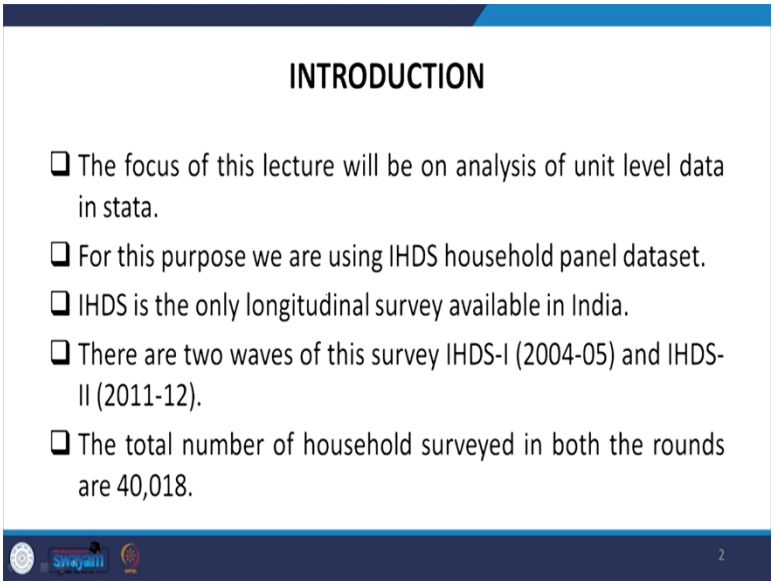
Handling Large-Scale Unit Level Data Using Stata
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences,
Indian Institute of Technology, Roorkee
Lecture 40
Analysis of Panel Data in Stata

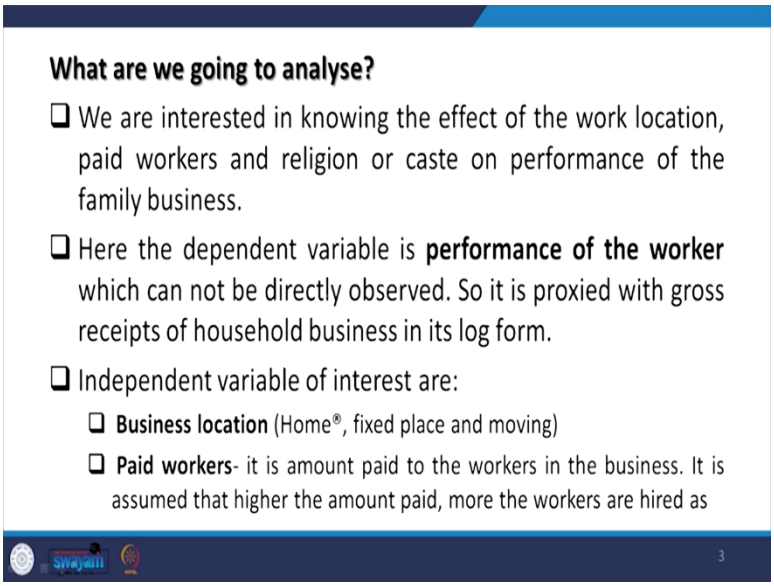
Welcome, friends, once again to the very last lecture of this module on Analysis of Panel Data in Stata. Since you have been following the lectures very seriously, I can infer that you can able to surely interpret the panel data in this format very correctly. And so far in last two lectures we have defined and constructed the panel data carefully and we mentioned using the IHDS both the rounds of the data. And in this final lecture, we will be analysing the panel data.

(Refer Slide Time: 01:20)



So, the focus here will be on the unit level data in Stata and for this purpose, we are using IHDS household panel data and IHDS is the only longitudinal data, as I already mentioned. There are two waves that is IHDS-I and IHDS-II. The total number of household survey in both the rounds is of 40,018 households.

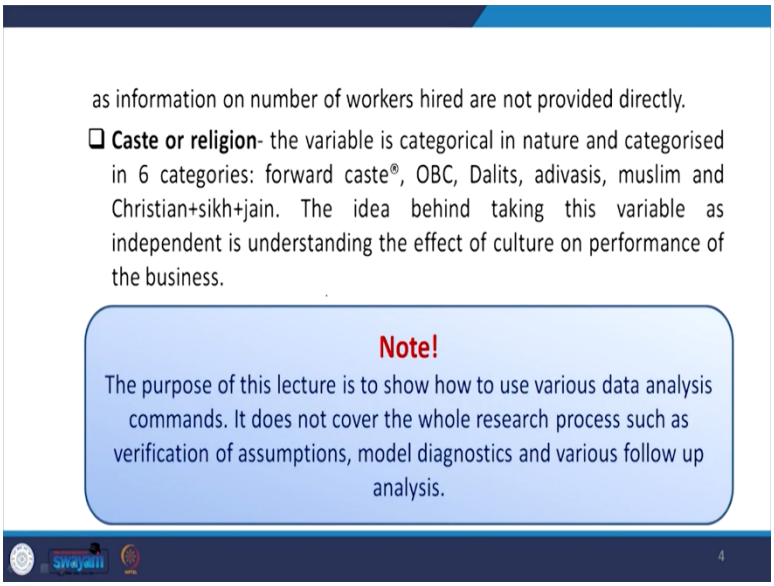
(Refer Slide Time: 01:38)



We are going to analyse the aspects such as we wish to know the effect of work location, then paid workers, religion or caste on performance of the family business. So, in this data, suppose we wanted to analyse the performance of the family business through their religion, caste, whether they are paid workers or their location. So, we are sticking to the performance of the workers since that is not directly observed, we take a proxy variable called a gross receipt of the household business in its log form.

So, gross receipt in absolute number is there. So we have converted into a logarithmic transformation of it. So, the independent variable of interest for us is, as I already mentioned, business location that is maybe home based or fixed place or moving type or paid workers are like, whether they are being paid for their business and it is assume that higher the amount paid more the workers are hired as informational number of workers hired are not provided directly.

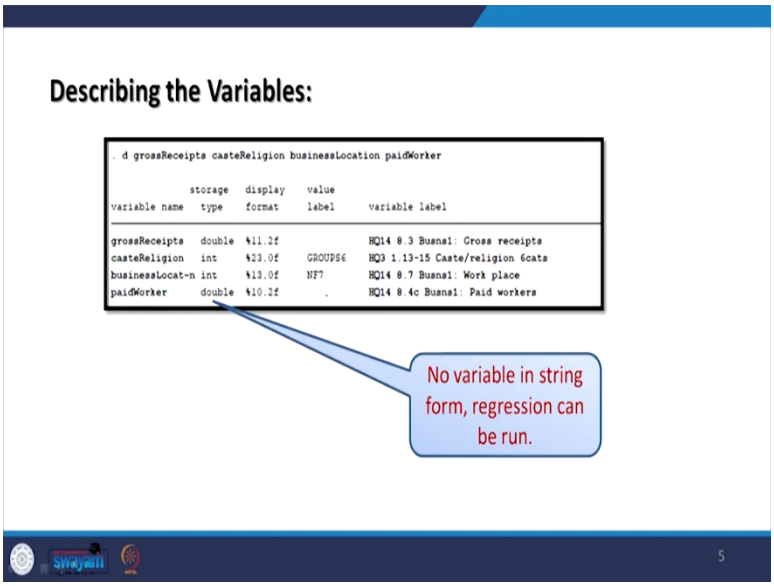
(Refer Slide Time: 02:54)



Relating to caste or religion, the variable is categorical in nature and categorized in six numbers or six categories, forward caste, OBC, Dalits, Adivasis, Muslims, Christians, Sikhs and Jain are combined together. The idea behind taking this variable as independent is understanding the effect of culture on performance of the business.

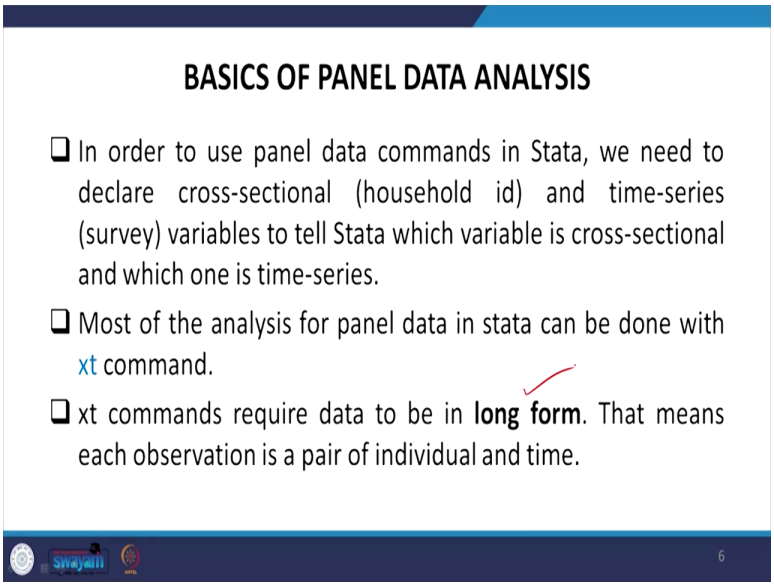
To be noted, the purpose of this lecture is to show how we use various data analysis and their commands. It does not give the whole research process, such as the verification of the assumption of the models, their diagnostics and various follow up analysis usually we do in our earlier lectures, but here we only give you how to use panel.

(Refer Slide Time: 03:58)



Basically, we need to check here that no variables are in string form. That is very important. And then only regression can be run. So, we need to check through the describe command.

(Refer Slide Time: 04:13)

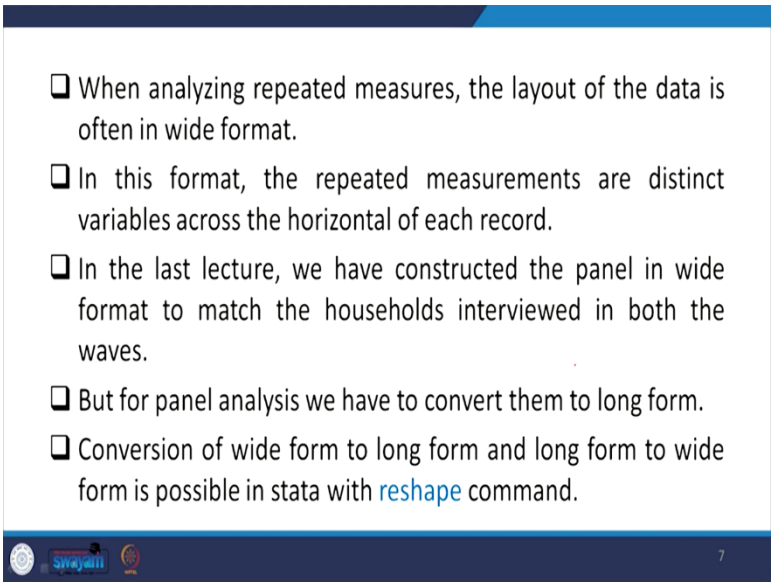


Some of the basics of the panel data analysis is that to use panel data commands in Stata, we need to declare cross-sectional household id, time series survey id. Here it is that survey id, in the name survey or the variable to be with time series component. We need to specify to Stata which variable is cross-sectional and which one is time series. Most of the analysis for panel data in

Stata can be done with xt command. Xt is highlighted here. Xt command is very relevant in most of the panel analysis.

So, xt commands require data to be a long form to make it very clear that the data has to be a long form. Long and wide, we have already mentioned, but we will also show it with the help of data. This means that each observation is a pair of individual and time. Each observation must be given with the time component with the individual information.

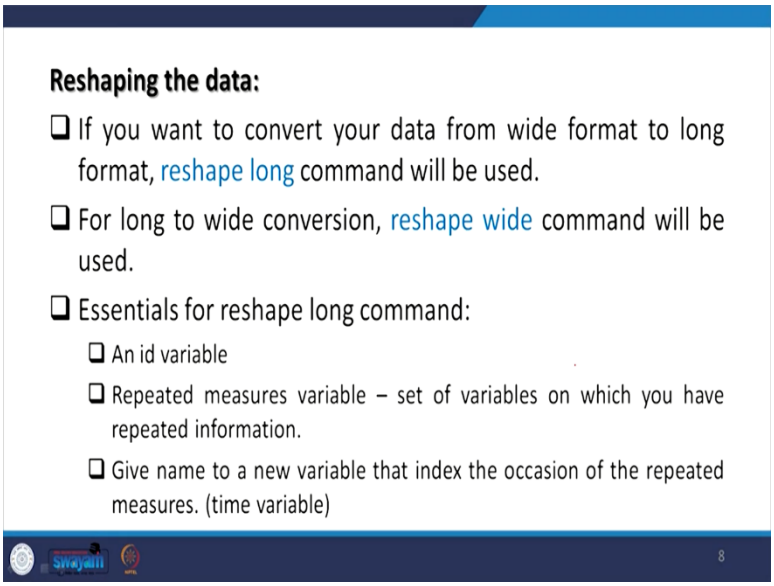
(Refer Slide Time: 05:43)



While analysing repeated measures the layout of the data is often in wide format. In this format the repeated measurements are distinct variables across horizontal of each record. Basically, each repeated information are given in horizontal series or horizontal entries.

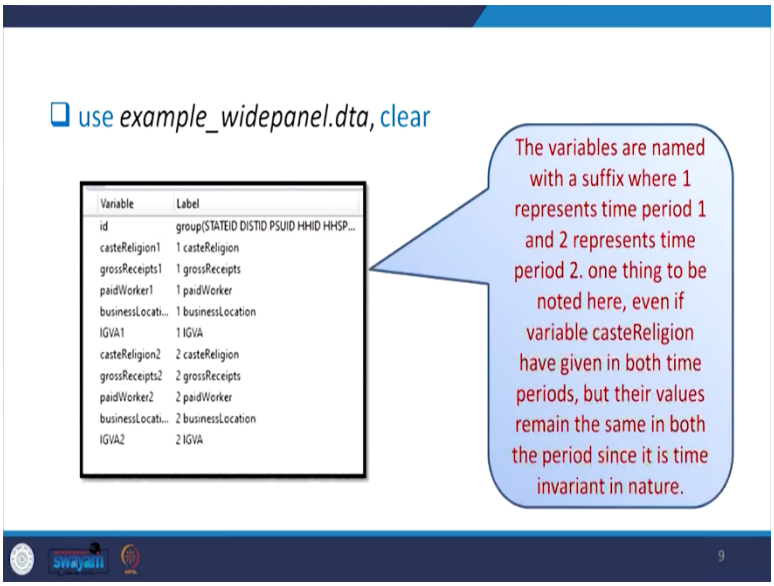
So, in the last lecture, we have constructed the panel in wide format to match the household interviewed in both the waves. But for panel analysis we have to convert them to a long form for analysis. The conversion of wide form to long and long to wide form is very much possible in Stata with a command called reshape. Reshape is highlighted in blue colour. So, we will do it right now.

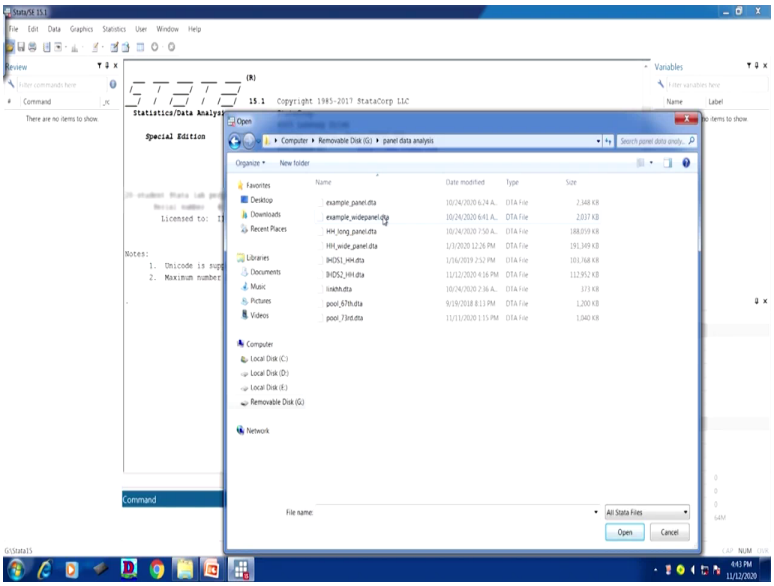
(Refer Slide Time: 06:17)

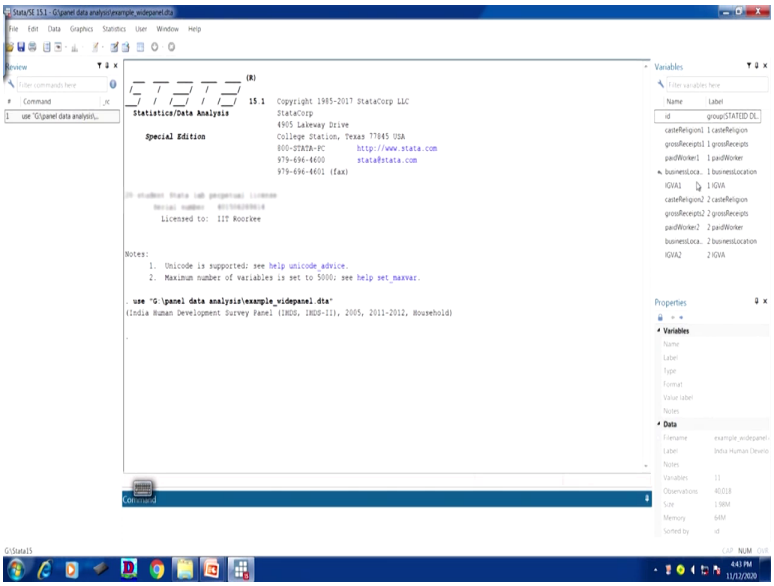


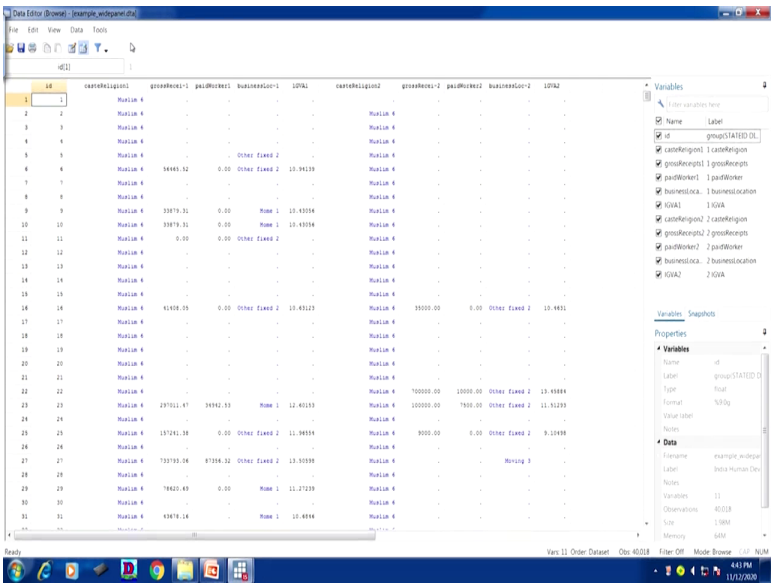
So, reshaping the data is necessary if we want to convert your data from wide format to long format. So, we will simply give reshape long command, reshape long and it will be converted very clearly. Similarly, if long is available, we need to convert it to wide, so simply we will give reshape wide as the command. So, essential for reshape long command is that an id variable. Those must be discussed like id variables, repeated measures variable, set of variables on which you have repeated information and give name to a new variable that is index the occasion of the repeated measures that is time component. So, time component and id information has to be very clearly specified.

(Refer Slide Time: 07:08)









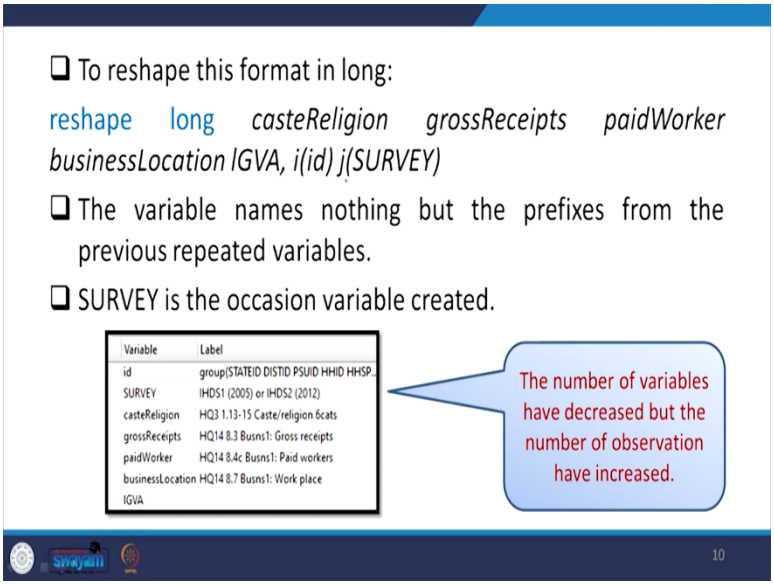
Like in the example we will also provide this information to you. We have an example wide panel data that we converted in the last class also. But we are discussing once again with the name example wide panel data. We will simply use that data right now, example wide panel data, and from there we will discuss like panel data example wide panel.

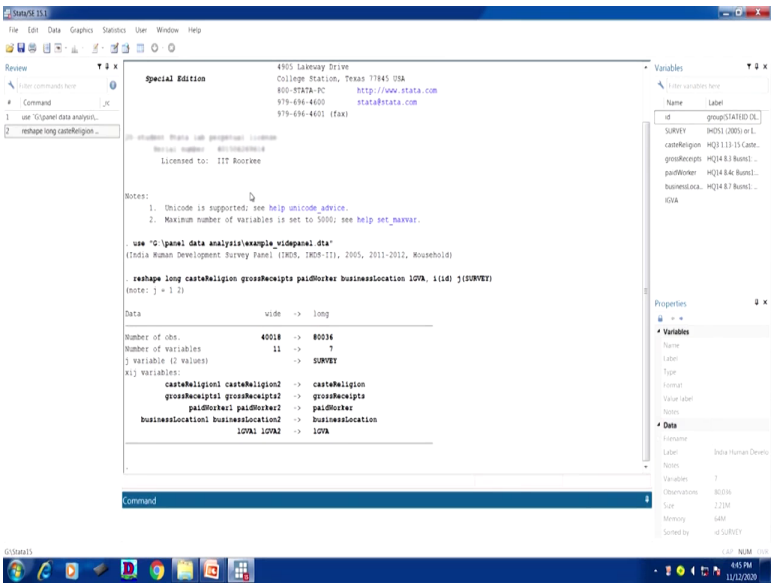
So, this has been opened. You can see the data is given. I will also interpret here. This is the variable and label is very clearly given at the right hand side. Variable name like caste, religion with one, gross receipt one and period one, then it is caste, religion and period two, gross receipt period two, paid worker in period two. So, those are given very clearly and you can check the

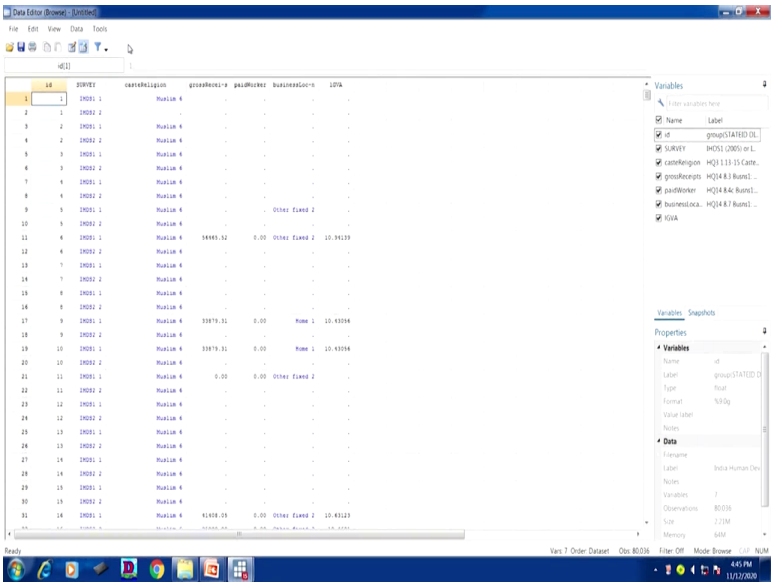
data also. And since this is a wide panel, because the repeated information related to caste, religion, it is given in horizontal order. But Stata requires for the analysis, we generally require a long format data we have to convert it.

So, this is what the data created. The variables are named with a suffix where 1 represents time period 1 and 2 represents time period 2. And like some variables are time invariant, like caste of the person generally of time invariant, but largely most of the variables are changing over time. So, better to rename with different variable because they are changing nature of those variables.

(Refer Slide Time: 09:21)







We will reshape it to long format. Since I have already shown that it is available in wide format. So, we will simply reshape it. But reshape with this particular name we have to specify what is our id variable because then only Stata can able to reshape it with id against time. So, we have to define what is our j that is time and id variables, who is the id and who is time. So, we will specify it. So, here will simply take it and will find out with this.

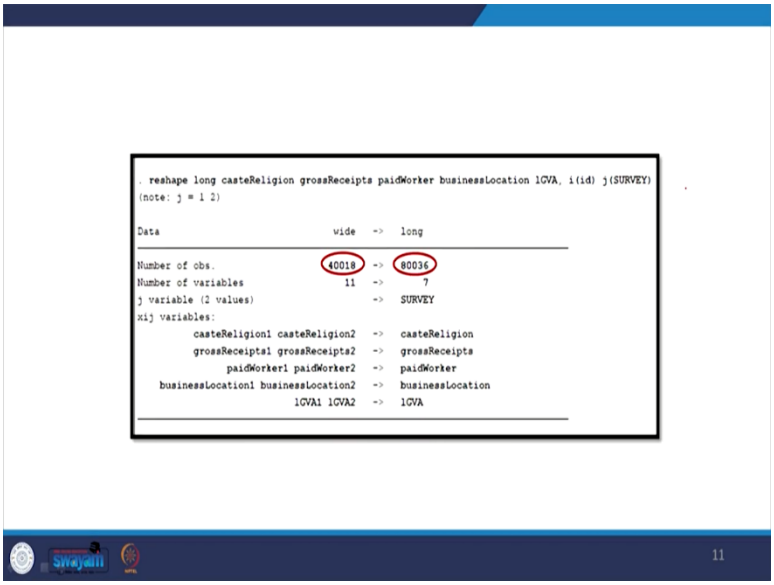
Once we specify it we enter it. We can see that these are information. I want to show it. Look at number of columns have been reduced. But the number of observation has been increased to 80,036. I will show you here. Look at this. Here the id is the observation who has responded 40,008 in total in one round. But since we have put it in long format, in vertical format, this is survey round one. This is for same id, the same person who is responded in survey time period one and time period two. The person three-time period one, and similarly, other information has been saved accordingly.

So, I wanted to just inform here that the number of variables were 11 earlier since it was of wide panel. Since we have made it clubbed into a vertical format, a long format, so the number of variables have been reduced, but the number of observation has been increased to 80,036 from 40,018.

So, we will go back to our PPT. So, the variable names nothing but the prefixes from the previous repeated variables. Survey is the occasion variable created that is the time variable

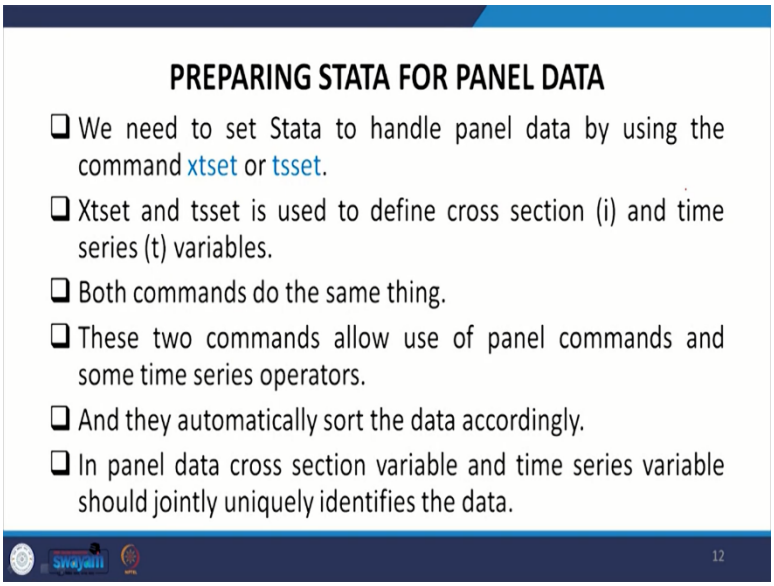
created. The number of variables have decreased, but the number of observation have increased as I already mentioned. So, here also against to our variable and their labels you can easily mark the difference as compared to the previous wide format data. We will come to the next.

(Refer Slide Time: 12:04)



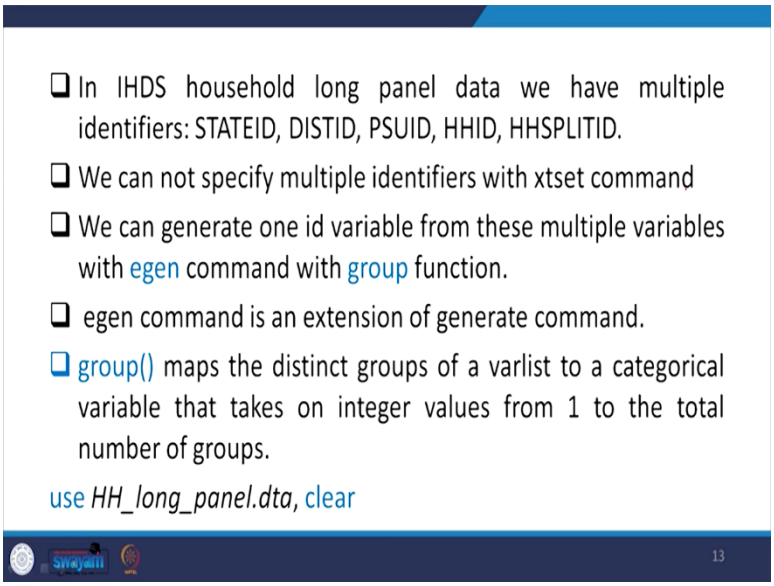
This is what we have explained.

(Refer Slide Time: 12:08)



We are preparing Stata for panel data analysis. We need to set Stata to handle panel data by using the command xtset or tsset. These two commands are interchangeable used xtset or tsset. So, xtset or tsset they both gives the same result. These commands are used to define cross section that is i and the time component t variables. We will show it right now. Both commands do the same thing. These two commands allow use of panel commands and sometime series operators and they automatically sort the data accordingly. In panel cross data section, panel data cross section variable and time series variable should jointly uniquely identify the data. We will tell you.

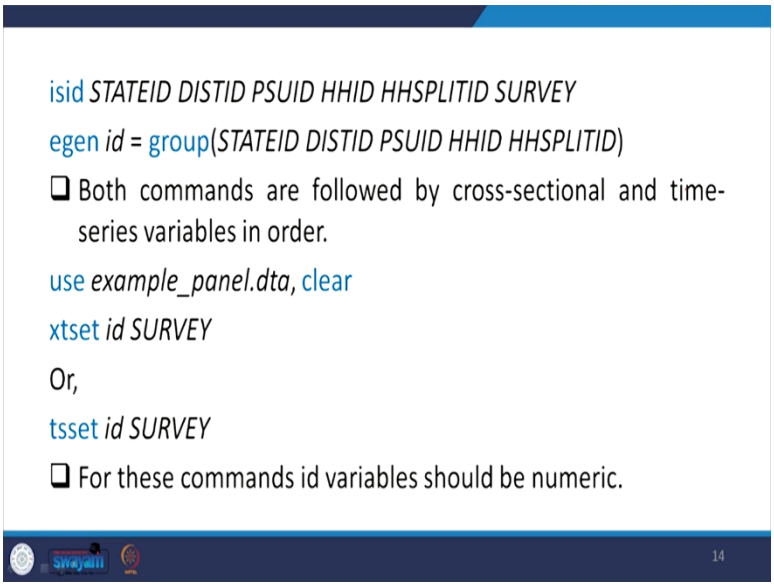
(Refer Slide Time: 13:10)

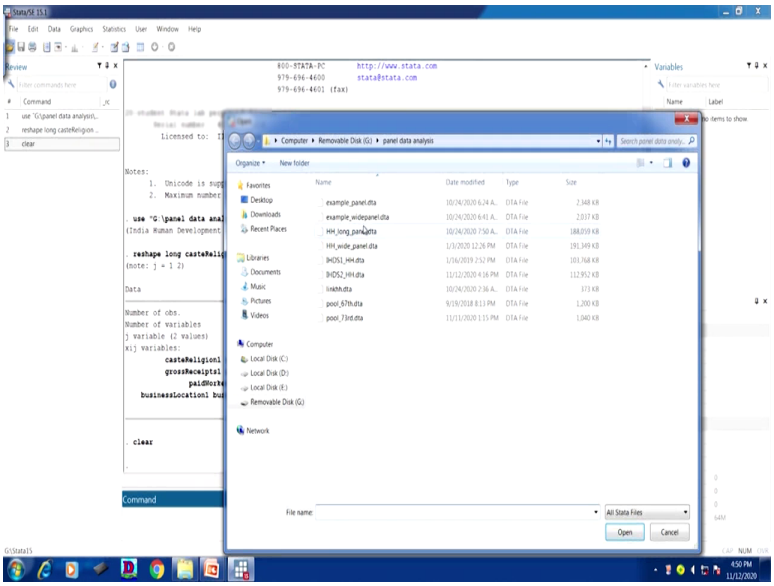


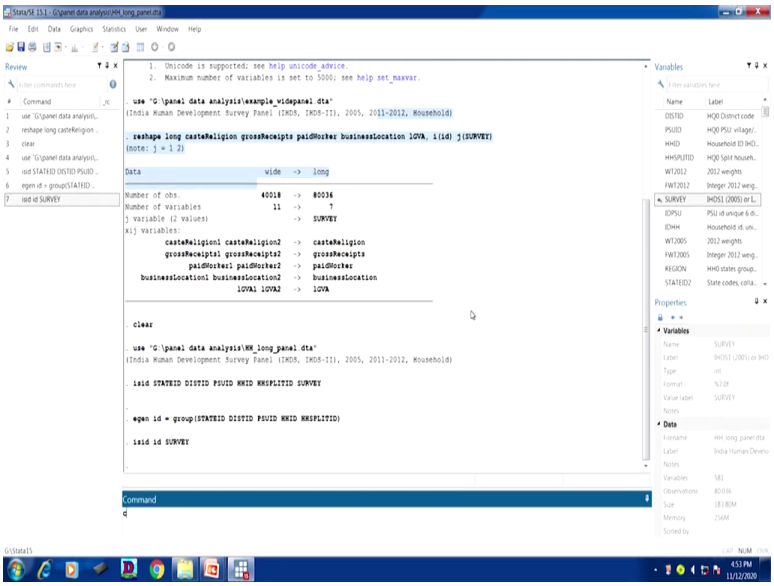
Like in IHDS household long panel data we have multiple identifiers that is state ID, district ID, PSUID, HHID and split ID. We cannot specify multiple identifiers with the xtset command. Xtset command we cannot have multiple identifiers. We have to make it a group. We can generate one id variable from these multiple variables with egen command. Egen command also we discussed earlier, but once again we are discussing that egen command with group function once we mention group with all those five variables it will combine into a particular variable.

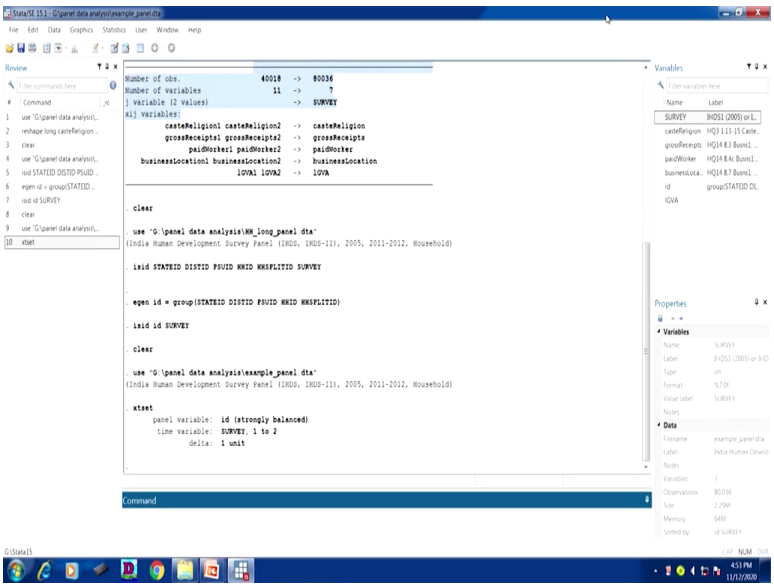
So, group maps like egen command is an extension of generate command group with the specific variables, maps the distinct groups of a list, of a variable is to a categorical variable that takes an integer values from 1 to the total number of groups.

(Refer Slide Time: 14:17)









And by using household long panel we will also show it here, like we will use a household long panel data we have already defined. We will clear it here first. Then we will use a household long panel data. So, this is getting opened. In between we will show it. It is 80,036 observations. We have opened this Stata for our explanation. So, we will define a group id, group variable with the name id with egen command. So, egen, that is ID and group. So, ID I think we already. So, we have defined.

First of all, we need to understand whether they are uniquely identifying or not. So, these are there. Just a minute. What I will do, we will take it to here. And so then come back to this. And

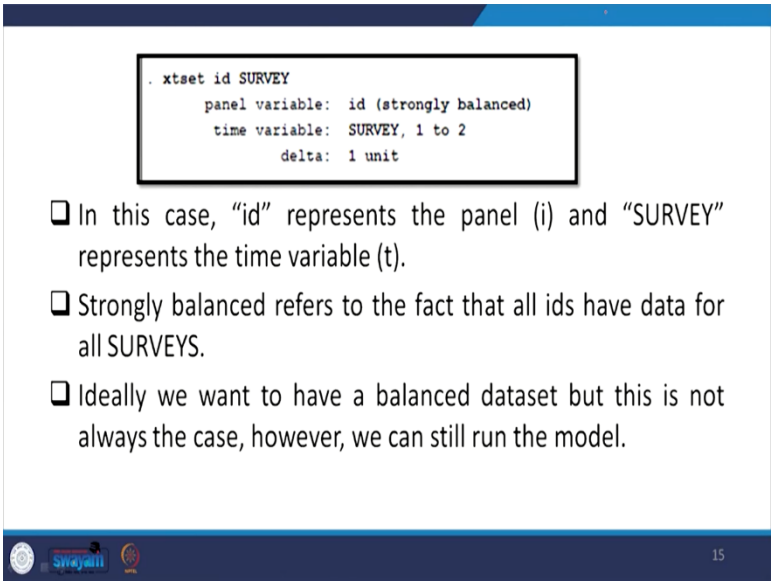
id variable has been generated. So, this id variable is going to be your uniquely identified variable that composed of five indicators.

So, you can also take ISID with id only so that also defined to be id going to be uniquely identified. Since it is composed of time as well, so we have not taken. So, survey was not included in the group. Both the commands are followed by cross sectional. I just wanted to mention that both the information, both the commands are followed by cross sectional and time series variables in order.

We will use once again the example panel data and we will also check it xtset, whether that is perfectly with panel format or not. So, we will use the example panel data for explanation. So we will clear it and we will use the example panel data. If you simply go by xtset or tsset, so this will give us the information. This clearly confirms that, it is a strongly balanced panel data and it has the survey content from first period to second period. There are two time periods with a change of time to be one.

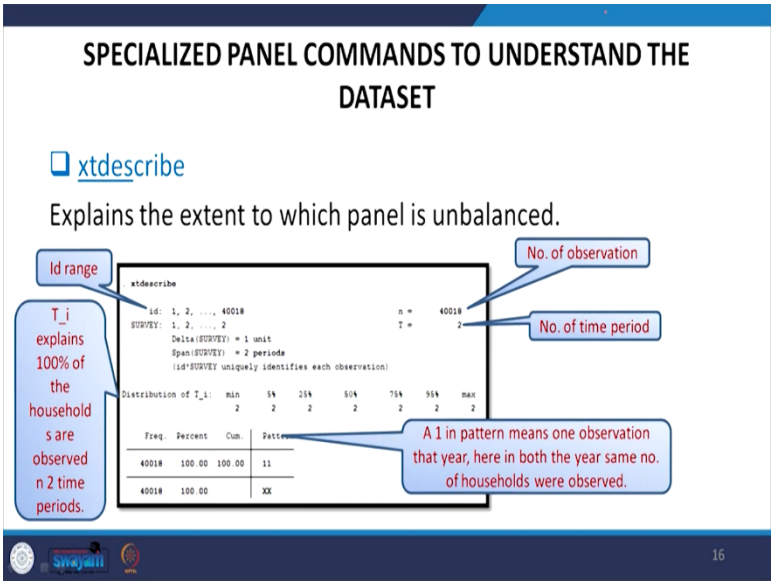
So, these, for these commands id variables would be numeric. We must be having id variables to be numerical. Usually, id variables are in string, but here xtset does not read the string variables.

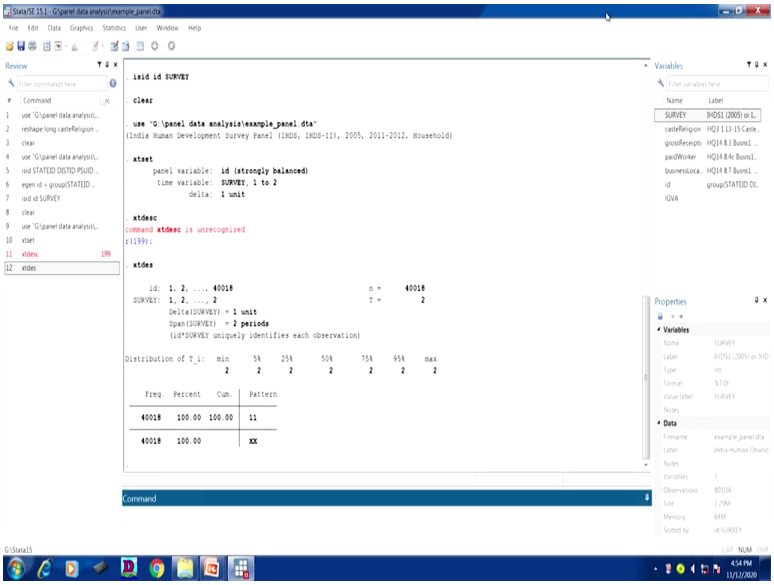
(Refer Slide Time: 18:19)



In this case id represents the panel id information and survey represents the time component and it is a strongly balanced panel data. Ideally, we want to have a balanced dataset in our analysis, but this is not always the case. However, we can still run the model without a balance data as well.

(Refer Slide Time: 19:15)

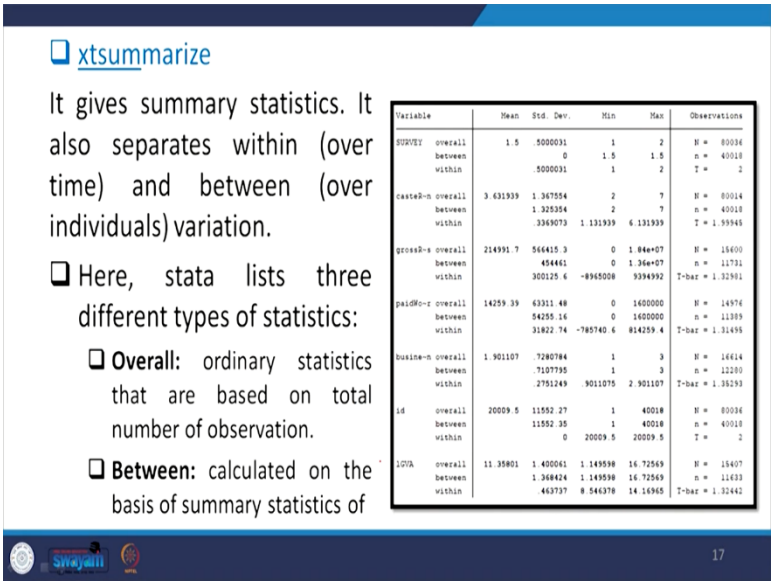


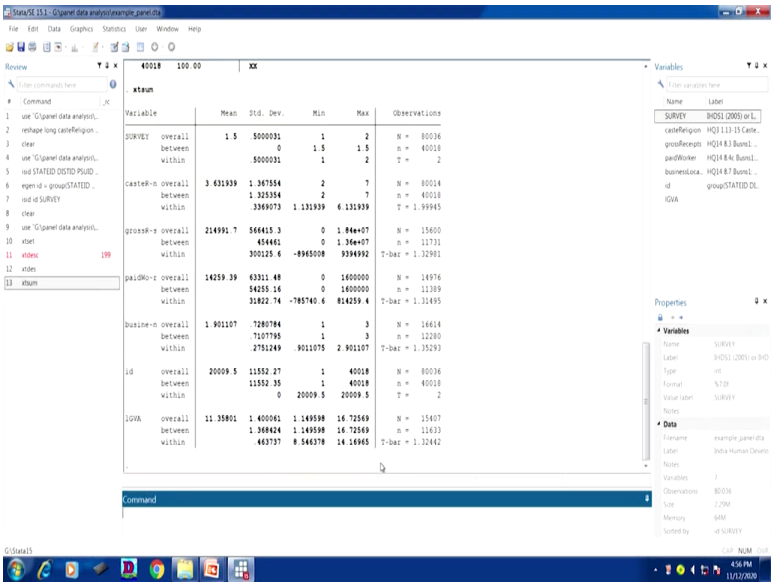


The usual way of explanation of almost all the dataset we do with their description. Here we do xtdescribe it simply gives us the explanation xtdes. It gives us explanation for how many observation 40,018 observations are there in total in two time periods. There t is two, over a span

of two periods is given. It also gives information of the pattern. Pattern suggests that it is one stands for one period information means one observation in that year and here in both the years same number of households are observed. So, that is why in both, here it is one and one is given. If same number of house information are not available then there must be a dot or not available indicators must have been given. I think I have already explained. I must proceed.

(Refer Slide Time: 19:45)



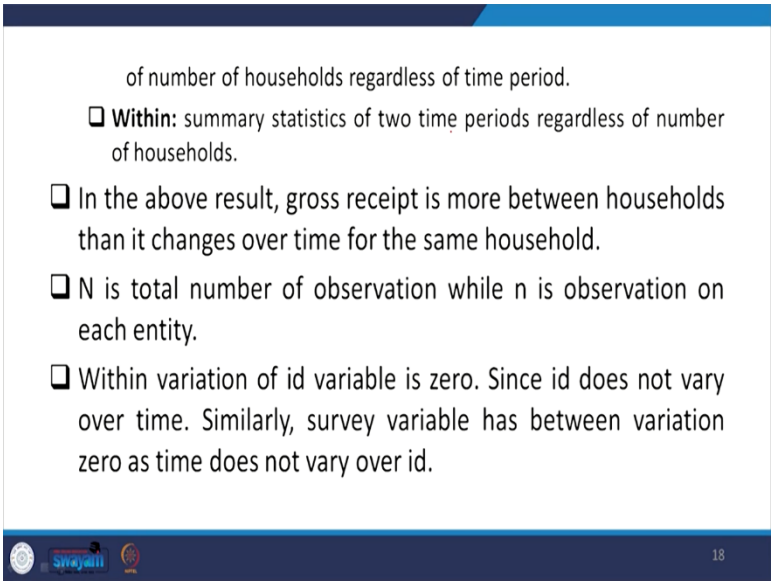


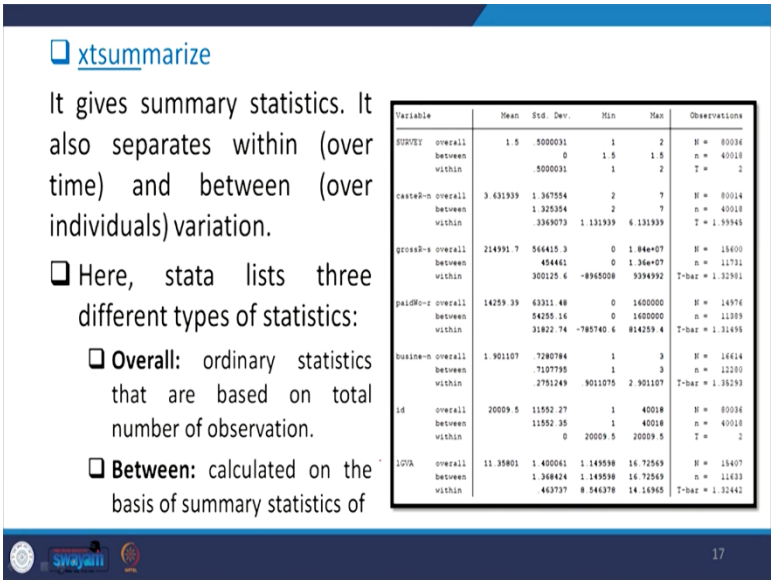
Coming to summarize the way we summarize, we generally summarize in the cross section, between analysis, basically, between the observation we compare in case of summarize. But in

panel data we have all the information basically, between the observation and within the time, within the household between the time period that is called within analysis is also important and overall explanation is also possible. So, xtsum if you do it, will find out the information of all that is being discussed here.

So, everything is mentioned. For your explanation I just wanted to give that overall description, like ordinary statistics that are based on total number of observation is given in overall indicators. Between is important so far as calculation of the summary statistics of number of households, regardless of time period. Basically, it is cross sectional information given so far as between is concerned, because time component is not considered while understanding their summary statistics.

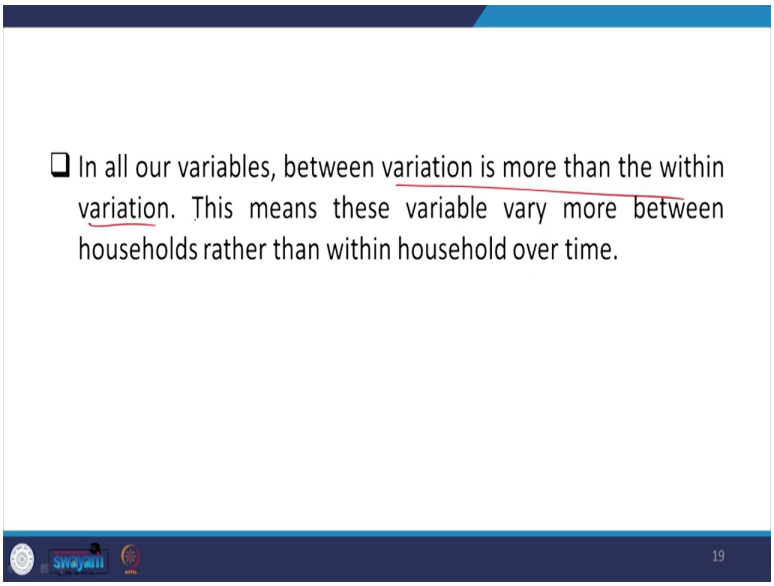
(Refer Slide Time: 20:46)





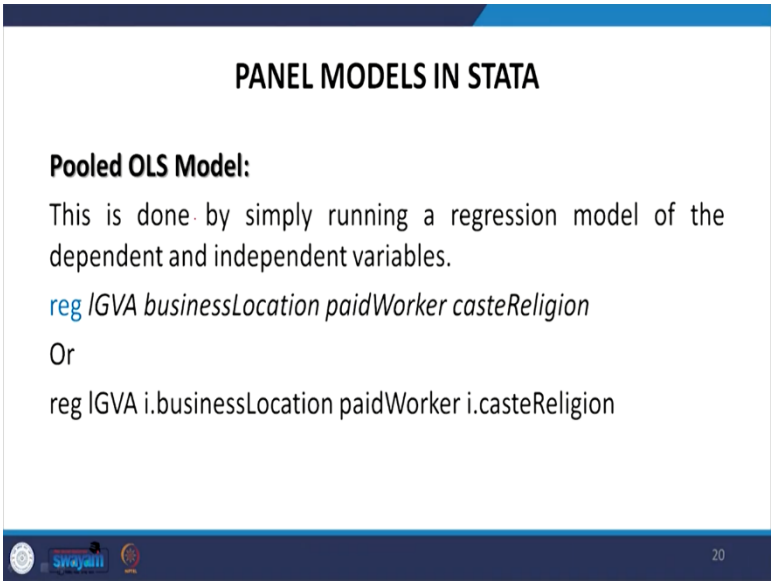
Within is given concerning the time period, regardless the number of households. So, in the above result, gross receipt is more between households than it changes over time for the same period, gross receipt which is shown here. between is much higher. Between basically across the household, the gross receipt is much higher as compared to the within. So, the N is total number of observation, while small n is observation on each entity. While understanding small n and capital N is giving, the right hand side and within variation of id variable is 0 and since id does not vary over time. Id remains same over time. So, the variation within id and their variation is 0. Their variation basically the standard deviation is 0. Coming to the survey variable has between variation 0 as time does not vary over time, over id.

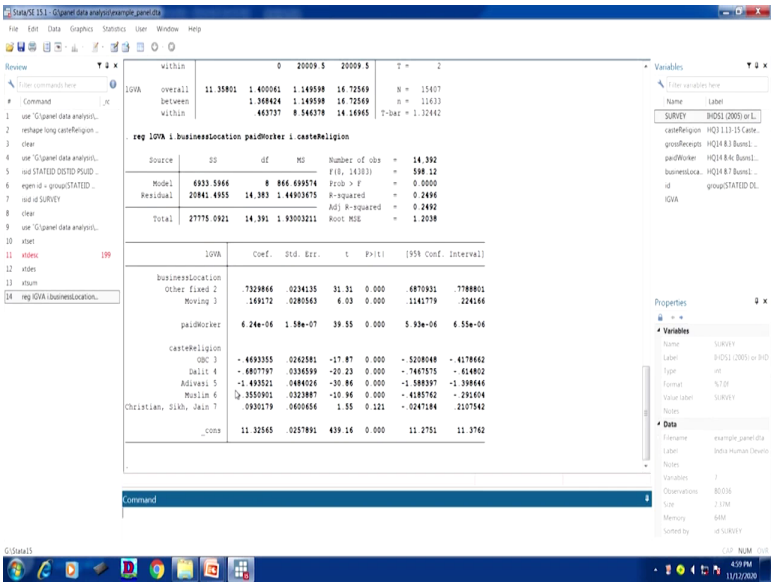
(Refer Slide Time: 22:18)



In all our variables between variation is more than the within variation. You please mark everywhere that we have already mentioned. This means these variables vary between households. Between households there are much variation, but the same household over time the variation is compared to lesser. So, this is one of the findings of the data we have so far summarized.

(Refer Slide Time: 22:51)

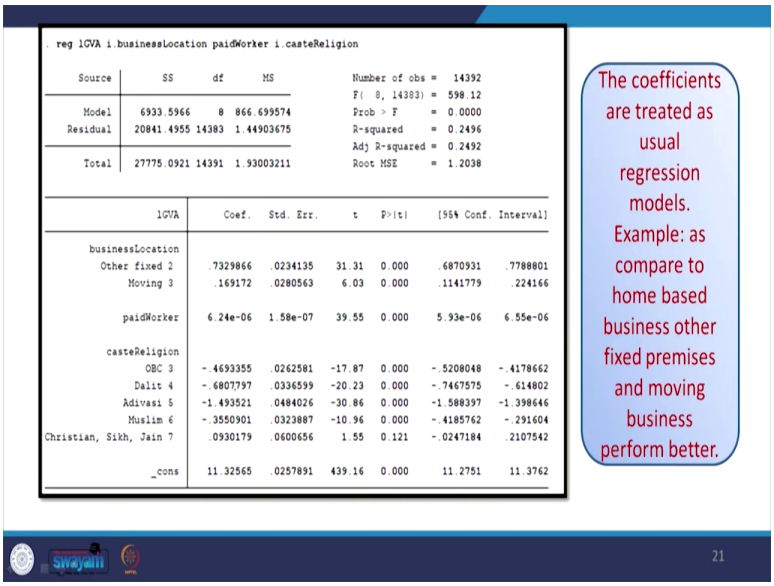




Coming to the panel models in Stata, we are now going to discuss three important aspects of analysis. One is through pooled panel, pooled OLS model that is simply the appending approach we discussed, but we are, in this panel only will simply give the ordinary regression technique. So, we will mention here and we will also operate it and we will take this command and that will be derived. We will operate and it is here.

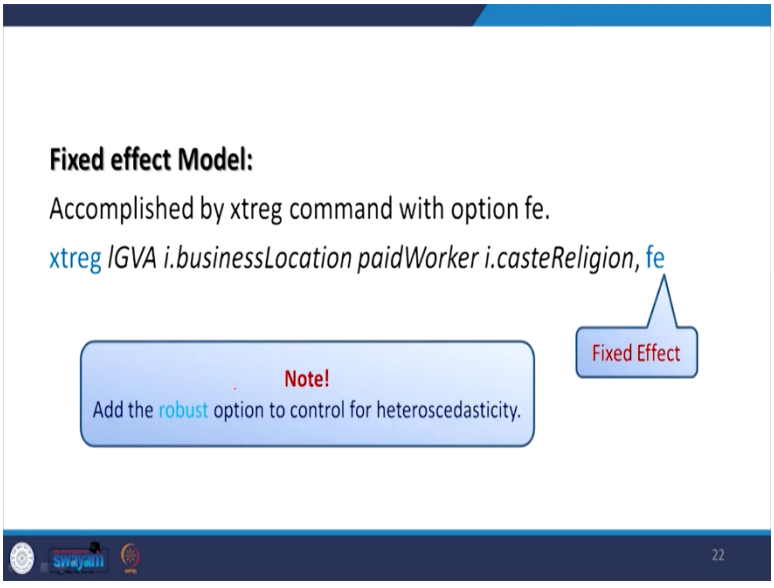
So the result is there. This is a simple OLS regression result, but our data is pooled type. It is in panel format, but it has only considered a pooled format information by considering entire observations. So, it has given the result accordingly.

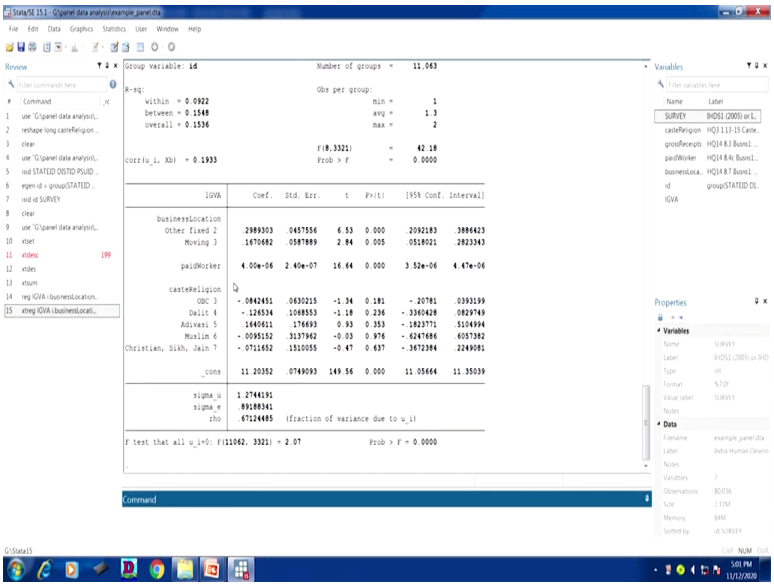
(Refer Slide Time: 23:51)



We will compare this. We will discuss this. The coefficients are treated as usual regression model. Example, as compared to home based business other fixed premises and moving business perform better that we discussed earlier.

(Refer Slide Time: 24:10)

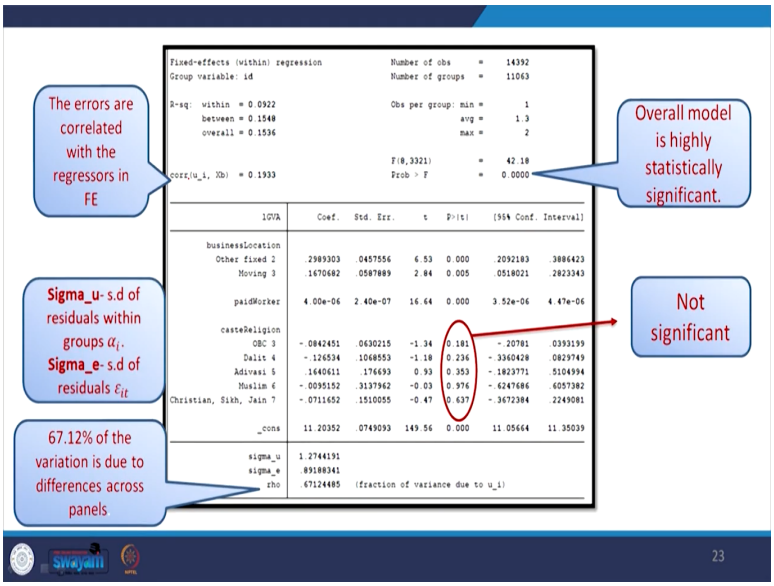




Coming to the fixed effect model. So, the fixed effect and random effect model and which one is going to be more suitable in this panel data is going to be discussed right now. Interestingly, while entering the syntax or the command, we have to specify whether it is fixed effect with fe. And if you also wanted to understand the heteroscedasticity information or the robust information, robust an option to be added here. But let us compare, understand the fixed effect regression model through xtreg with this. Let me mention that, it is here and we will do that mentioned it. And this is all fixed effect model.

The result is in front of you. And the correlation by assumption we know that the correlation between the error term and the independent variable or the explanatory variables are non-zero. Basically in the random effect model we have mentioned that there occurs a stochastic relationship between the random, the error term and the explanatory variables. But the assumption here in the correlation, there exist certain correlation between these two. So, that is the reason why the correlation value here is having some positive number.

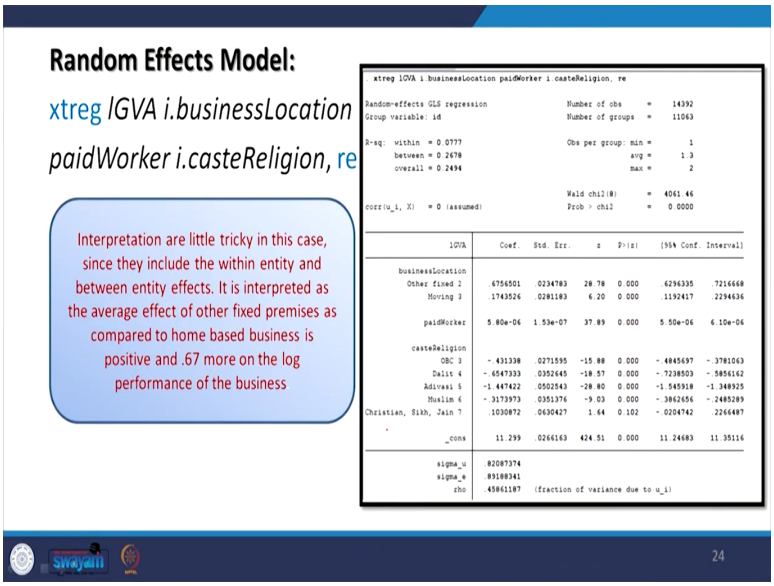
(Refer Slide Time: 25:48)

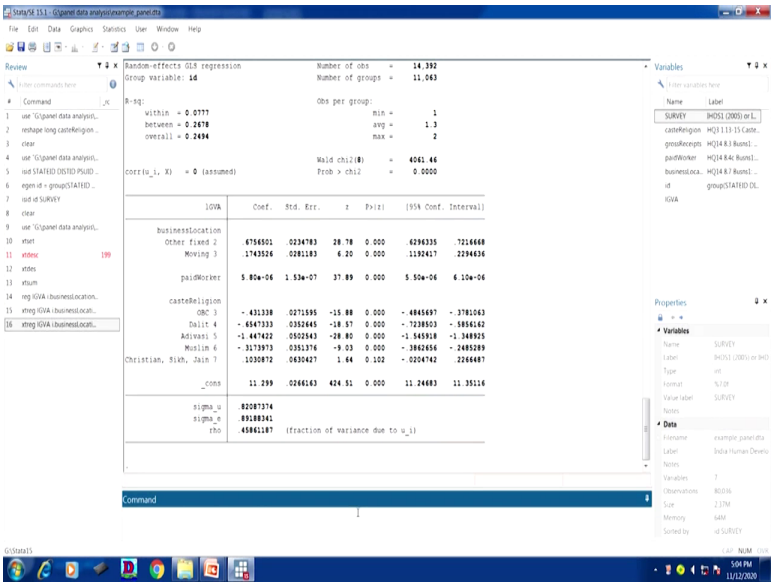


So, coming to the errors, are correlated with the regressors in fixed effect model. In case of random, by assumption, it will be 0 and we will show it. Here other important aspect that the overall model is statistically significant as mentioned with this and coming to the sigma u and sigma e as we already pointed out in our theoretical portion of panel data that sigma u is the standard deviation of residual within the group of alpha i the constant effect, the fixed effect without the time component. And it is, the deviation, the sigma value is given. And sigma e is having the standard deviation of the residuals with the time component. That is epsilon. Both the time component is there. So, the value is there.

The rho gives the value of 67.12 percent that explain the variation is due to differences across panel, the difference is of 67.12 percent. And there are some indicators we are highlighting here having non-significant. So, as per our model, since we did not specify the model so systematically we have simply shown the results. So, many variables are not coming out to be significant.

(Refer Slide Time: 27:21)



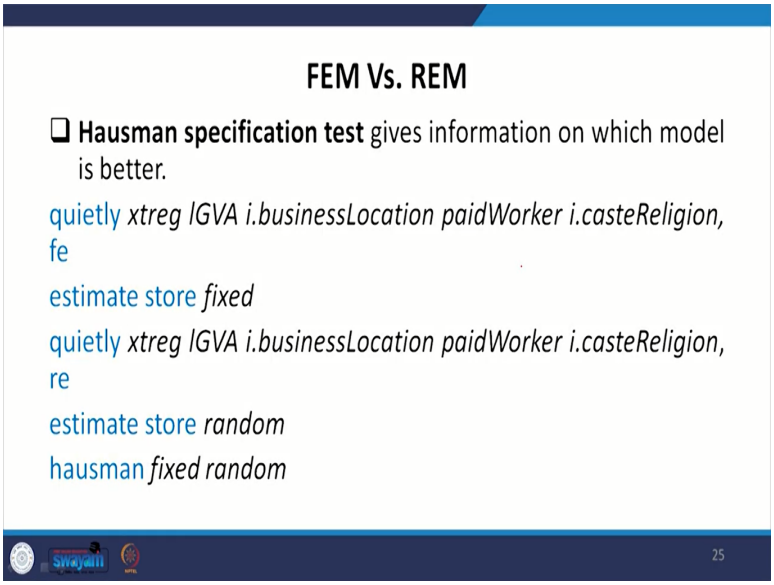


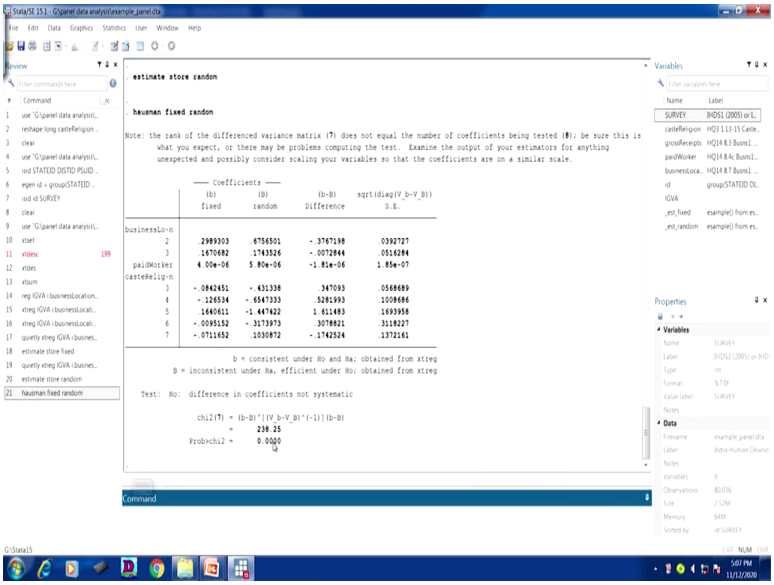
Coming to the random effect model, we will only add re there to our model. Random effect, we simply changed to re instead of fe. It gives the random effect results. As I already mentioned from the beginning that the correlation by assumption is 0. There occurs a stochasticity between these two variables. So, this is assumed to be 0.

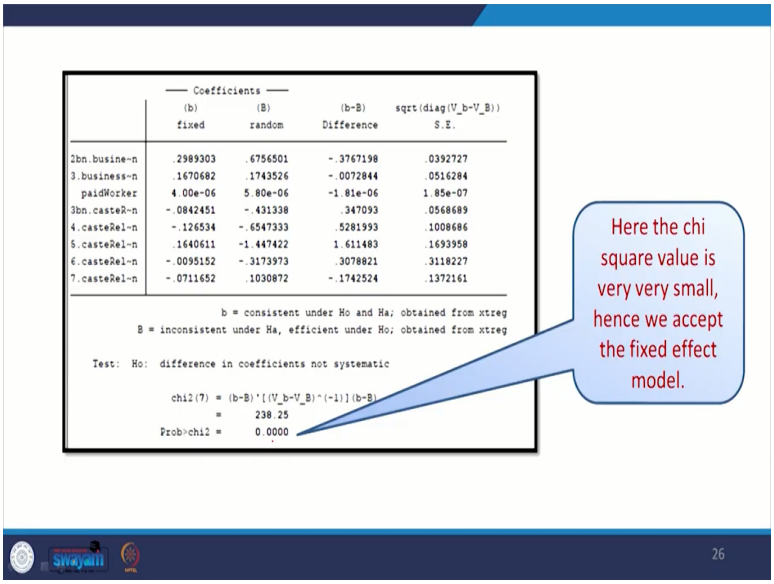
Coming to the explanation once again, the interpretations are little tricky in this case since they include within entity and between entity effect. It is interpreted as the average effect of other fixed premises as compared to the home-based business is positive and home-based and the fixed

premises is positive and 0.67, I already given in the coefficient, 0.67 more on the log performance on the business, 0.67 more so far as the fixed business as compared to home business is concerned. So, that is more important and rest details you will get it from our document.

(Refer Slide Time: 28:47)







We will come to the comparison between the random effect and the fixed effect model. As we repeatedly mentioned earlier that we require a Hausman specification test that really identifies which model to be better. We will compare these the way we did it earlier. We will compare also the same thing. We will quietly as the command we stated earlier that we will take the quietly command every time the way we taken quietly. Then we will also estimate its store value and store it. Basically, we will estimate and store it in every cases. Every time we will go by that. Then our results will be derived very quickly.

So, we derived fixed effect. We derived random effect. Then we also check a Hausman test. So, all the results are derived in front of you. After estimating all those detail, the comparison through the Hausman test is very clearly mentioned. As I told you, fixed effect and its coefficients are mentioned, random effect and its coefficients are mentioned, as per our Hausman effect if you remember on the 36th lecture we discussed that why Hausman is required and what is the formula for it.

We take the chi square estimation of these two differences. So, we take the difference and divide it with their standard deviation. There square root of these variation. So, these variations are defined and defined to be significant. So, this significance really statistically significantly violating our null hypothesis. So, null hypothesis is that the difference in coefficients are systematic. So, there are not systematic, sorry. The difference in coefficient, these two coefficients are non-systematic.

So, they are very random, but we are rejecting that null hypothesis based on the significance level. Our significance level confirms that our model is statistically significant and we are rejecting a null hypothesis. So, that means, the difference in coefficients are systematic. So, there is a very clear understanding based in the model. And based on that, we say that when they are very systematic that means the correlation between the error term and the explanatory variables are there and they are not random.

So, the assumption of the random effect is violated. So, this model, Hausman test confirms to the fact that we have to apply the fixed effect model. So, that is what we have derived. And this test confirms to the fact that the fixed effect model is more appropriate.

So, that is all about the discussion in the entire lectures of the 40 modules. This is, in fact, our 40th module and we have discussed everything about almost all basics of panel model. There are many details of panel models also. I think we have given a very good base for all of you to understand the unit level data and their structure of making a panel and how to analyse panel with the basic guidance have already been given. So, with this, I think it is the time to conclude and mention that we have systematically proceeded from the unit level data and the basics of Stata, then conceptualizing Stata, we codify Stata.

Then we enter with various components of panel and we also analyse panel. I hope you guys enjoy in between and any kind of doubts persist please do not hesitate to raise it and we will be very happy to address it. With this, I am closing the lecture and expecting your better participation and performance in the exam. Thank you.