Exploring Survey Data on Health Care

Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 35
Construction of Pseudo Panel

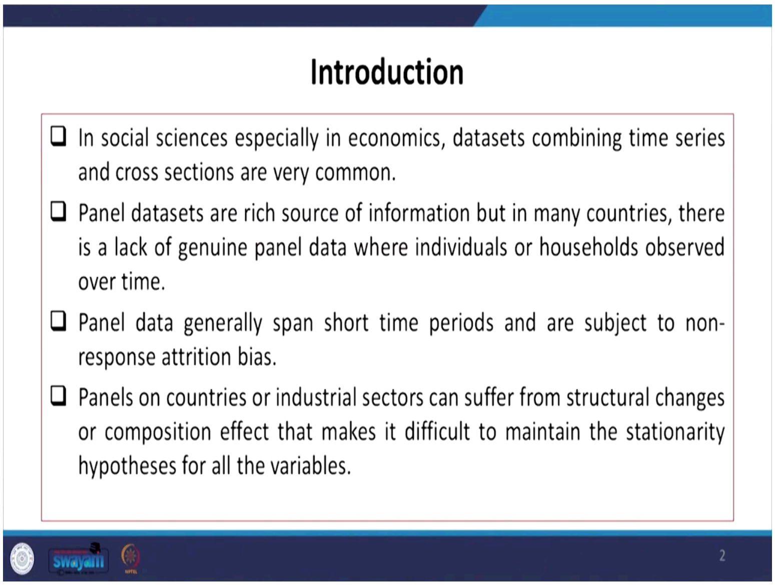
Welcome participants, once again to my NPTEL MOOC module on panel survey data on healthcare. We are on the 7th week of understanding panel survey data and the title of the NPTEL MOOC program of my course is exploring healthcare survey data.

The title of this particular week is panel survey data and here we are explaining how to have pseudo panel data. This is quite rare in the existing studies and very less number of studies which could have guided you about the pseudo panel and its construction.

Therefore, we kept this as a specific lecture. In giving you a certain direction about understanding the construction of pseudo panel data. So many large scales cross-sectional data sets could be actually presented with a panel structure.

In social sciences especially in economics, datasets combining time-series and cross-sections are very common. Panel data sets are rich source of information, but in many countries, there is a lack of genuine panel data where individuals or households observe over time.

(Refer Slide Time: 01:56)



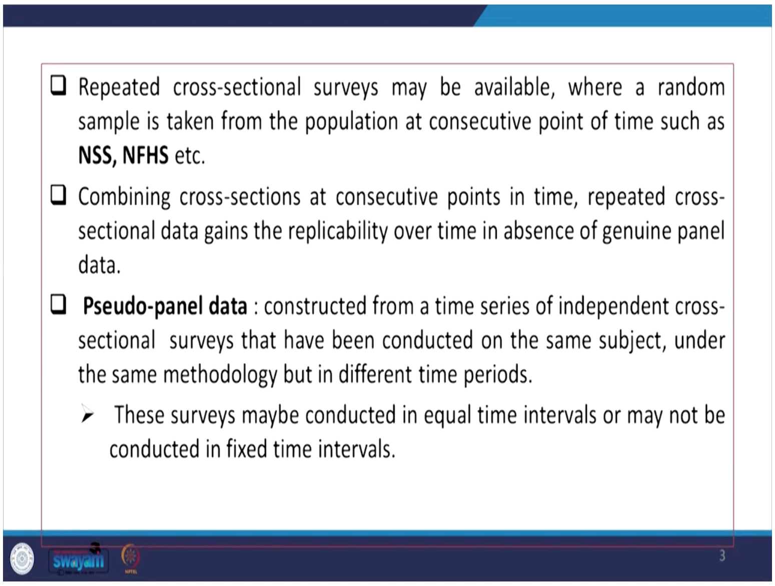
Introduction

- ❑ In social sciences especially in economics, datasets combining time series and cross sections are very common.
- ❑ Panel datasets are rich source of information but in many countries, there is a lack of genuine panel data where individuals or households observed over time.
- ❑ Panel data generally span short time periods and are subject to non-response attrition bias.
- ❑ Panels on countries or industrial sectors can suffer from structural changes or composition effect that makes it difficult to maintain the stationarity hypotheses for all the variables.

Panel data generally span a short period of time and are subject to non-response attrition bias.

Panels on countries or industrial sectors can suffer from structural changes or composition effect that makes it very difficult to maintain the stationarity hypotheses for all the variables that are carried forward in the next round.

(Refer Slide Time: 02:19)



- ❑ Repeated cross-sectional surveys may be available, where a random sample is taken from the population at consecutive point of time such as **NSS, NFHS** etc.
- ❑ Combining cross-sections at consecutive points in time, repeated cross-sectional data gains the replicability over time in absence of genuine panel data.
- ❑ **Pseudo-panel data** : constructed from a time series of independent cross-sectional surveys that have been conducted on the same subject, under the same methodology but in different time periods.
  - ➢ These surveys maybe conducted in equal time intervals or may not be conducted in fixed time intervals.

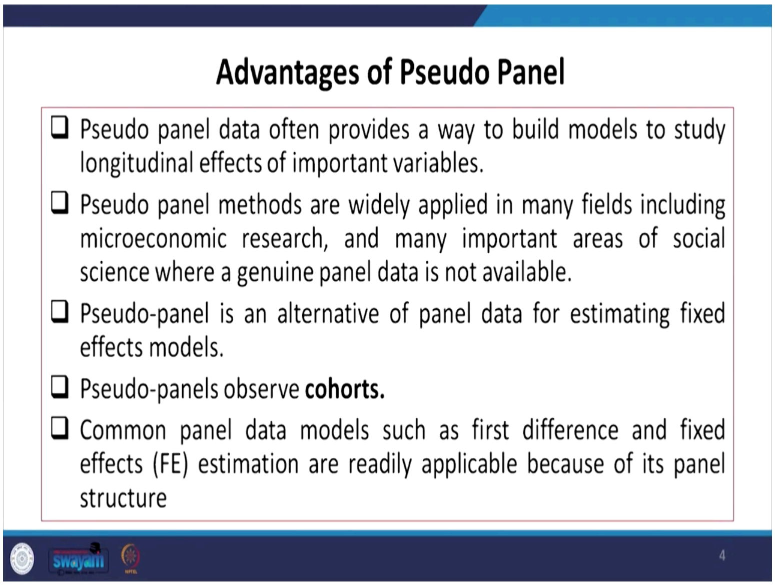
Repeated cross-sectional surveys may be available, where a random sample is taken from the population at a consecutive point of time such as NSS or NFHS etc.

Combining cross-sections at consecutive points in time repeated, cross-sectional data gains replicability over time in the absence of genuine panel data. Even if there is general panel data, repeated studies are taken in different time periods. Therefore, a pseudo panel can be constructed.

So, the pseudo panel is something where a time series of independent cross-sectional surveys that have been conducted on the same subject under the same methodology, but in different time periods could be conceptualized in a dataset called pseudo panel.

These surveys may be conducted in equal time intervals or may not be conducted in the fixed time interval but a time component is required.

(Refer Slide Time: 03:25)



Advantages of Pseudo Panel

□ Pseudo panel data often provides a way to build models to study longitudinal effects of important variables.

□ Pseudo panel methods are widely applied in many fields including microeconomic research, and many important areas of social science where a genuine panel data is not available.

□ Pseudo-panel is an alternative of panel data for estimating fixed effects models.

□ Pseudo-panels observe **cohorts.**

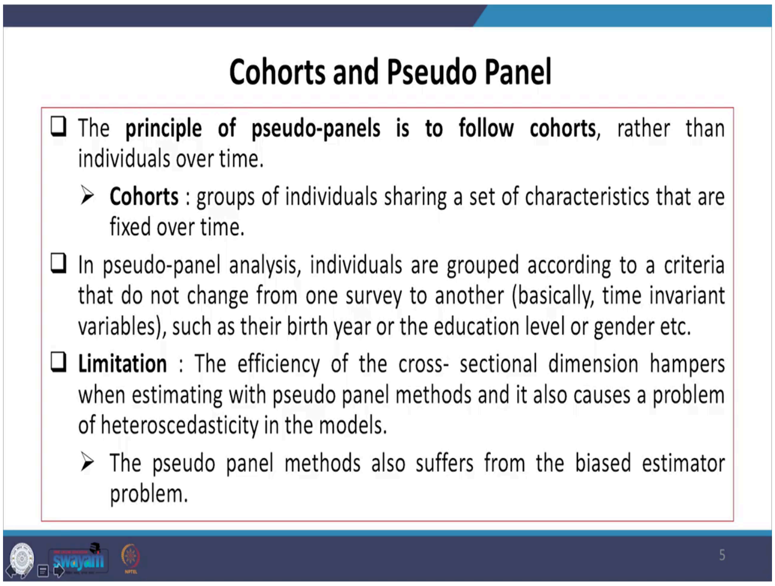
□ Common panel data models such as first difference and fixed effects (FE) estimation are readily applicable because of its panel structure

So, the advantage of the pseudo panel is that pseudo panel data often provides a way to build models to study the longitudinal effects of important variables. So, longitudinal components of important variables are continued in this setup.

Pseudo panel methods are widely applied in many fields including microeconomic research and many important areas of social sciences where genuine panel data is not actually available. The pseudo panel is an alternative to panel data for estimating fixed-effect models.

Pseudo panels observe cohorts instead of individual observations. Common panel data models such as first difference and fixed effect estimations are readily applicable because of their panel structure. So, accordingly, methods will be employed.

(Refer Slide Time: 04:34)



Cohorts and Pseudo Panel

❑ The **principle of pseudo-panels is to follow cohorts**, rather than individuals over time.

- **Cohorts** : groups of individuals sharing a set of characteristics that are fixed over time.

❑ In pseudo-panel analysis, individuals are grouped according to a criteria that do not change from one survey to another (basically, time invariant variables), such as their birth year or the education level or gender etc.

❑ **Limitation** : The efficiency of the cross- sectional dimension hampers when estimating with pseudo panel methods and it also causes a problem of heteroscedasticity in the models.

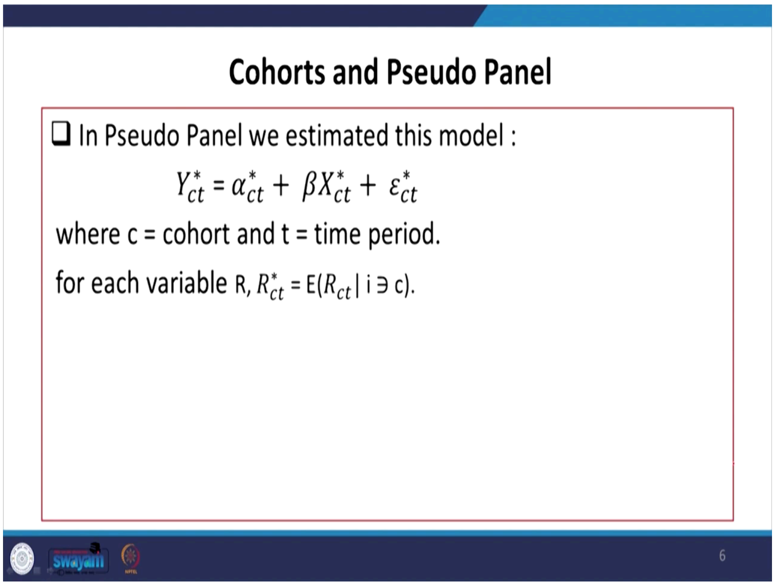
- The pseudo panel methods also suffers from the biased estimator problem.

Now, we are clarifying the cohorts and pseudo panel. The principle of pseudo panels is to follow the cohorts instead of the individual as the individual persons over time. So, the cohorts which I have just emphasized are going to be defined. These are something defined in groups of individuals sharing a set of characteristics that are actually fixed over time.

In pseudo panel analysis, individuals are grouped according to a criterion that do not change from one survey to another basically time-invariant variable are considered to be the cohorts and these are like birth year or education level or gender etc. Usually, there is not much change over time.

So, the limitations of this data are that the efficiency of cross-sectional dimension hampers when estimating with pseudo panel methods and it also occurs a problem of heteroscedasticity in the models. So, there might be some heteroscedasticity problems since and accordingly the efficiency of the model is compromised to some extent. The pseudo panel models also suffer from a biased estimator problem because of these issues.

(Refer Slide Time: 06:07)



Cohorts and Pseudo Panel

❑ In Pseudo Panel we estimated this model :

$$Y^*_{ct} = \alpha^*_{ct} + \beta X^*_{ct} + \varepsilon^*_{ct}$$

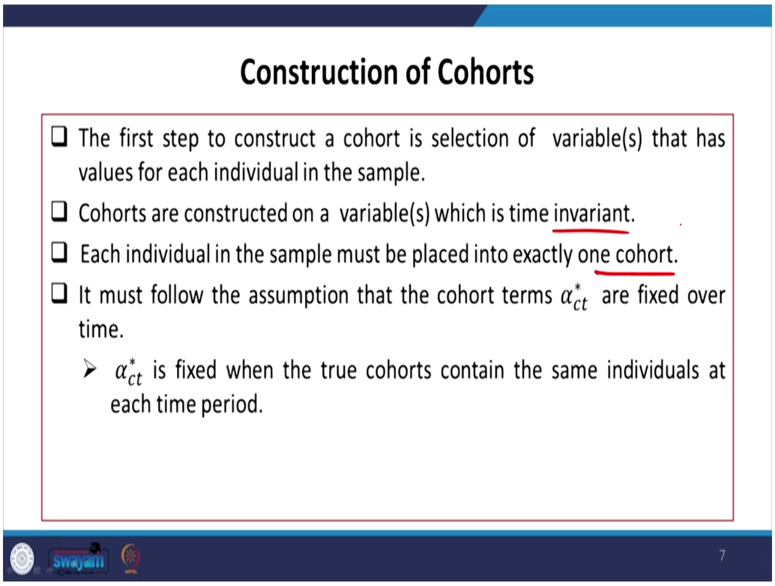
where c = cohort and t = time period.

for each variable R, $R^*_{ct} = E(R_{ct} | i \ni c)$.

In the pseudo panel instead of the individual component, we have a cohort component. Cohort we have taken c stands for cohort and t is the time dimension for each of the variables.

So, we are basically looking at the R and Rct. So, c stands for cohort and t stands for time variation and accordingly we will estimate its R-value based on these two components.

(Refer Slide Time: 06:44)



Construction of Cohorts

❑ The first step to construct a cohort is selection of variable(s) that has values for each individual in the sample.

❑ Cohorts are constructed on a variable(s) which is time invariant.

❑ Each individual in the sample must be placed into exactly one cohort.

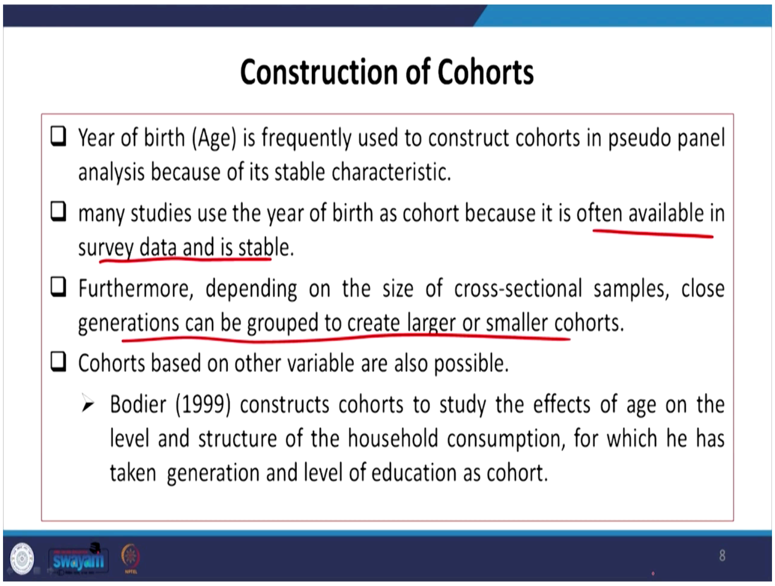
❑ It must follow the assumption that the cohort terms α^*_{ct} are fixed over time.

➢ α^*_{ct} is fixed when the true cohorts contain the same individuals at each time period.

We are now guiding the construction of cohorts. The first step is to construct a cohort-based on the selection of the variables. Those variables' values usually do not have many changes over time and that is and cohorts are constructed on a variable which is actually time-invariant.

Each individual in the sample must be placed with exactly one cohort, without that it is difficult to define. It must follow the assumption that the cohort term α is fixed over time. So, α is fixed when the true cohorts contain the same individuals at each point time period.

(Refer Slide Time: 07:48)



Construction of Cohorts

- ❑ Year of birth (Age) is frequently used to construct cohorts in pseudo panel analysis because of its stable characteristic.
- ❑ many studies use the year of birth as cohort because it is often available in survey data and is stable.
- ❑ Furthermore, depending on the size of cross-sectional samples, close generations can be grouped to create larger or smaller cohorts.
- ❑ Cohorts based on other variable are also possible.
  - ➢ Bodier (1999) constructs cohorts to study the effects of age on the level and structure of the household consumption, for which he has taken generation and level of education as cohort.

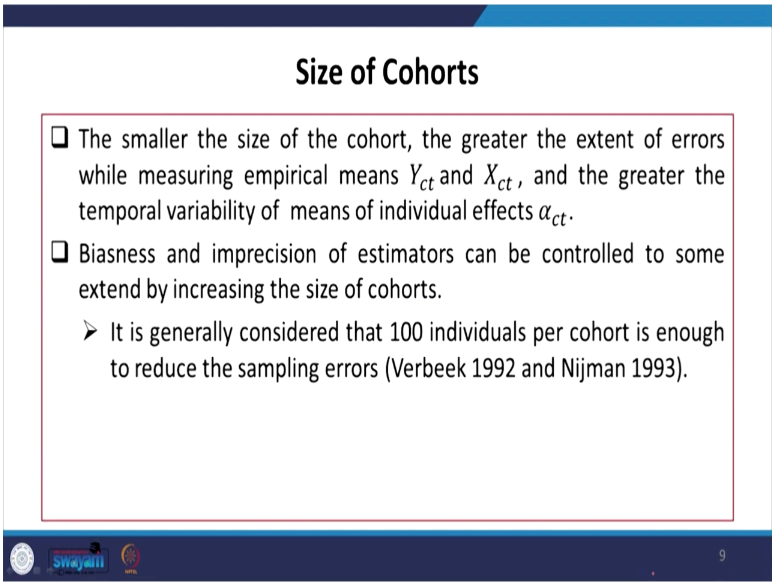
8

In the construction of the cohort year of birth is frequently used to construct cohorts in pseudo panel analysis because of its stable characteristics. Many studies use the year of birth as the cohort because of it is often available in survey data and is usually stable.

Further, depending on the size of the cross-sectional samples, close generations can be grouped to create larger or smaller cohorts. Based on the size we can be able to define bigger cohorts or smaller cohorts.

Cohorts based on other variables are also possible like Bodier 1999 constructs cohorts to study the effects of age on the level and structure of the household consumption, for which he has taken generation and level of education as a cohort.

(Refer Slide Time: 08:59)



Size of Cohorts

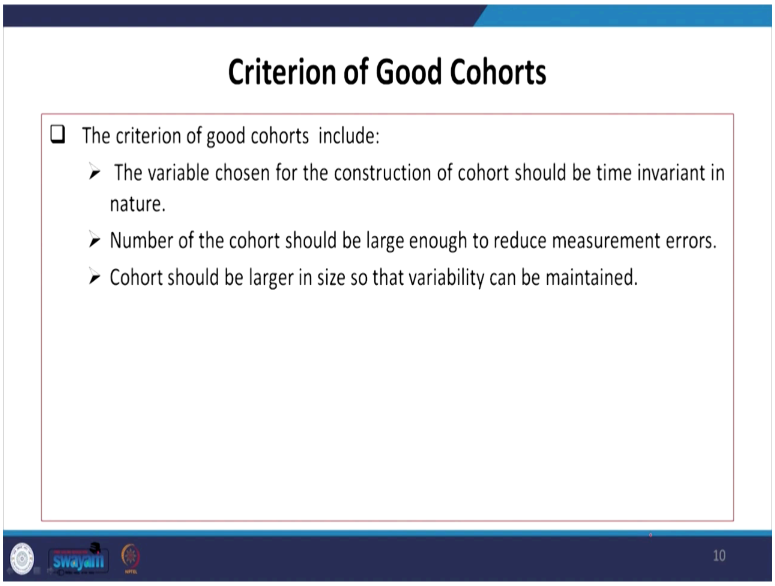
- ❑ The smaller the size of the cohort, the greater the extent of errors while measuring empirical means Y_{ct} and X_{ct}, and the greater the temporal variability of means of individual effects α_{ct}.
- ❑ Biasness and imprecision of estimators can be controlled to some extend by increasing the size of cohorts.
  - ➢ It is generally considered that 100 individuals per cohort is enough to reduce the sampling errors (Verbeek 1992 and Nijman 1993).

So, what about the size of cohorts, what should be the ideal size some suggestions will count here. The smaller the size of the cohort the greater the extent of errors while measuring empirical means such as Yct and Xct and the greater the temporal variability of means of the individual variability of the means of the individual effects.

So, there will be a greater temporal variability if the size of the cohort is actually smaller. Biasness and imprecision of estimators can be controlled to some extent by increasing the size of the cohort.

It is generally considered that 100 individuals per cohort are enough to reduce the sampling errors as per the suggestions given by these 2 authors.

(Refer Slide Time: 10:13)



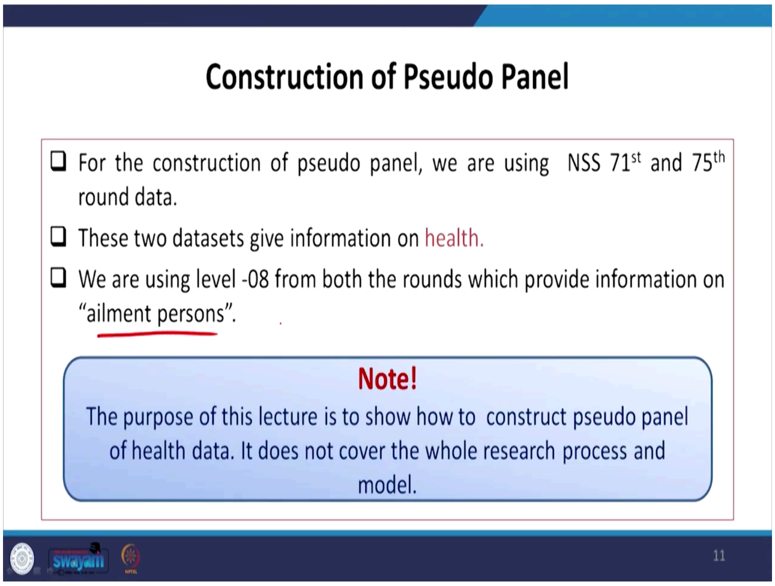
Criterion of Good Cohorts

❑ The criterion of good cohorts include:

- ➢ The variable chosen for the construction of cohort should be time invariant in nature.
- ➢ Number of the cohort should be large enough to reduce measurement errors.
- ➢ Cohort should be larger in size so that variability can be maintained.

The criteria of good cohorts include the variable chosen for the construction of the cohort should be time-invariant in nature. The number of cohorts should be large enough to reduce the measurement errors which we have just guided and the cohort should be larger in size so that the variability can be also maintained in the data.

(Refer Slide Time: 10:51)



Construction of Pseudo Panel

❑ For the construction of pseudo panel, we are using NSS 71st and 75th round data.

❑ These two datasets give information on health.

❑ We are using level -08 from both the rounds which provide information on "ailment persons".

Note!

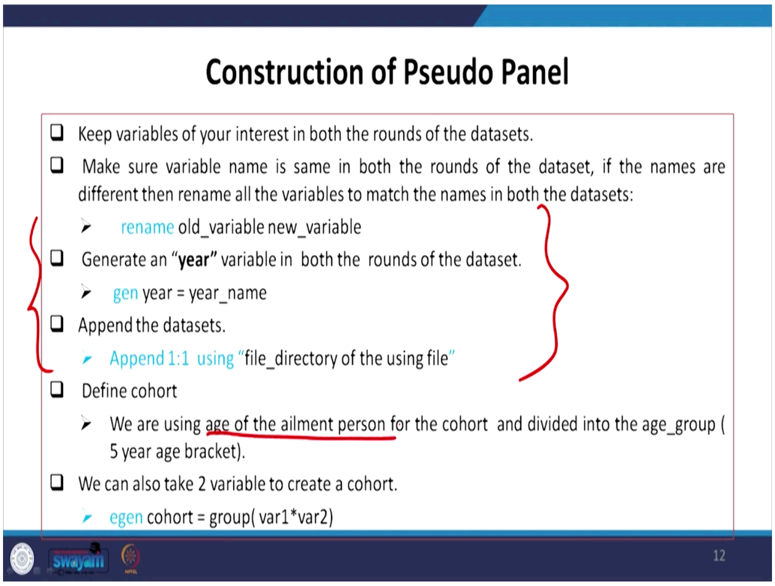
The purpose of this lecture is to show how to construct pseudo panel of health data. It does not cover the whole research process and model.

Now we are guiding about the construction of the pseudo panel. For the construction of the pseudo panel, we are using two rounds of data that is NSS 71st round and the NSS 75th round on healthcare.

These two data sets give information on health. We are using level 8 from both rounds, which provides information about the ailment of the person. The purpose of this lecture is to show how to construct a pseudo panel of healthcare data. It does not cover the whole research process and model.

We are not interpreting everything or not to guide about to complete analysis. The construction of this panel will now guide you, some of the processes we are not operating these are very common.

(Refer Slide Time: 11:44)



Construction of Pseudo Panel

□ Keep variables of your interest in both the rounds of the datasets.

□ Make sure variable name is same in both the rounds of the dataset, if the names are different then rename all the variables to match the names in both the datasets:

➢ rename old_variable new_variable

□ Generate an "**year**" variable in both the rounds of the dataset.

➢ gen year = year_name

□ Append the datasets.

➢ Append 1:1 using "file_directory of the using file"

□ Define cohort

➢ We are using age of the ailment person for the cohort and divided into the age_group (5 year age bracket).

□ We can also take 2 variable to create a cohort.

➢ egen cohort = group(var1*var2)

We need to keep variables of your interest that are going to be used in the panel data of both rounds. We need to make sure that the variable name is same in both rounds of the data sets. So, that panel can be made otherwise some errors might be generated.

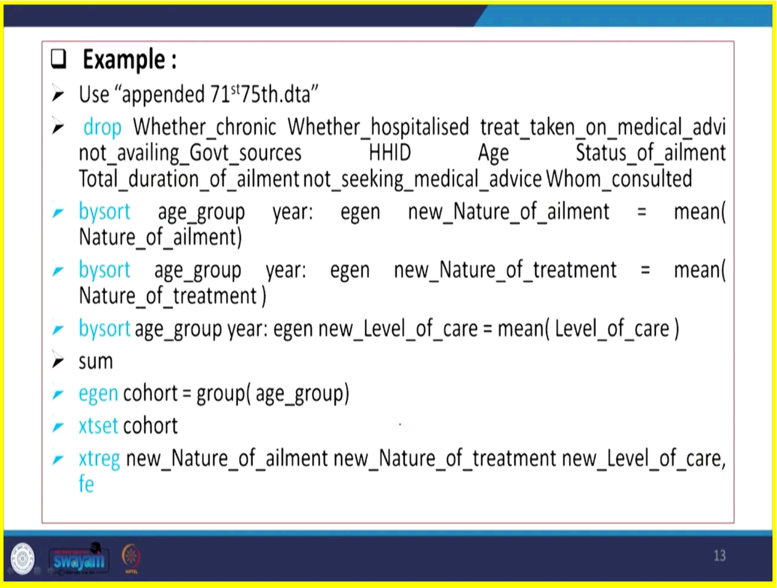
If the names are different, then rename all the variables to match the names in both datasets which is essential. So, we can rename by the command "rename old variable new variable" and we need to generate the year variable in both the rounds of the dataset. So, like year 1 or year 2 or year that particular year name we can write it down and the year variable is also generated.

Then we need to append these two datasets with 1 is to 1 like "append using file_directory". So, one is to one append without director name you can give it.

Now, we are guiding very particularly on cohort generation. So, to define cohort we are using the age of the ailment persons for cohort and dividing it into the age group. Once we define that then we are actually dividing them or classifying them into age groups with the 5-year age bracket.

So, all of them with certain different 5 years’ age brackets we can do that, then we can also take 2 variables as well. Instead of just 1 variable as a cohort, you can also generate 2 variables simultaneously. In that case, the command is “egen”.

(Refer Slide Time: 13:40)



❑ **Example :**

- Use “appended 71st75th.dta”
- drop Whether_chronic Whether_hospitalised treat_taken_on_medical_advi not_availing_Govt_sources HHID Age Status_of_ailment Total_duration_of_ailment not_seeking_medical_advice Whom_consulted
- bysort age_group year: egen new_Nature_of_ailment = mean(Nature_of_ailment)
- bysort age_group year: egen new_Nature_of_treatment = mean(Nature_of_treatment)
- bysort age_group year: egen new_Level_of_care = mean(Level_of_care)
- sum
- egen cohort = group(age_group)
- xtset cohort
- xtreg new_Nature_of_ailment new_Nature_of_treatment new_Level_of_care, fe

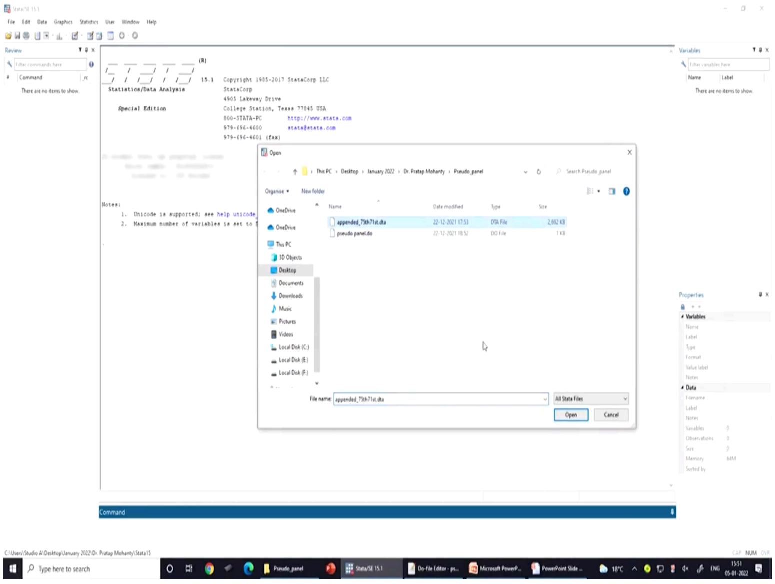
So, the example dataset we are using as I already mentioned is NSS 71 and 75th round. Now, I am just giving you the direction at this moment then we will use it directly from there. First, we drop the unnecessary variable then we will sort the age variable.

Therefore, the bysort command is used. Then bysort year with age group year and also we have taken the mean value of the variable that is nature of ailment and then we have defined these steps in front of you level of care then level of nature of the treatment, then nature of ailment three important aspects we have also expressed in by their mean values.

We have sort the data by the group year of that age variable, then we will summarize that and we will generate the cohort by the age group and then we will move to understand or define the data as a cohort as the id variable.

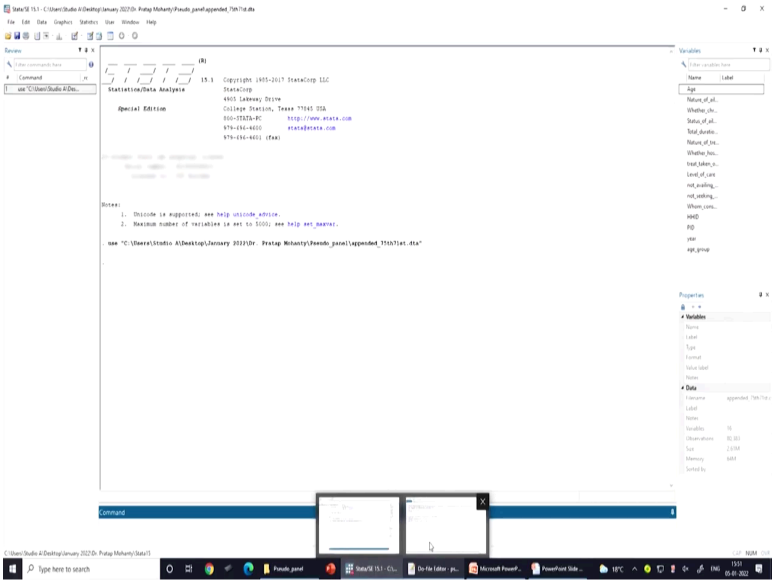
Then accordingly we will run the regression with a fixed-effect model because it is time invariance in nature.

(Refer Slide Time: 15:23)



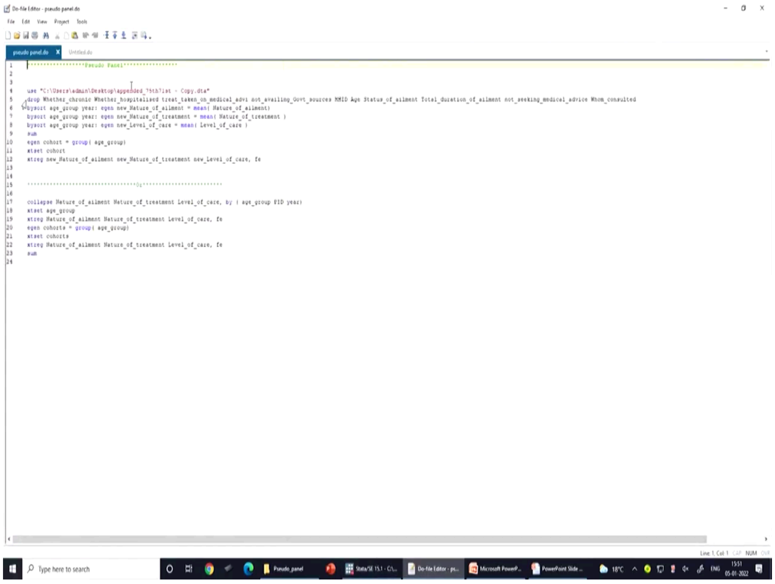
```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   15.1   Copyright 1985-2017 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas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

Notes:
      1.  Unicode is supported; see help unicode_
      2.  Maximum number of variables is set to 5
```

(Refer Slide Time: 15:31)



```
  ___  ____  ____  ____  ____ (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)

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

. use "C:\Users\Studio A\Desktop\January 2022\Dr. Pratap Mohanty\Pseudo_panel\appended_75th71st.dta"

.
```

(Refer Slide Time: 15:32)

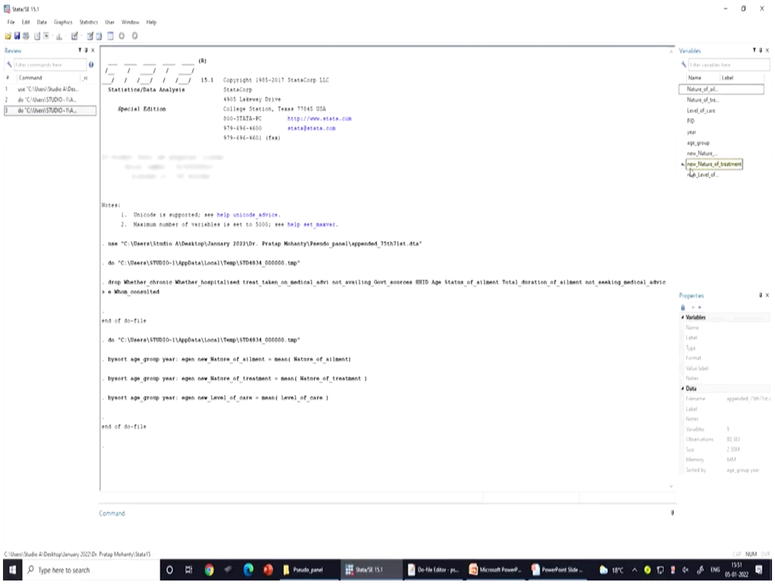


```
****************Pseudo Panel****************


use "C:\Users\admin\Desktop\appended_75th71st - Copy.dta"
drop Whether_chronic Whether_hospitalised treat_taken_on_medical_advi not_availing_Govt_sources HHID Age Status_of_ailment Total_duration_of_ailment not_seeking_medical_advice Whom_consulted
bysort age_group year: egen new_Nature_of_ailment = mean( Nature_of_ailment)
bysort age_group year: egen new_Nature_of_treatment = mean( Nature_of_treatment )
bysort age_group year: egen new_Level_of_care = mean( Level_of_care )
sum
egen cohort = group( age_group)
xtset cohort
xtreg new_Nature_of_ailment new_Nature_of_treatment new_Level_of_care, fe


*********************************Or************************

collapse Nature_of_ailment Nature_of_treatment Level_of_care, by ( age_group PID year)
xtset age_group
xtreg Nature_of_ailment Nature_of_treatment Level_of_care, fe
egen cohorts = group( age_group)
xtset cohorts
xtreg Nature_of_ailment Nature_of_treatment Level_of_care, fe
sum
```

Now, these are the steps which I am just going to show on the screen. We will open the do file on your screen and we can show that these are all important and how you can use them for your work directly.

So, we can drop the variables which are not relevant for our analysis, you can also drop on your own way. So, we have dropped it, and now the number of variables have been dropped. Now we will sort it by age group, years and mean value generated for each of the variables.

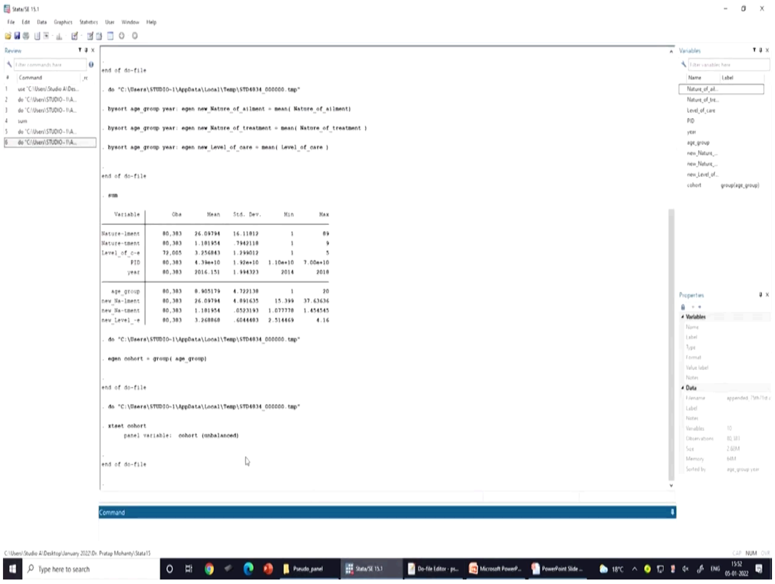
(Refer Slide Time: 16:02)



```
  ___  ____  ____  ____  ____ (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)

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

. use "C:\Users\Studio A\Desktop\January 2022\Dr. Pratap Mohanty\Pseudo_panel\appended_75th71st.dta"

. do "C:\Users\STUDIO~1\AppData\Local\Temp\STD4834_000000.tmp"

. drop Whether_chronic Whether_hospitalised treat_taken_on_medical_advi not_availing_Govt_sources HHID Age Status_of_ailment Total_duration_of_ailment not_seeking_medical_advic
> e Whom_consulted

.
end of do-file

. do "C:\Users\STUDIO~1\AppData\Local\Temp\STD4834_000000.tmp"

. bysort age_group year: egen new_Nature_of_ailment = mean( Nature_of_ailment)

. bysort age_group year: egen new_Nature_of_treatment = mean( Nature_of_treatment )

. bysort age_group year: egen new_Level_of_care = mean( Level_of_care )

.
end of do-file

.
```

Now, three you can see the nature of the treatment, nature of ailment and level of care alright. Now next is our summarizing our data.

(Refer Slide Time: 16:18)



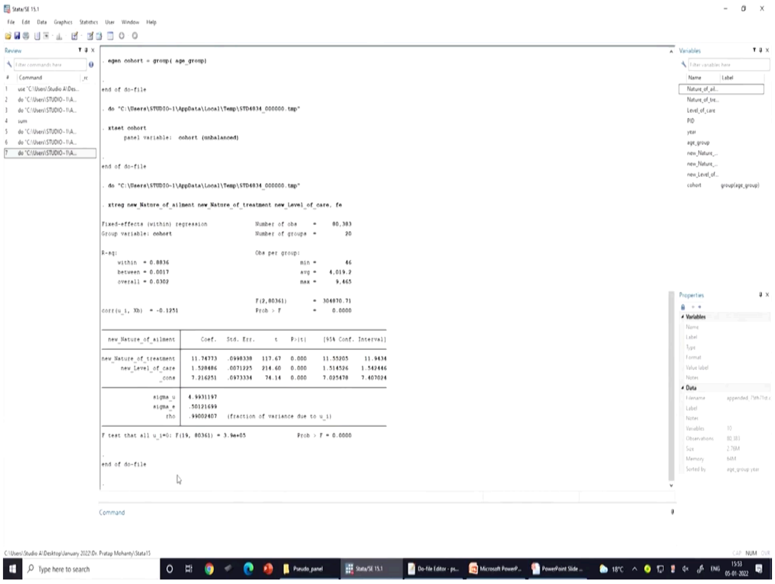
```
.
end of do-file

. do "C:\Users\STUDIO~1\AppData\Local\Temp\STD4034_000000.tmp"

. bysort age_group year: egen new_Nature_of_ailment = mean( Nature_of_ailment)

. bysort age_group year: egen new_Nature_of_treatment = mean( Nature_of_treatment )

. bysort age_group year: egen new_Level_of_care = mean( Level_of_care )

.
end of do-file

. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
Nature~lment |     80,383    26.09794    16.11012          1         89
Nature~tment |     80,383    1.101954    .7942118          1          9
Level_of_c~e |     72,005    3.256843    1.299012          1          5
         PID |     80,383    4.39e+10    1.92e+10   1.10e+10   7.00e+10
        year |     80,383    2016.151    1.994323       2014       2018
-------------+---------------------------------------------------------
   age_group |     80,383    8.905179    4.722138          1         20
new_Na~lment |     80,383    26.09794    4.891635     15.399   37.63636
new_Na~tment |     80,383    1.101954    .0523193   1.077778   1.454545
new_Level_~e |     80,383    3.268868    .6044403   2.514469       4.16

. do "C:\Users\STUDIO~1\AppData\Local\Temp\STD4034_000000.tmp"

. egen cohort = group( age_group)

.
end of do-file

. do "C:\Users\STUDIO~1\AppData\Local\Temp\STD4034_000000.tmp"

. xtset cohort
       panel variable:  cohort (unbalanced)

.
end of do-file

.
```

In the summary, you can see each of the observations and their mean value and their standard deviation and everything is presented on your screen.

The variation is not much higher here, because the mean value has been taken in the case. The next one is to group the variable that is the cohort. We are generating an age group and it has it is confined to the age variable in the grouping of the variable and that is considered as the panel component.

We will define the command here with the xtset cohort as the panel variable and this is how the stata has read that this is the variable panel variable and this is an unbalance data. Now we can run the regression with its fixed effect model because we know that we have taken a time-invariant model.

(Refer Slide Time: 17:35)



```
. egen cohort = group( age_group)

.
end of do-file

. do "C:\Users\STUDIO~1\AppData\Local\Temp\STD4034_000000.tmp"

. xtset cohort
       panel variable:  cohort (unbalanced)

.
end of do-file

. do "C:\Users\STUDIO~1\AppData\Local\Temp\STD4034_000000.tmp"

. xtreg new_Nature_of_ailment new_Nature_of_treatment new_Level_of_care, fe

Fixed-effects (within) regression               Number of obs     =     80,383
Group variable: cohort                          Number of groups  =         20

R-sq:                                           Obs per group:
     within  = 0.8836                                         min =         46
     between = 0.0017                                         avg =    4,019.2
     overall = 0.0302                                         max =      9,465

                                                F(2,80361)        =  304870.71
corr(u_i, Xb)  = -0.1251                        Prob > F          =     0.0000

------------------------------------------------------------------------------
new_Nature_of_ailment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------+-------------------------------------------------------
new_Nature_of_treatment |   11.74773   .0998338   117.67   0.000     11.55205     11.9434
    new_Level_of_care |   1.528486   .0071225   214.60   0.000     1.514526    1.542446
                _cons |   7.216251   .0973334    74.14   0.000     7.025478    7.407024
----------------------+-------------------------------------------------------
              sigma_u |  4.9931197
              sigma_e |  .50121699
                  rho |  .99002407   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(19, 80361) = 3.9e+05                Prob > F = 0.0000

.
end of do-file

.
```

The fixed-effect model has been drawn and you can now see that the p-value is significant and the rest of the interpretation we usually do is also applied here as well. This is how it works and as we already said that we are not going to interpret everything as like a paper.

We have generated or constructed a pseudo panel with the same technique you can apply with so many other databases and once your panel data is generated you can also work for the number of methodologies like policy evaluation techniques in your paper and those are going to be highly appreciated by the review because of its policy implications.

So, these are all for today and we have covered the pseudo panel for you. If there are any sort of queries in this regard, please do not hesitate and try to come to us. We will try to deal with our team and will be most happy to address it.

Thank you.