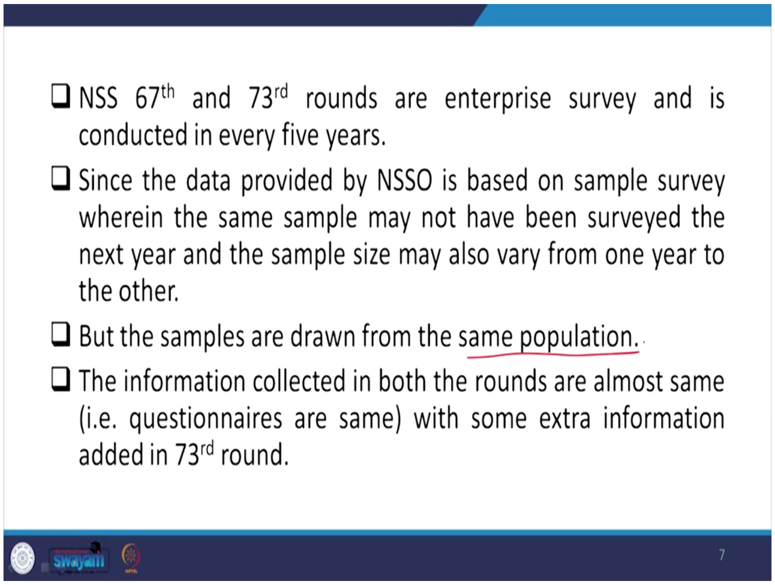
CONSTRUCTION OF POOLED CROSS SECTION DATA

- ❑ In India, **National Sample Survey** (NSS) data provides cross section data periodically on various socio economic heads.
- ❑ Here we are using NSS 67th (2010-11) and 73rd round (2015-16).
- ❑ NSS 67th and 73rd round are surveys on **unincorporated non-agricultural enterprises** that provides data on various indicators of economic and operational characteristics of such enterprises in manufacturing, trade and other services sector (excluding construction) at the national and state level.

6

Let us understand the construction of the pooled cross-sectional data. We are considering NSS data that provides cross-section data periodically. As we already discussed in NSS 67th. That was in 2010-11 and 73rd round of NSS of 2015-16 are going to be discussed now. In these two rounds, we are emphasizing on the unincorporated non-agricultural enterprises, which we already discussed in earlier modules as well. Especially we have the interest variables like enterprising types such as manufacturing, trade and other services. But make sure that these data do not consider the construction sector. So, coming to the survey, this is conducted in every 5 years.

(Refer Slide Time: 06:51)

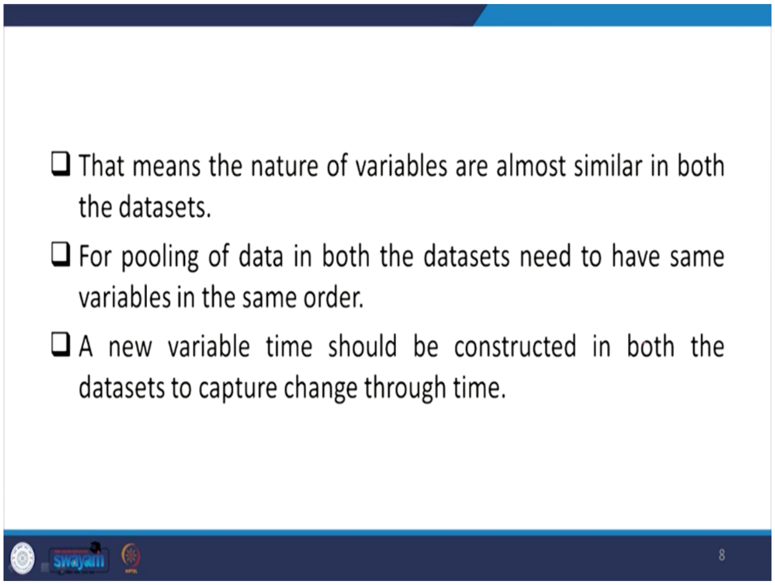


- ❑ NSS 67th and 73rd rounds are enterprise survey and is conducted in every five years.
- ❑ Since the data provided by NSSO is based on sample survey wherein the same sample may not have been surveyed the next year and the sample size may also vary from one year to the other.
- ❑ But the samples are drawn from the same population.
- ❑ The information collected in both the rounds are almost same (i.e. questionnaires are same) with some extra information added in 73rd round.

And since the data provided by NSS is based on sample survey where in the same sample may not have been surveyed the next year. And the sample size may also vary from one year to another year. But the samples are drawn from the same population. What do you mean by same population? We wanted to mention that the population base, the enterprising base of concern for the survey were taken from the same districts in India.

So, the base of the population is same. But the random selection of the samples is different. The information collected in both rounds are almost same, especially relating to the variables of conscience. And besides that the 73rd round added some extra information, some of them we discussed earlier in our earlier modules.

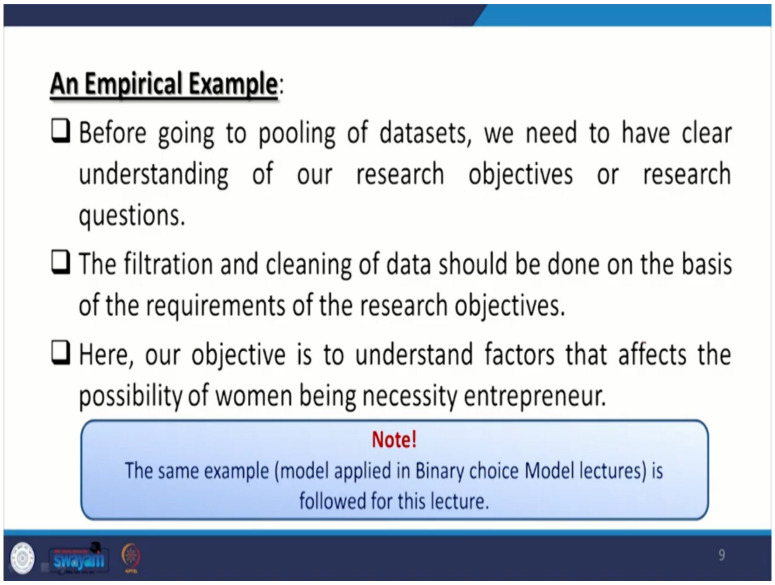
(Refer Slide Time: 07:56)



- ❑ That means the nature of variables are almost similar in both the datasets.
- ❑ For pooling of data in both the datasets need to have same variables in the same order.
- ❑ A new variable time should be constructed in both the datasets to capture change through time.

So, coming to the nature of variable, as we already discussed that they are similar in both the datasets. So, pooling of datasets specially, requires the same variables almost in same order. So, new variable created is time and we wanted to capture the time effect of other variables after pooling the data.

(Refer Slide Time: 08:18)



An Empirical Example:

- ❑ Before going to pooling of datasets, we need to have clear understanding of our research objectives or research questions.
- ❑ The filtration and cleaning of data should be done on the basis of the requirements of the research objectives.
- ❑ Here, our objective is to understand factors that affects the possibility of women being necessity entrepreneur.

Note!

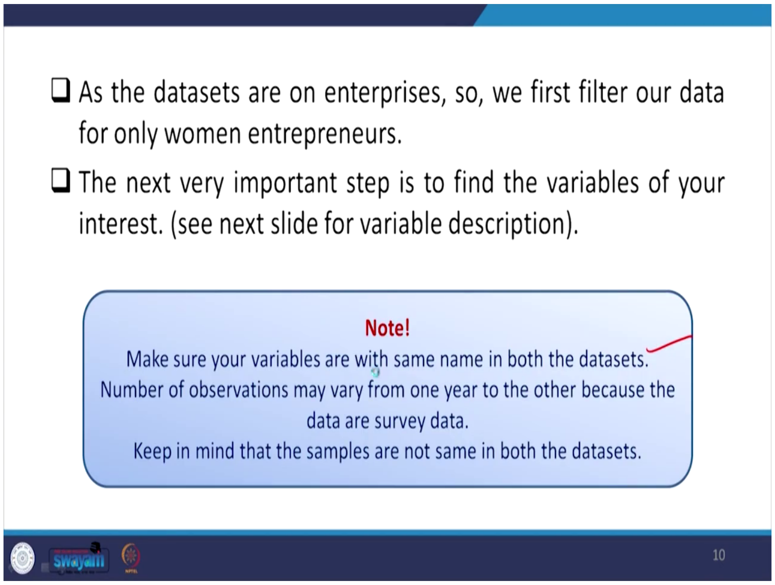
The same example (model applied in Binary choice Model lectures) is followed for this lecture.

Empirically, we wanted to explain this pooling of dataset using both the rounds of data. The objective of pooling would also be explained while explaining the pooling, the filtration and cleaning of data should also be done on the basis of the objectives we are targeting.

Here our objective is to understand the factors which affect the possibility of women being a necessity entrepreneur. So, we already clarified that whether necessity or opportunity-based entrepreneur, we are going to also suggest you one article which we published in the Standard ABDC ranked Journal, you can refer at the end. We are going to suggest that.

So, our interest here is to identify the important factors, alright the same example was also given during our lecture on Binary Choice model, especially on the week which we discuss about qualitative variables. Limited dependent variable models also. So, you please follow that and find out other clarification if you have desired to understand.

(Refer Slide Time: 09:46)



❑ As the datasets are on enterprises, so, we first filter our data for only women entrepreneurs.

❑ The next very important step is to find the variables of your interest. (see next slide for variable description).

Note!

Make sure your variables are with same name in both the datasets.

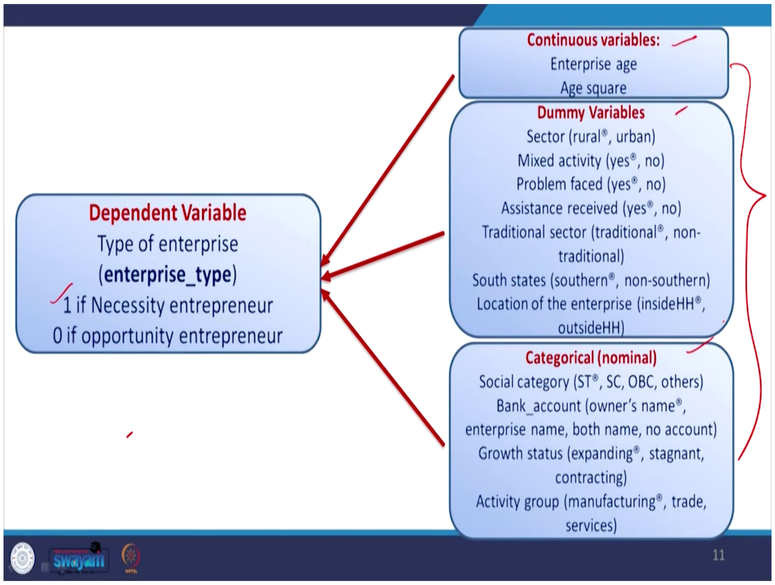
Number of observations may vary from one year to the other because the data are survey data.

Keep in mind that the samples are not same in both the datasets.

As the datasets are on enterprises, we first filter our data. For only women entrepreneur since our objective is to discuss on necessity women entrepreneur. So, we are emphasizing and filtering on that angle. The next important step is to get the variables of our interest which affect the women entrepreneur to be necessarily driven.

There is a note here, please mark it carefully, please make sure that the variables are with the same name in both the datasets. That is important variables should be with the same name, number or observation may vary from one year to another because the datasets are survey data. And keep in mind that the samples are not the same in both the data sets.

(Refer Slide Time: 10:40)



Dependent Variable

Type of enterprise

(**enterprise_type**)

1 if Necessity entrepreneur

0 if opportunity entrepreneur

Continuous variables:

Enterprise age

Age square

Dummy Variables

Sector (rural®, urban)

Mixed activity (yes®, no)

Problem faced (yes®, no)

Assistance received (yes®, no)

Traditional sector (traditional®, non-traditional)

South states (southern®, non-southern)

Location of the enterprise (insideHH®, outsideHH)

Categorical (nominal)

Social category (ST®, SC, OBC, others)

Bank_account (owner's name®, enterprise name, both name, no account)

Growth status (expanding®, stagnant, contracting)

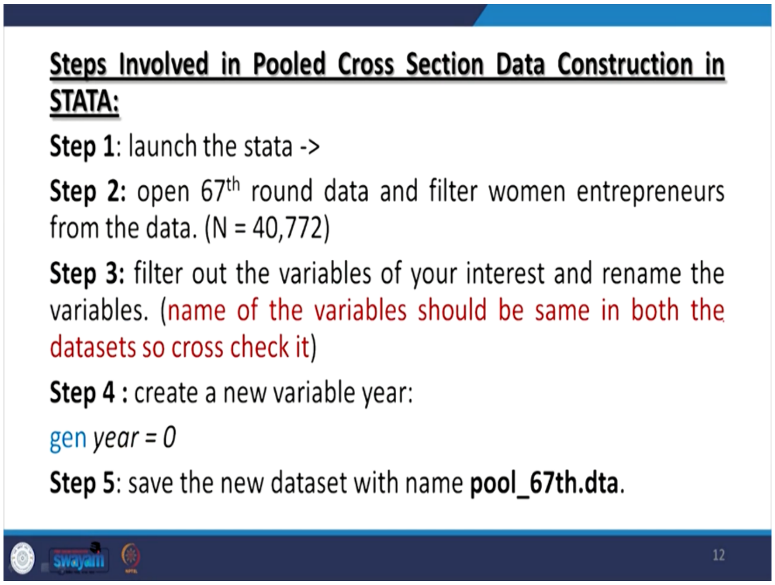
Activity group (manufacturing®, trade, services)

11

Since different periods are considered. So, samples need not be same but variables with the same name must carry. As we already mentioned, that our dependent variable is women entrepreneur. So, enterprise type is our name of the variable so one is the code if it is of necessity type and 0 if it is opportunity type. The other variable that we already given emphasis of this aspect earlier like some continuous variables.

We have dummy variables in the model, we have categorical variables, we have some random variables that we mentioned, like we discussed about Enterprise age. And we also take the age square of that variable then from the dummy like sector dummy, mixed activity dummy and so on. Some categorical variable like social categories, bank accounts, etcetera. You go through and you will certainly find out how these variables affect the enterprise type.

(Refer Slide Time: 11:49)



Steps Involved in Pooled Cross Section Data Construction in STATA:

Step 1: launch the stata ->

Step 2: open 67th round data and filter women entrepreneurs from the data. (N = 40,772)

Step 3: filter out the variables of your interest and rename the variables. (name of the variables should be same in both the datasets so cross check it)

Step 4 : create a new variable year:

gen *year = 0*

Step 5: save the new dataset with name **pool_67th.dta**.

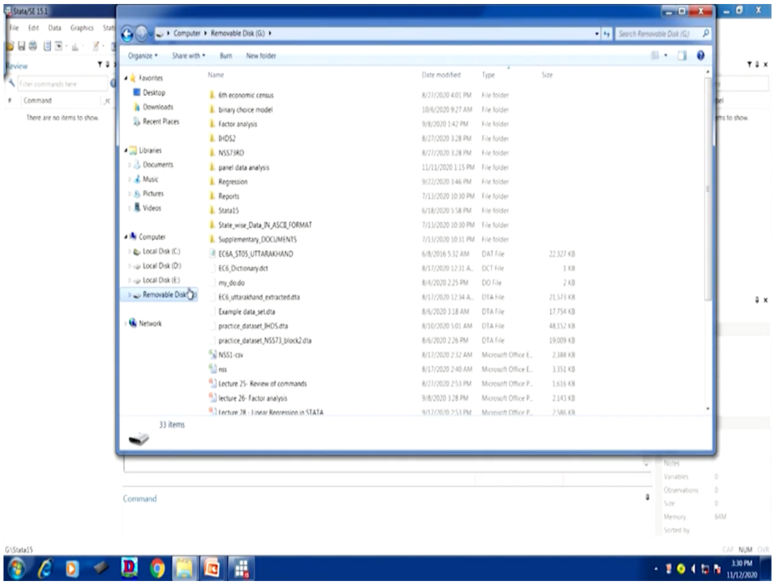
We from the pool we can cross check and verify. And coming to the necessary steps involved in the construction of pooled data and their explanation, in this slide, we are mentioning the most important five steps. Launching the data in STATA. We start with 67th round data and then you can do on your own. But I am just telling you these important step and whichever is required direct operation and explanation we will use the software.

You simply open the 67th round data which already been guided earlier. It contains 40772 observations, there are women entrepreneurs and now filter the variables of your interest. And rename them like the variable which we already listed you can also change on your own, but try to give the name same in both the rounds. And create a new variable.

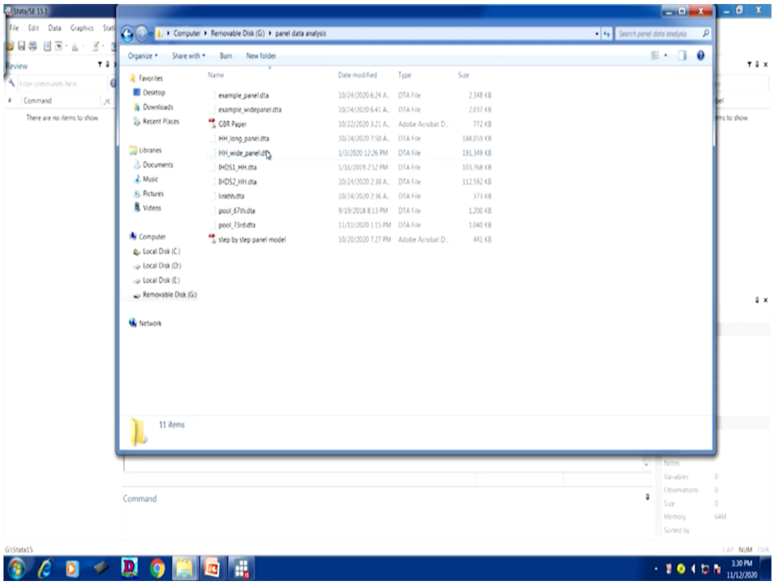
This variable is not there in the data we need to generate a variable that is here since two time period that is 67th and 73rd we are pooling, We are appending. So, only one variable is missing from the data. That is your time period. So, time we can give, it is equal to 0 generate a variable here equal to 0.

You can create it for the first period that is 67th round. We can save the dataset with the name pool 67th data will also show it and will also provide the sample of it for your easy understanding. I think that is there I am not going to open it but I will just show it to you, like here.

(Refer Slide Time: 13:50)



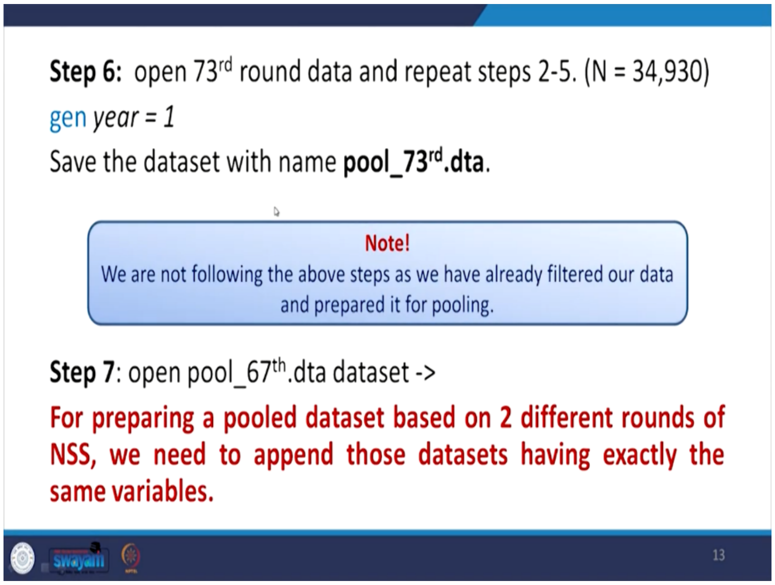
Name	Date modified	Type	Size
6th economic census	8/27/2020 4:01 PM	File folder	
binary choice model	10/6/2020 9:27 AM	File folder	
Factor analysis	9/8/2020 1:42 PM	File folder	
IHDS2	8/27/2020 3:28 PM	File folder	
NSS73RD	8/27/2020 3:28 PM	File folder	
panel data analysis	11/11/2020 1:15 PM	File folder	
Regression	9/22/2020 3:46 PM	File folder	
Reports	7/13/2020 10:30 PM	File folder	
Stata15	6/18/2020 5:58 PM	File folder	
State_wise_Data_IN_ASCII_FORMAT	7/13/2020 10:30 PM	File folder	
Supplementary_DOCUMENTS	7/13/2020 10:31 PM	File folder	
EC6A_ST05_UTTARAKHAND	6/8/2016 5:32 AM	DAT File	22,327 KB
EC6_Dictionary.dct	8/17/2020 12:31 A...	DCT File	1 KB
my_do.do	8/4/2020 2:25 PM	DO File	2 KB
EC6_uttarakhand_extracted.dta	8/17/2020 12:34 A...	DTA File	21,573 KB
Example data_set.dta	8/6/2020 3:18 AM	DTA File	17,754 KB
practice_dataset_IHDS.dta	8/10/2020 5:01 AM	DTA File	48,152 KB
practice_dataset_NSS73_block2.dta	8/6/2020 2:26 PM	DTA File	19,009 KB
NSS1-csv	8/17/2020 2:32 AM	Microsoft Office E...	2,388 KB
nss	8/17/2020 2:40 AM	Microsoft Office E...	3,351 KB
Lecture 25- Review of commands	8/27/2020 2:53 PM	Microsoft Office P...	1,616 KB
lecture 26- Factor analysis	9/8/2020 3:28 PM	Microsoft Office P...	2,143 KB
Lecture 28 - Linear Regression in STATA	9/17/2020 2:53 PM	Microsoft Office P...	2,586 KB



Name	Date modified	Type	Size
example_panel.dta	10/24/2020 6:24 A...	DTA File	2,348 KB
example_widepanel.dta	10/24/2020 6:41 A...	DTA File	2,037 KB
GBR Paper	10/22/2020 3:21 A...	Adobe Acrobat D...	772 KB
HH_long_panel.dta	10/24/2020 7:50 A...	DTA File	188,059 KB
HH_wide_panel.dta	1/3/2020 12:26 PM	DTA File	191,349 KB
IHDS1_HH.dta	1/16/2019 2:52 PM	DTA File	103,768 KB
IHDS2_HH.dta	10/24/2020 2:30 A...	DTA File	112,592 KB
linkhh.dta	10/24/2020 2:36 A...	DTA File	373 KB
pool_67th.dta	9/19/2018 8:13 PM	DTA File	1,200 KB
pool_73rd.dta	11/11/2020 1:15 PM	DTA File	1,040 KB
step by step panel model	10/20/2020 7:27 PM	Adobe Acrobat D...	441 KB

So, it is there now pool 67 will provide this, and we also saw it once. Do you pool 73rd. So, that is going to be explained to you shortly? I am just giving the clarification of the concepts.

(Refer Slide Time: 14:10)



Step 6: open 73rd round data and repeat steps 2-5. (N = 34,930)

gen *year = 1*

Save the dataset with name **pool_73rd.dta**.

Note!
We are not following the above steps as we have already filtered our data and prepared it for pooling.

Step 7: open pool_67th.dta dataset ->

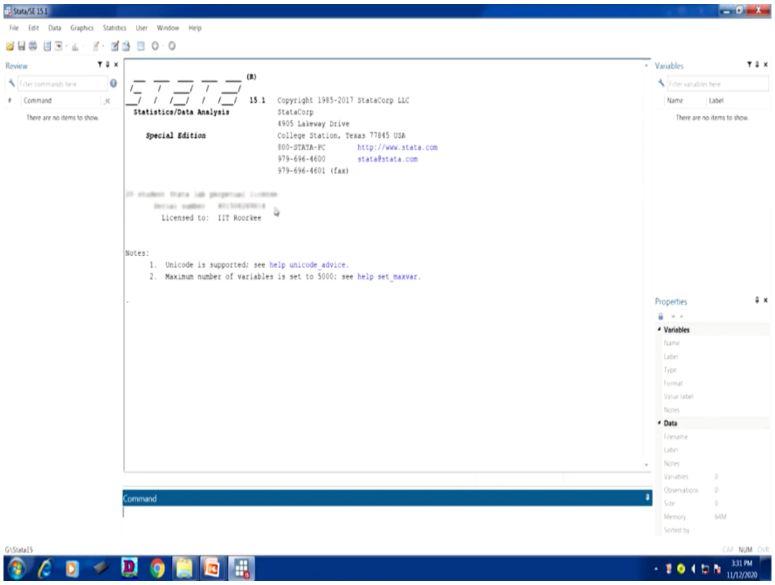
For preparing a pooled dataset based on 2 different rounds of NSS, we need to append those datasets having exactly the same variables.

13

Similarly, we open the another round data that is seven third and we again repeat all those five steps that have been already mentioned here now and then two to five steps. That is very important. And to mention here that after filtration the women enterprises here for us for the explanation is of 34930. And we again generate the same year as the variable name to be one, because this is another time period data. And will save with the pool 73rd data.

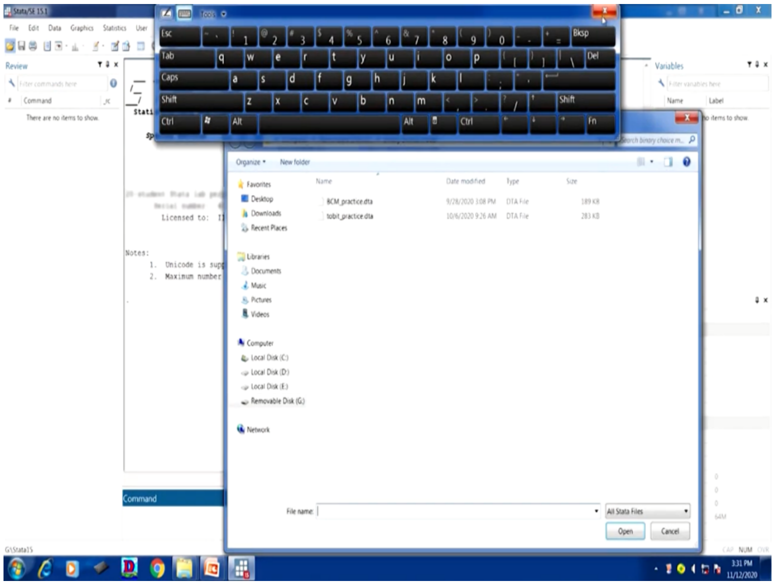
We are not following the above steps as we have already filtered out our data, the data which we have kept here for recognition, we already done all those steps. So, we are not repeating on each step that we have mentioned. Onwards we are going to experiment in front you like we are opening the data that is pooled data of 67th round. Let me open that first for you.

(Refer Slide Time: 15:24)



```
  ____  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

[redacted]
       Licensed to:  IIT Roorkee


Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
```

Stata/SE 15.1
File Edit Data Graphics Statistics User Window Help
Review
Command
There are no items to show.
Variables
Name Label
There are no items to show.
Properties
Variables
Data
Command
G:\Stata15

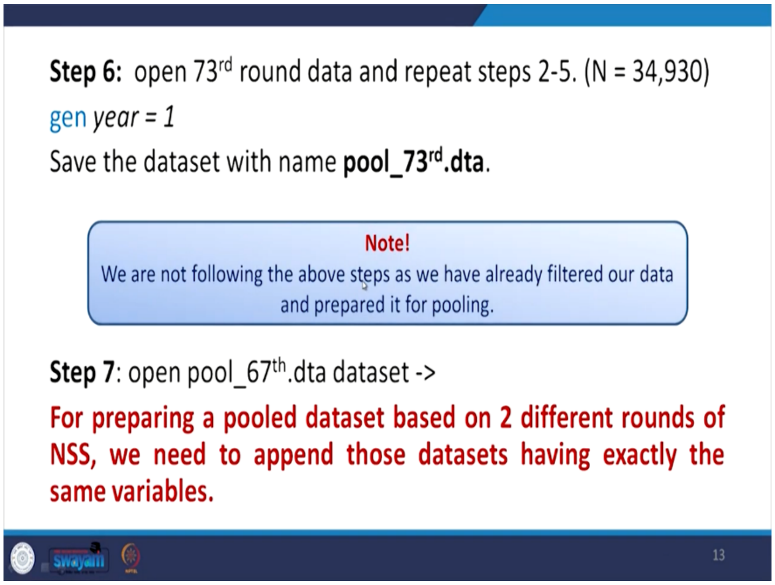


Name	Date modified	Type	Size
BCM_practice.dta	9/28/2020 3:08 PM	DTA File	189 KB
tobit_practice.dta	10/6/2020 9:26 AM	DTA File	283 KB

File name:

It is here now I am just going to open it, it is here. So, the pooled 67th data, as I told you it has already been opened now. what I will do.

(Refer Slide Time: 15:43)



Step 6: open 73rd round data and repeat steps 2-5. (N = 34,930)

gen *year = 1*

Save the dataset with name **pool_73rd.dta**.

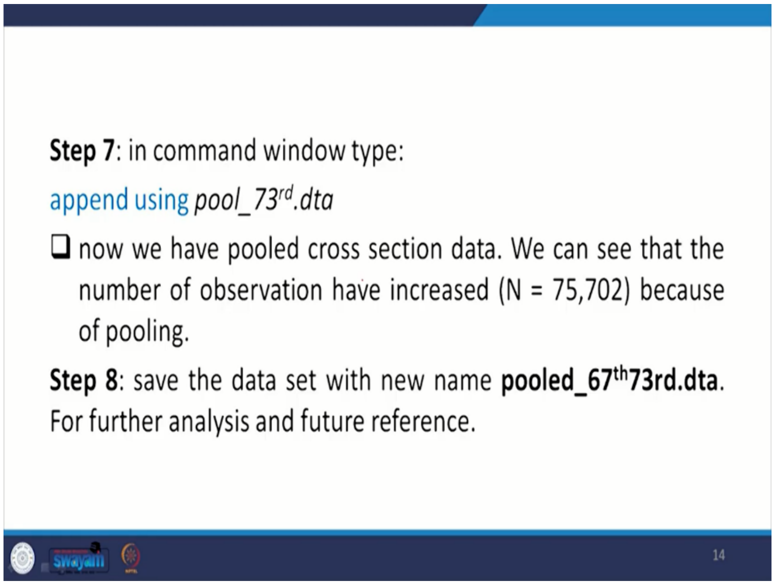
Note!
We are not following the above steps as we have already filtered our data and prepared it for pooling.

Step 7: open pool_67th.dta dataset ->

For preparing a pooled dataset based on 2 different rounds of NSS, we need to append those datasets having exactly the same variables.

I will go by, just read the red colour for the text. That is for preparing a pooled dataset based on two different rounds of NSS. We need to append those datasets like two data sets. We are discussing 67th 73rd we need to append, having the exactly same variables then only it will simply append and basically it is a vertical appending procedure. The observation of 67th will be from the beginning and at the end part the 73rd observation will be simply appended or added vertically. Append is the right word add sometimes you may infer that it is a numerical operation, but we are simply trying to say that it gets kept in different rules of the data.

(Refer Slide Time: 16:33)



Step 7: in command window type:

append using *pool_73rd.dta*

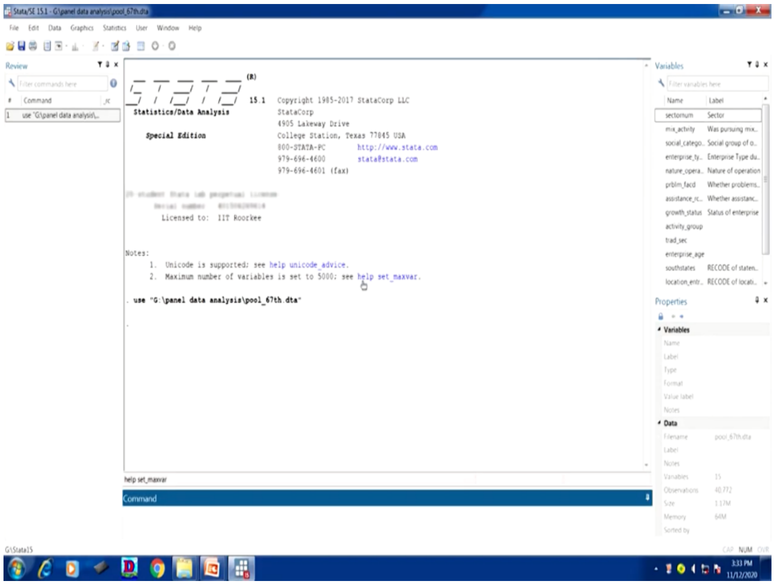
□ now we have pooled cross section data. We can see that the number of observation have increased (N = 75,702) because of pooling.

Step 8: save the data set with new name **pooled_67th73rd.dta**. For further analysis and future reference.

14

Also everything right now. So, what we are doing, we are trying to show you the append.

(Refer Slide Time: 16:41)



#	Command	_rc
1	use "G:\panel data analysis\...	

```
  ____  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

[redacted]
         Licensed to:  IIT Roorkee


Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

. use "G:\panel data analysis\pool_67th.dta"

.
```

```
help set_maxvar
```

Name	Label
sectornum	Sector
mix_activity	Was pursuing mix...
social_catego...	Social group of o...
enterprise_ty...	Enterprise Type du...
nature_opera...	Nature of operation
prblm_facd	Whether problems...
assistance_rc...	Whether assistanc...
growth_status	Status of enterprise
activity_group	
trad_sec	
enterprise_age	
southstates	RECODE of staten...
location_entr...	RECODE of locati...

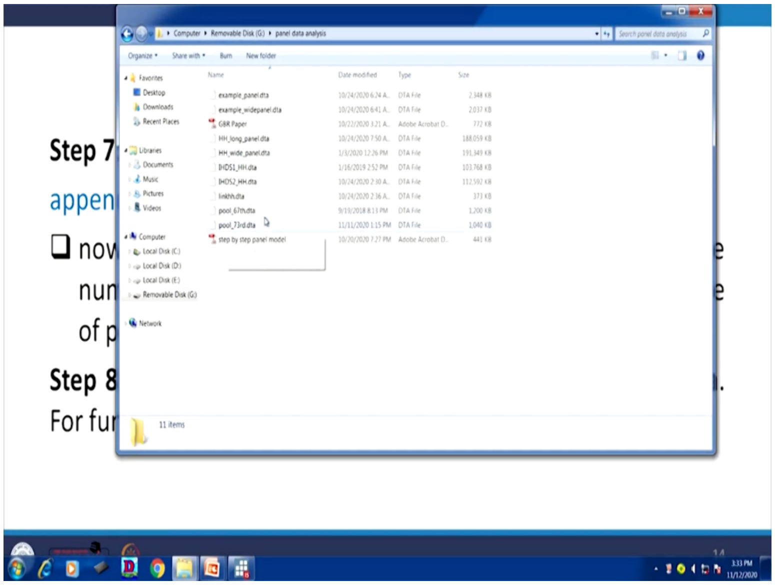
Variables
Name
Label
Type
Format
Value label
Notes

Data

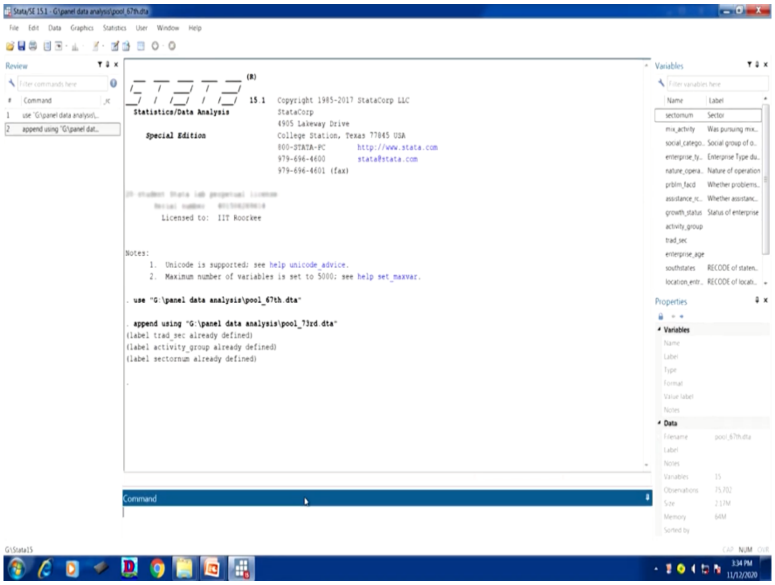
Filename	pool_67th.dta
Label	
Notes	
Variables	15
Observations	40,772
Size	1.17M
Memory	64M
Sorted by	

Since we have already used the data 67th to open. So, you are opening append using 73rd data. The location of it is here.

(Refer Slide Time: 17:02)



Name	Date modified	Type	Size
example_panel.dta	10/24/2020 6:24 A...	DTA File	2,348 KB
example_widepanel.dta	10/24/2020 6:41 A...	DTA File	2,037 KB
GBR Paper	10/22/2020 3:21 A...	Adobe Acrobat D...	772 KB
HH_long_panel.dta	10/24/2020 7:50 A...	DTA File	188,059 KB
HH_wide_panel.dta	1/3/2020 12:26 PM	DTA File	191,349 KB
IHDS1_HH.dta	1/16/2019 2:52 PM	DTA File	103,768 KB
IHDS2_HH.dta	10/24/2020 2:30 A...	DTA File	112,592 KB
linkhh.dta	10/24/2020 2:36 A...	DTA File	373 KB
pool_67th.dta	9/19/2018 8:13 PM	DTA File	1,200 KB
pool_73rd.dta	11/11/2020 1:15 PM	DTA File	1,040 KB
step by step panel model	10/20/2020 7:27 PM	Adobe Acrobat D...	441 KB



#	Command	_rc
1	use "G:\panel data analysis\...	
2	append using "G:\panel dat...	

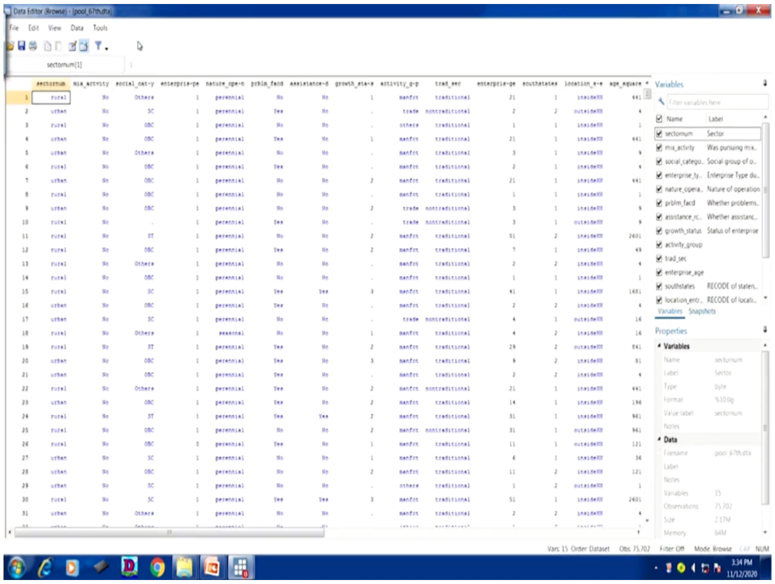
```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

[redacted]
         Licensed to:  IIT Roorkee


Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

. use "G:\panel data analysis\pool_67th.dta"

. append using "G:\panel data analysis\pool_73rd.dta"
(label trad_sec already defined)
(label activity_group already defined)
(label sectornum already defined)

.
```

Name	Label
sectornum	Sector
mix_actvity	Was pursuing mix...
social_catego...	Social group of o...
enterprise_ty...	Enterprise Type du...
nature_opera...	Nature of operation
prblm_facd	Whether problems...
assistance_rc...	Whether assistanc...
growth_status	Status of enterprise
activity_group	
trad_sec	
enterprise_age	
southstates	RECODE of staten...
location_entr...	RECODE of locati...

Variables
Name
Label
Type
Format
Value label
Notes

Data

Filename	pool_67th.dta
Label	
Notes	
Variables	15
Observations	75,702
Size	2.17M
Memory	64M
Sorted by	



Data Editor (Browse) - [pool_67th.dta]

sectornum[1]

	sectornum	mix_actvity	social_cat~y	enterpris~pe	nature_ope~n	prblm_facd	assistance~d	growth_sta~s	activity_g~p	trad_sec	enterpris~ge	southstates	location_e~e	age_square
1	rural	No	Others	1	perennial	No	No	1	manfct	traditional	21	1	insideHH	441
2	urban	No	SC	1	perennial	Yes	No	.	trade	nontraditional	2	2	outsideHH	4
3	rural	No	OBC	1	perennial	No	No	.	others	traditional	1	1	insideHH	1
4	urban	No	OBC	1	perennial	Yes	No	1	manfct	traditional	21	1	insideHH	441
5	urban	No	Others	1	perennial	No	No	.	manfct	traditional	3	1	insideHH	9
6	rural	No	OBC	1	perennial	Yes	No	.	manfct	traditional	2	1	insideHH	4
7	urban	No	OBC	1	perennial	No	No	2	manfct	traditional	21	1	insideHH	441
8	rural	No	OBC	1	perennial	No	No	.	manfct	traditional	1	1	insideHH	1
9	urban	No	OBC	1	perennial	No	No	2	trade	nontraditional	3	1	insideHH	9
10	rural	No	.	1	perennial	Yes	No	.	trade	nontraditional	3	1	outsideHH	9
11	rural	No	ST	1	perennial	No	No	2	manfct	traditional	51	2	insideHH	2601
12	rural	No	OBC	1	perennial	Yes	No	2	manfct	traditional	7	1	insideHH	49
13	rural	No	Others	1	perennial	No	No	.	manfct	traditional	2	2	insideHH	4
14	rural	No	OBC	1	perennial	No	No	.	manfct	traditional	1	1	insideHH	1
15	rural	No	SC	1	perennial	Yes	Yes	3	manfct	traditional	41	1	insideHH	1681
16	urban	No	OBC	1	perennial	Yes	No	.	manfct	traditional	2	2	insideHH	4
17	urban	No	SC	1	perennial	No	No	.	trade	nontraditional	4	1	outsideHH	16
18	rural	No	Others	1	seasonal	No	No	1	manfct	traditional	4	2	insideHH	16
19	rural	No	ST	1	perennial	Yes	No	2	manfct	traditional	29	2	outsideHH	841
20	urban	No	OBC	1	perennial	Yes	No	3	manfct	traditional	9	2	insideHH	81
21	urban	No	OBC	1	perennial	Yes	No	.	manfct	traditional	2	2	insideHH	4
22	rural	No	Others	1	perennial	Yes	No	2	manfct	nontraditional	21	1	insideHH	441
23	urban	No	OBC	1	perennial	Yes	No	2	manfct	traditional	14	1	insideHH	196
24	rural	No	ST	1	perennial	Yes	Yes	2	manfct	traditional	31	1	insideHH	961
25	rural	No	OBC	1	perennial	No	No	2	manfct	nontraditional	31	1	outsideHH	961
26	rural	No	OBC	0	perennial	Yes	No	1	manfct	traditional	11	1	outsideHH	121
27	urban	No	SC	1	perennial	No	No	1	manfct	traditional	6	1	insideHH	36
28	urban	No	OBC	1	perennial	No	No	2	manfct	traditional	11	2	insideHH	121
29	urban	No	SC	1	perennial	No	No	.	others	traditional	1	2	outsideHH	1
30	rural	No	SC	1	perennial	Yes	Yes	3	manfct	traditional	51	1	insideHH	2601
31	urban	No	Others	1	perennial	No	No	.	manfct	traditional	2	2	insideHH	4

Name	Label
sectornum	Sector
mix_activity	Was pursuing mix...
social_catego...	Social group of o...
enterprise_ty...	Enterprise Type du...
nature_opera...	Nature of operation
prblm_facd	Whether problems...
assistance_rc...	Whether assistanc...
growth_status	Status of enterprise
activity_group	
trad_sec	
enterprise_age	
southstates	RECODE of staten...
location_entr...	RECODE of locati...

Properties

Variables

Name	sectornum
Label	Sector
Type	byte
Format	%10.0g
Value label	sectornum
Notes	

Data

Filename	pool_67th.dta
Label	
Notes	
Variables	15
Observations	75,702
Size	2.17M
Memory	64M

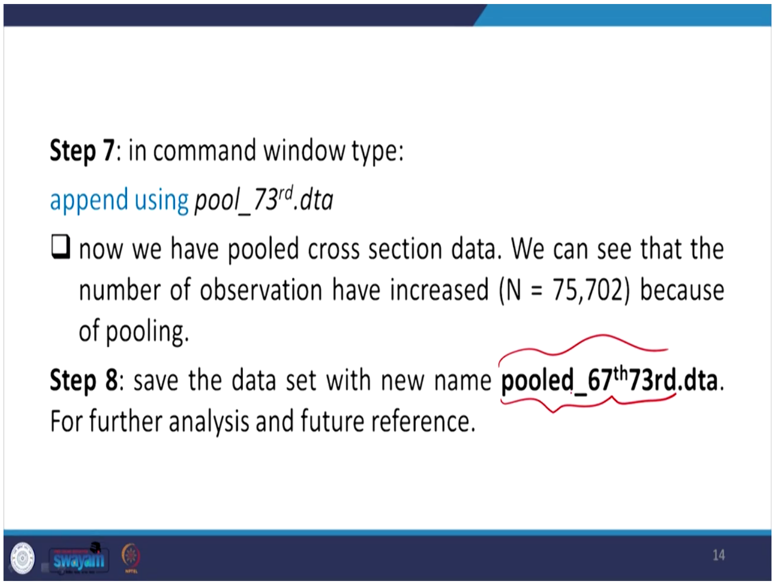
Vars: 15 Order: Dataset Obs: 75,702 Filter: Off Mode: Browse CAP NUM

We are copying the location of it. 67 already opened, we are simply we are adding the 73rd round. 73rd so it is here. Alright. We have appended the data. You can check the number of the observation recently the case has been added, increased to 75702. And the variables are same that those are 15 we have carried from the 67th round and since the number of variables are of 15 in both the rounds. So, number of variables remain the same.

And our observation increased by another the number of observations in the next round. So, these are the data we have added. And we have appended now just to show you that your variable have already created and we have deliberately made it a dummy one to identify that zero indicates your 67th round data and whenever one is there, one indicates you a 73rd round data.

So, suppose I just wanted to mix like age of the enterprise times year effect of it, or location of the enterprise. And year effect we can simply multiply this data. I can find out the location with effect on the dependent variable so there is a better interpretation possible in our ppt.

(Refer Slide Time: 19:04)



Step 7: in command window type:

append using *pool_73rd.dta*

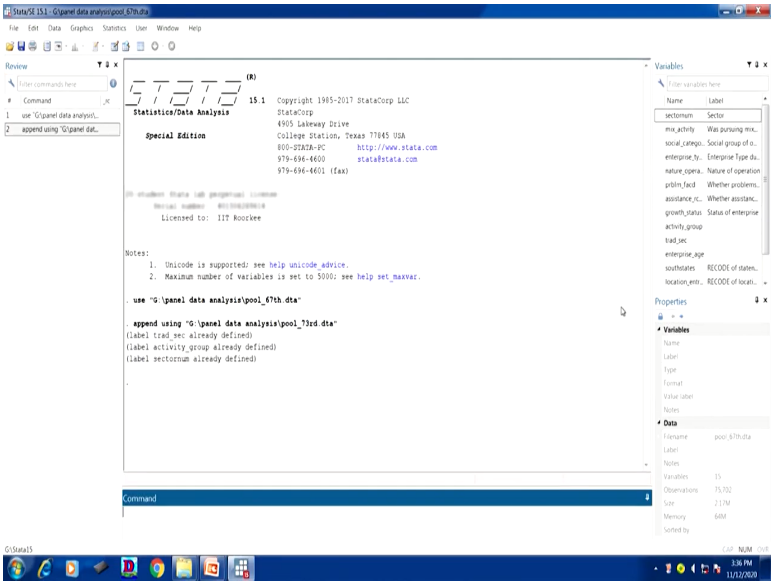
- ❑ now we have pooled cross section data. We can see that the number of observation have increased (N = 75,702) because of pooling.

Step 8: save the data set with new name **pooled_67th73rd.dta**. For further analysis and future reference.

Let me go through the exact steps that we have already followed. After pooling one thing for sure, we have to save it, save the data with a convenient name that can identify pooled 67th as well as 73rd dot dta will come by default since dta file has already been opened. So, you please take a name with a convenient identification 67 and 73 both are shown here. You can take any other name, but for your clarification, we are just trying to keep some suggestions.

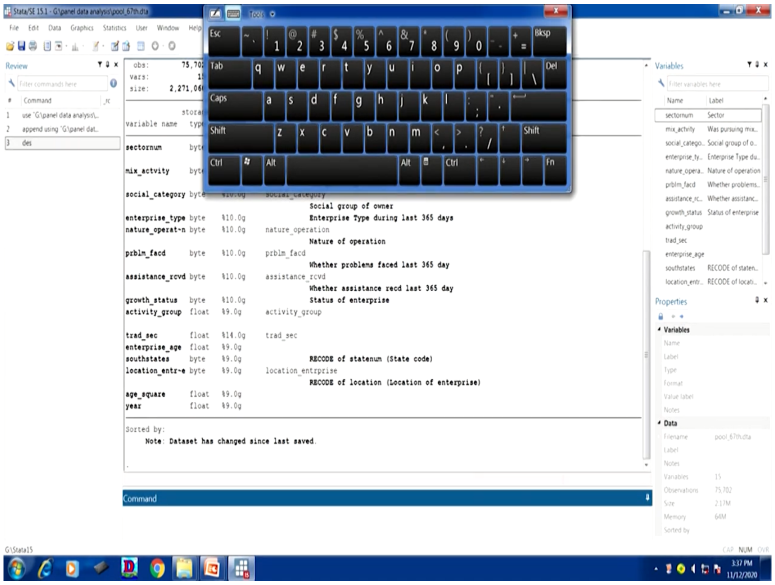
So, as we mentioned the number of observation has been increased already after pooling, alright. So, analysis can be done, we can do n number of explanation but some just sample we have done those explanations many times likewise we do in the earlier statistical operations.

(Refer Slide Time: 20:05)



Stata/SE 15.1 - G:\panel data analysis\pool_67th.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
use 'G:\panel data analysis\...
append using 'G:\panel dat...
Variables
Properties
Command

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

[redacted]
       Licensed to:  IIT Roorkee


Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.

. use "G:\panel data analysis\pool_67th.dta"

. append using "G:\panel data analysis\pool_73rd.dta"
(label trad_sec already defined)
(label activity_group already defined)
(label sectornum already defined)

.
```

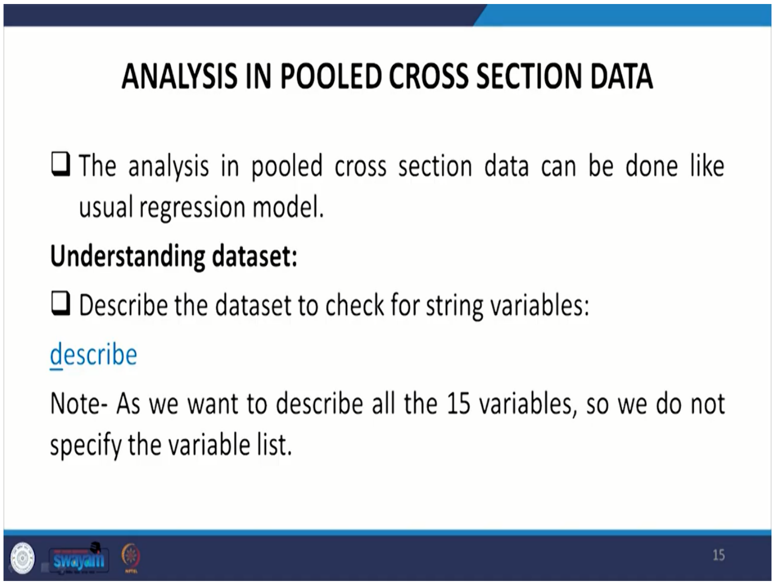


Stata/SE 15.1 - G:\panel data analysis\pool_67th.dta
File Edit Data Graphics Statistics User Window Help
Tools
Review
Command
use 'G:\panel data analysis\...
append using 'G:\panel dat...
des
Variables
Properties
Command

```
  obs:          75,70
 vars:             1
 size:     2,271,06

              storag
variable name   type

sectornum       byte

mix_actvity     byte

social_category byte

                                          Social group of owner
enterprise_type byte    %10.0g            Enterprise Type during last 365 days
nature_operat~n byte    %10.0g     nature_operation
                                          Nature of operation
prblm_facd      byte    %10.0g     prblm_facd
                                          Whether problems faced last 365 day
assistance_rcvd byte    %10.0g     assistance_rcvd
                                          Whether assistance recd last 365 day
growth_status   byte    %10.0g            Status of enterprise
activity_group  float   %9.0g      activity_group

trad_sec        float   %14.0g     trad_sec
enterprise_age  float   %9.0g
southstates     byte    %9.0g             RECODE of statenum (State code)
location_entr~e byte    %9.0g      location_entrprise
                                          RECODE of location (Location of enterprise)

age_square      float   %9.0g
year            float   %9.0g

Sorted by:
     Note: Dataset has changed since last saved.

.
```

We can simply do the same thing we can do describe. Simply describe, can clarify. It gives the information about the kind of storage the byte or float or numeric or string. And also the value label as well as variable label can also be clarified.

(Refer Slide Time: 20:32)



ANALYSIS IN POOLED CROSS SECTION DATA

❑ The analysis in pooled cross section data can be done like usual regression model.

Understanding dataset:

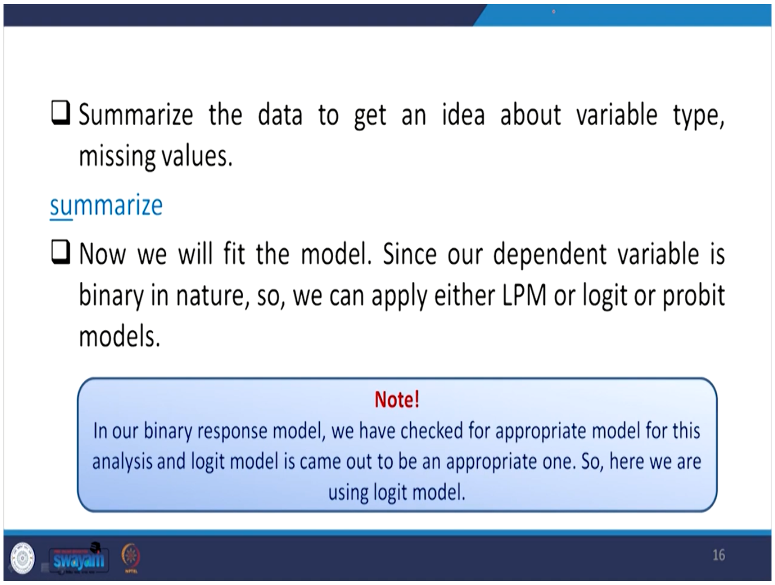
❑ Describe the dataset to check for string variables:

describe

Note- As we want to describe all the 15 variables, so we do not specify the variable list.

Similarly, the 15 variables, we have already appended. So, the 15 variable information will come.

(Refer Slide Time: 20:42)



❑ Summarize the data to get an idea about variable type, missing values.

summarize

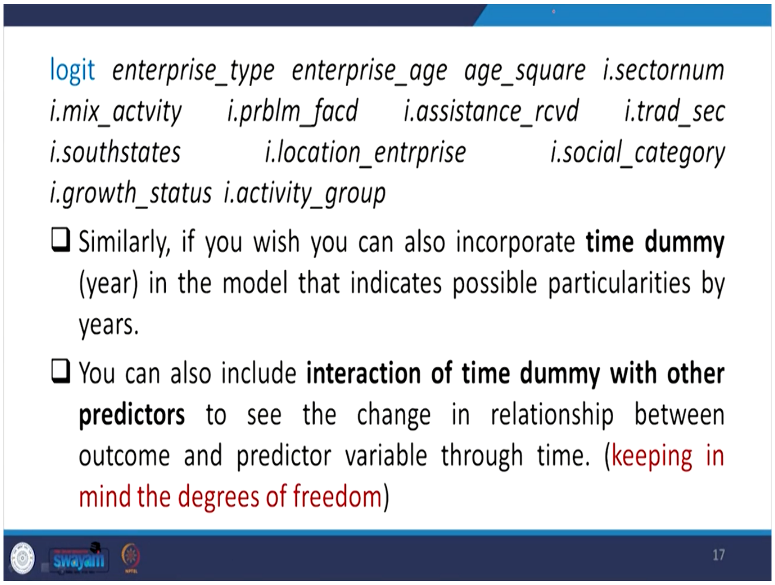
❑ Now we will fit the model. Since our dependent variable is binary in nature, so, we can apply either LPM or logit or probit models.

Note!
In our binary response model, we have checked for appropriate model for this analysis and logit model is came out to be an appropriate one. So, here we are using logit model.

Similarly, you can do summarize, you can just go through on your own. But I am not operating here much. But I think it is too simple to understand, just to mention that since our dependent variable is binary in nature, because we already pointed out that it is whether the enterprise is of necessity type or opportunity type. So, we can apply either of the 3 models we discuss in the binary choice model LPM model or logit or profit model.

So, you go to that section once again and verify and I am sure you will enjoy going through the data. In our binary response model, we have already check the appropriate model and it will be very useful for you. The same model we can apply to do it here.

(Refer Slide Time: 21:37)

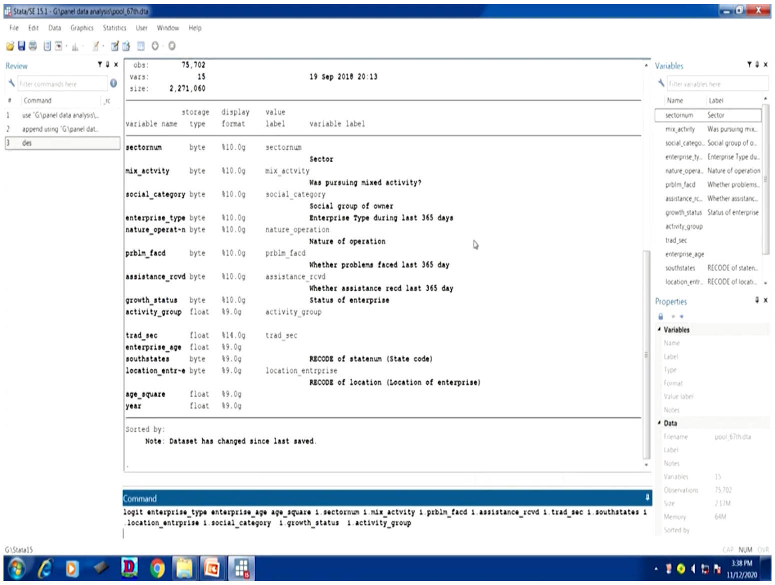


logit *enterprise_type enterprise_age age_square i.sectornum i.mix_actvity i.prblm_facd i.assistance_rcvd i.trad_sec i.southstates i.location_entrprise i.social_category i.growth_status i.activity_group*

- ❑ Similarly, if you wish you can also incorporate **time dummy** (year) in the model that indicates possible particularities by years.
- ❑ You can also include **interaction of time dummy with other predictors** to see the change in relationship between outcome and predictor variable through time. (keeping in mind the degrees of freedom)

We can operate it here just for your another round of confidence. Let me do it. Let me just do it once to that particular model. And just for your confidence, only though usually we suggested logit or profit based on some logic and we will once again operate here.

(Refer Slide Time: 22:00)



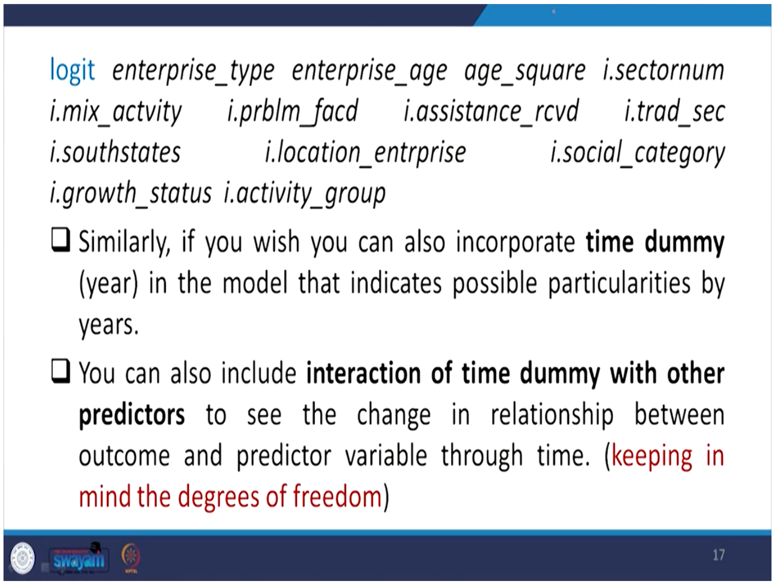
```
  obs:        75,702
 vars:            15                          19 Sep 2018 20:13
 size:     2,271,060
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
sectornum       byte    %10.0g     sectornum
                                              Sector
mix_actvity     byte    %10.0g     mix_actvity
                                              Was pursuing mixed activity?
social_category byte    %10.0g     social_category
                                              Social group of owner
enterprise_type byte    %10.0g                Enterprise Type during last 365 days
nature_operat~n byte    %10.0g     nature_operation
                                              Nature of operation
prblm_facd      byte    %10.0g     prblm_facd
                                              Whether problems faced last 365 day
assistance_rcvd byte    %10.0g     assistance_rcvd
                                              Whether assistance recd last 365 day
growth_status   byte    %10.0g                Status of enterprise
activity_group  float   %9.0g      activity_group

trad_sec        float   %14.0g     trad_sec
enterprise_age  float   %9.0g
southstates     byte    %9.0g                 RECODE of statenum (State code)
location_entr~e byte    %9.0g      location_entrprise
                                              RECODE of location (Location of enterprise)

age_square      float   %9.0g
year            float   %9.0g
--------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

Command

```
logit enterprise_type enterprise_age age_square i.sectornum i.mix_actvity i.prblm_facd i.assistance_rcvd i.trad_sec i.southstates i
.location_entrprise i.social_category  i.growth_status  i.activity_group
```

With the same data and the result is in front of your screen. It is I think walking now the result has come. Alright. So accordingly, different variables and the interpretation everything is there and our dependent variable is enterprise type. As I told you age and by square of that age whether it is significant or not. And about pseudo r square is not that important. We already discussed during our lecture you go through and surely you can able to interpret it. So, let me proceed further.

(Refer Slide Time: 22:43)

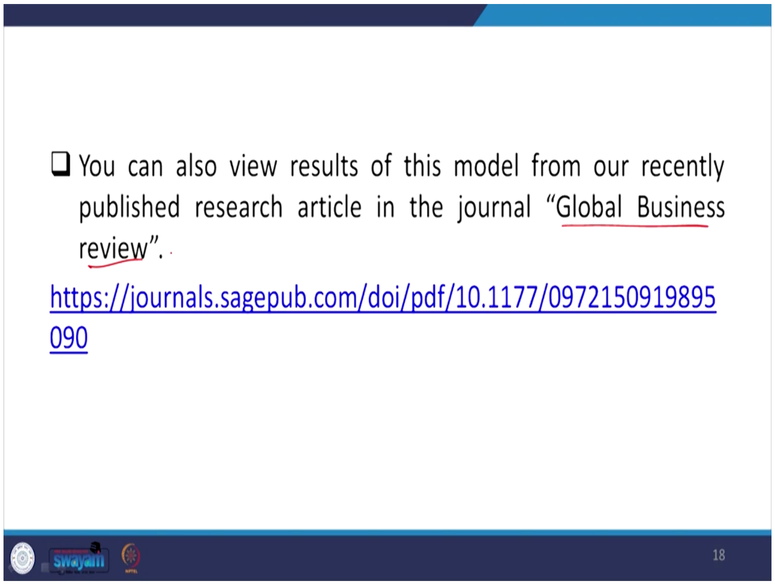


logit *enterprise_type enterprise_age age_square i.sectornum i.mix_actvity i.prblm_facd i.assistance_rcvd i.trad_sec i.southstates i.location_entrprise i.social_category i.growth_status i.activity_group*

- ❑ Similarly, if you wish you can also incorporate **time dummy** (year) in the model that indicates possible particularities by years.
- ❑ You can also include **interaction of time dummy with other predictors** to see the change in relationship between outcome and predictor variable through time. (keeping in mind the degrees of freedom)

Coming to that time dummy we have already discussed that. We can have interaction dummy also. We wanted to get the change in relationship between outcome and predicted variable through time. Basically, wherever you multiply the time effect, it will give important information. But since dummy has already been introduced, so the degrees of freedom gets reduced in the model. So, we need to be careful about that while analysing the dummies.

(Refer Slide Time: 23:24)



❑ You can also view results of this model from our recently published research article in the journal "Global Business review".

https://journals.sagepub.com/doi/pdf/10.1177/0972150919895090

18

Most importantly, if you wanted to understand the model very correctly and the interpretation, everything is there in this link. This is our article published in Global Business Review in the year 2020. So, you download and read. I am sure you can find better interpretation and a systematic interpretation of it with this. I thank you all, will now look forward to your involvement for the next lecture because next lecture we are going to give you a systematic guidance on fixed effect, panel effect discussions. Thank you so much.