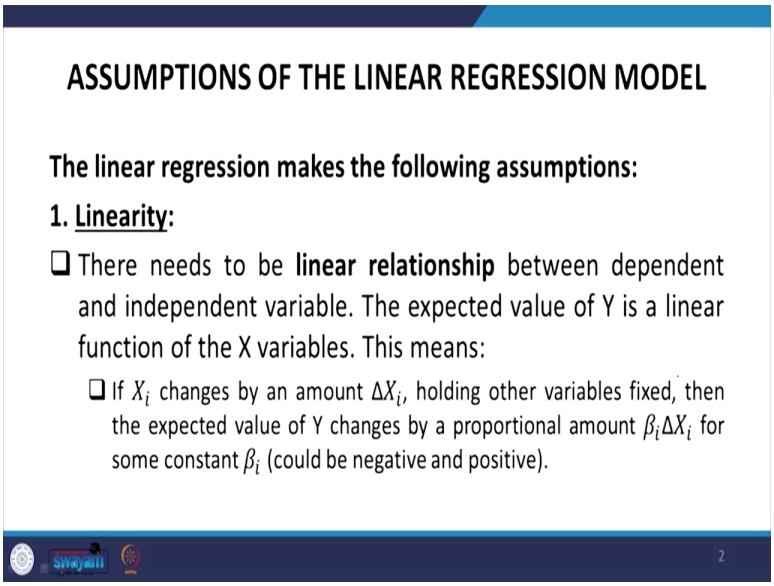
Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology Roorkee
Lecture 30
Linear Regression Analysis in STATA-III

Welcome friends once again to the NPTEL MOOC module on Handling Large-Scale Data Using STATA and large-scale we mean we are trying to handle the large scale unit level data which our students, our social science, management students largely use unit level data for their research.

So, this course is accordingly meant and in last two lectures particularly I tried to explain that linear regression issues and on the very third lecture, basically in the previous two lectures we discussed the prerequisites of the linear regression and where you must be very cautious enough and today as I promised in the last lecture that we are going to have at least for half an hour to 45 minutes on the commands or the STATA operations of linear regressions, pre as well as post estimations of this particular analysis.

(Refer Slide Time: 01:49)

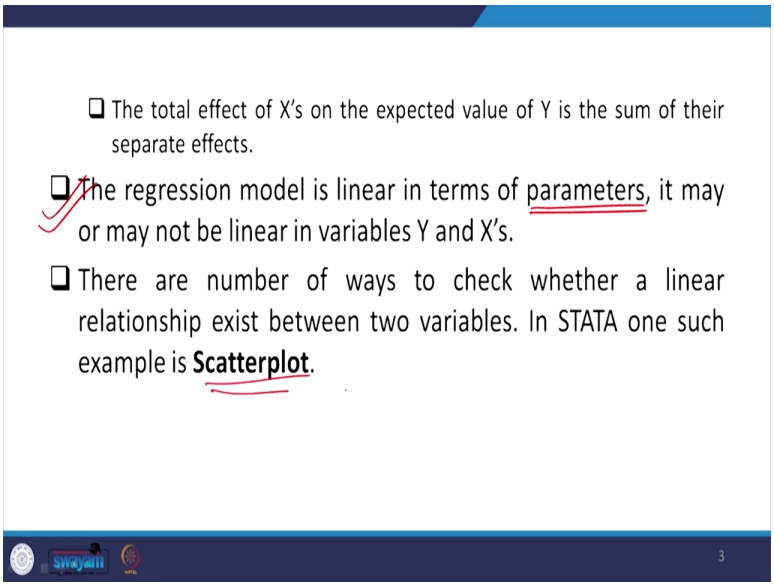


So, let us start. To begin the linear regression model, we are supposed to understand the very basic assumptions of the model. Some assumptions start with linearity, that is the very most and beginning assumption of the linear regression model. Even almost all the models in econometrics assumes linearity or after some approximation they linearize the model and then interpretation

gets easy. So, we need to understand that, it is the need of the model to have a linear relationship between dependent and independent variables.

The expected value of Y is a linear function of the X variable, which we wanted to say. That means that if X_i changes by an amount of unit change of delta X, holding other variables fixed, then the expected value of Y changes by a proportional amount that will be with the coefficient called beta i delta X_i for some constant beta i that could be negative or positive depending upon our relationship of the variables and in the model whatever we are going to get it.

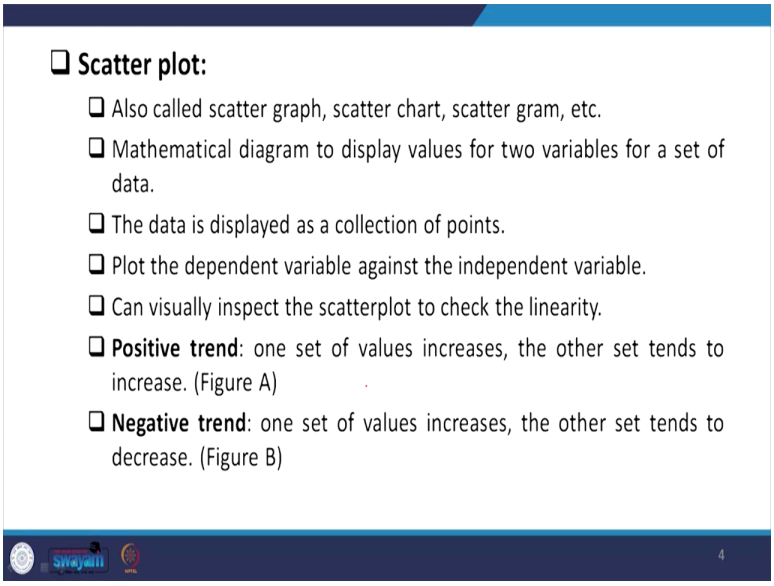
(Refer Slide Time: 03:04)



So, the total effect of X on the expected value of Y is the sum of their separate effects, so then only the total effect can be calculated. The regression model is linear in terms of parameters that is more important and I mentioned several times. So, that is linear in parameter, but not necessarily for its variables. So, this is one of the very very important assumption you must be very careful and try to understand the logic that we explained earlier.

There are number of ways to check whether a linear relationship exists between two variables. In STATA, one such example is through Scatterplot. So, we will try to establish this and through understanding whether there exist linear relationship or not. We will clarify what we mean by linear relationship.

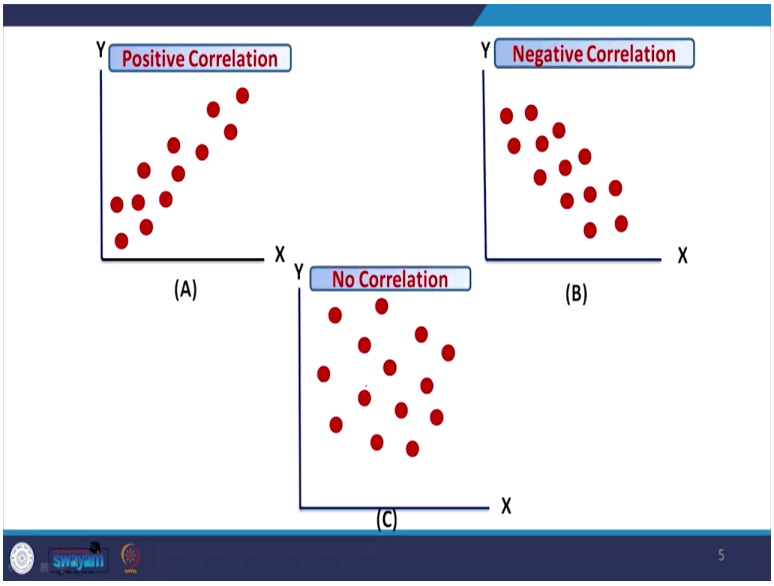
(Refer Slide Time: 03:55)



So, the scatter plots give us scatter chart, scatter graphs, scatter grams. These are also called like graphs, scatter charts or grams, etc. The mathematical diagram to display values for two variables for a set of data. The data is displayed as a collection of points. Different points will be plotted and that is being displayed with the scatter plot.

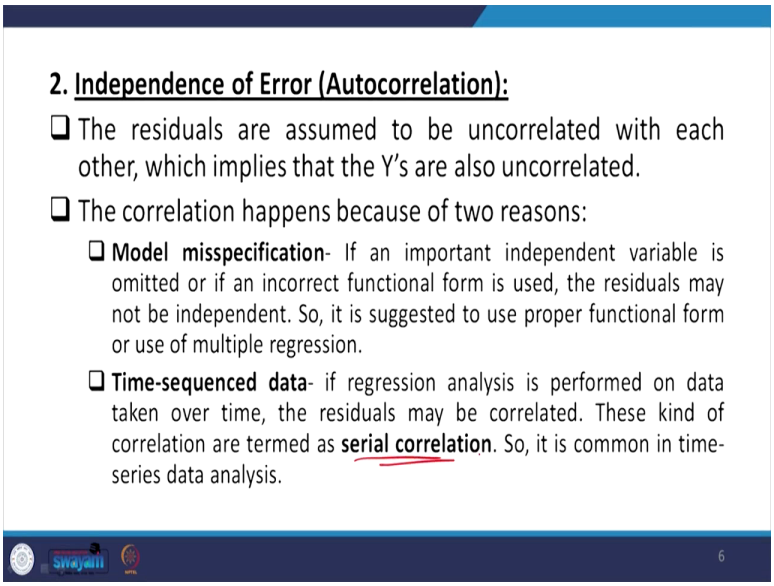
It plots the dependent variable against the independent variable. So, this can visually inspect the scatter plot to check the linearity. So, positive trend means there exists a positive relationship which is clarified as 1 set of values increases the other set also increases, whereas vice versa is valid.

(Refer Slide Time: 05:00)



These are explained in figure A and B. A refers to increasing trend and B refers to negative trend, whereas in C, there is no trend. So, accordingly we can define that A explains positive correlation and B negative and there is spurious correlation in case of C.

(Refer Slide Time: 05:14)



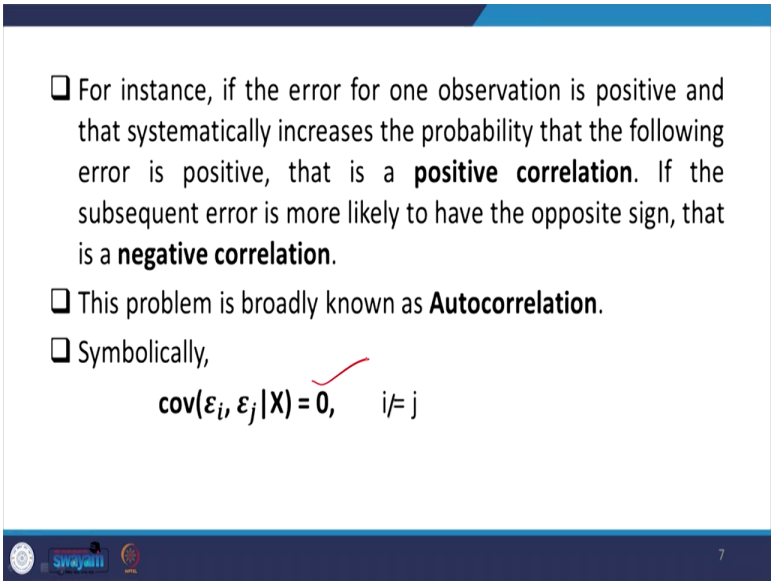
Coming to other backdrop of understanding the linear regression and its applications. The residuals are assumed to be; we are now going by another assumption called independence of error. So, independence of error also referred to as autocorrelation. Let us clarify autocorrelation

that the residuals function is mainly dealt. The error functions are assumed to be uncorrelated with each other, like errors function of the period T or T minus 1 should not be correlated, which implies that your Y_i are also uncorrelated.

The correlation happens because of some reasons. Mainly two reasons we are referring. One is related to model misspecification or time-sequenced data, which I just said. If an important independent variable is either omitted or of any incorrect functional form, the residuals may not be independent, if by any chance there are incorrect functions taken. So, it is suggested to use proper functional form or use a multiple regression.

In case of time-sequenced data, if regression analysis is performed on data taken over time, the residuals may be correlated. Residuals over time correlated because generally they go by average time and that average trend keeps on continuing over time. So, there occurs correlation. These kind of correlations are termed as serial correlation also. When time series component are there and it is common in time series data as I just mentioned.

(Refer Slide Time: 07:02)

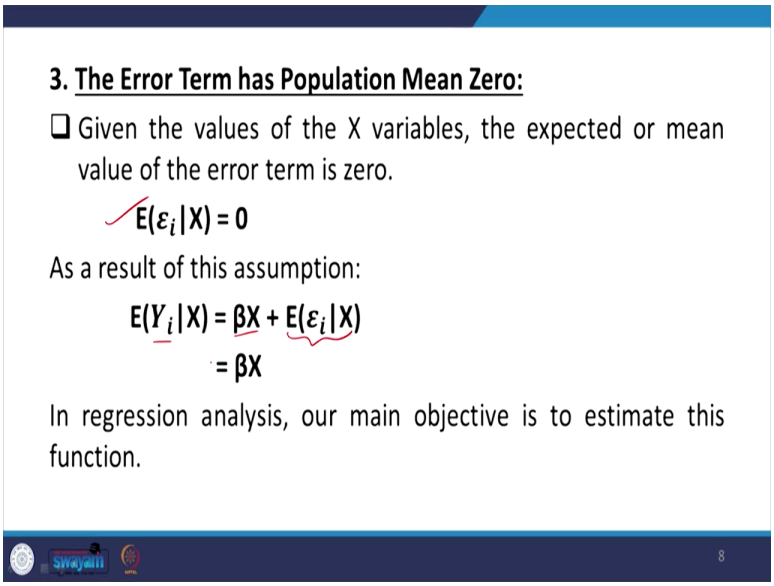


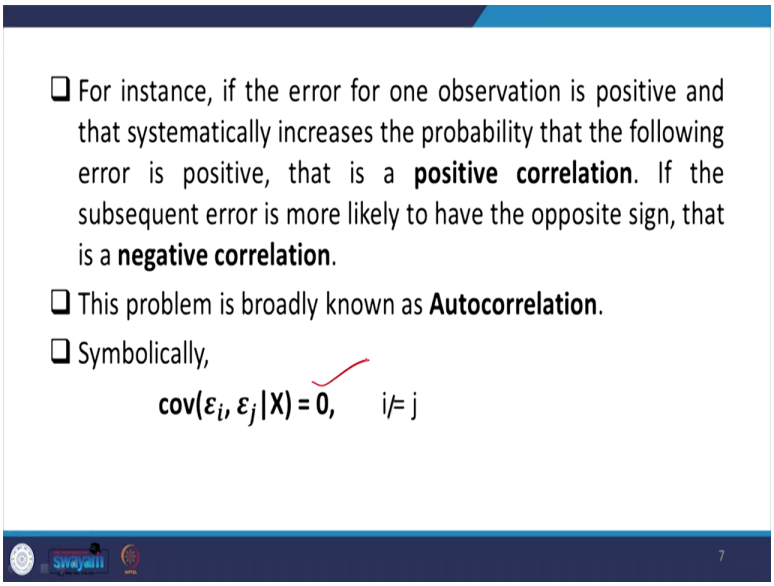
For instance, there exist positive, for instance if the error for one observation is positive and that systematically increases the probability that the following error is positive that is a positive correlation. If the subsequent error is more likely to have the opposite sign that means it is

negative correlation. The problem is broadly known as autocorrelation, which we have just clarified.

Now, symbolically we present that covariance, basically correlation between two error term, the covariance between e_i and epsilon i or population error in case of population data otherwise small e_i and e_j given the X, given conditional upon the independent variable from where the error term is generated, the covariance must be equal to 0. So, i and j are not the same. It should be a different series of data.

(Refer Slide Time: 08:07)

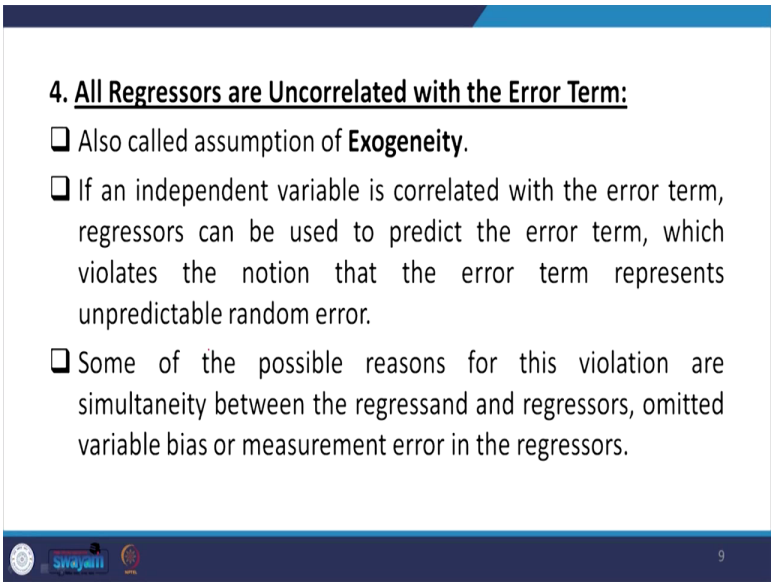




So, the error term has a population mean 0 that is given, the third option is more important as well as the average error, the average mean error should be 0 and there should not be autocorrelation. So, the covariance should be equal to 0. Then, we are saying the estimated value of the error term given X would be also 0, which is explained by here expected value of e_i is equal to 0.

As a result of these assumptions, when the expected error term is 0, if you take the dependent in the standard regression equation where Y_i is equal to beta X_i plus alpha i or beta X_i plus e_i, given conditional upon X if I take it and then I take the expected value, these component boils down to 0, so that means the expected value of this is nothing but the beta of X. So, expected value of the dependent variable is nothing but the beta of X and accordingly we can estimate the function. The beta value can be estimated.

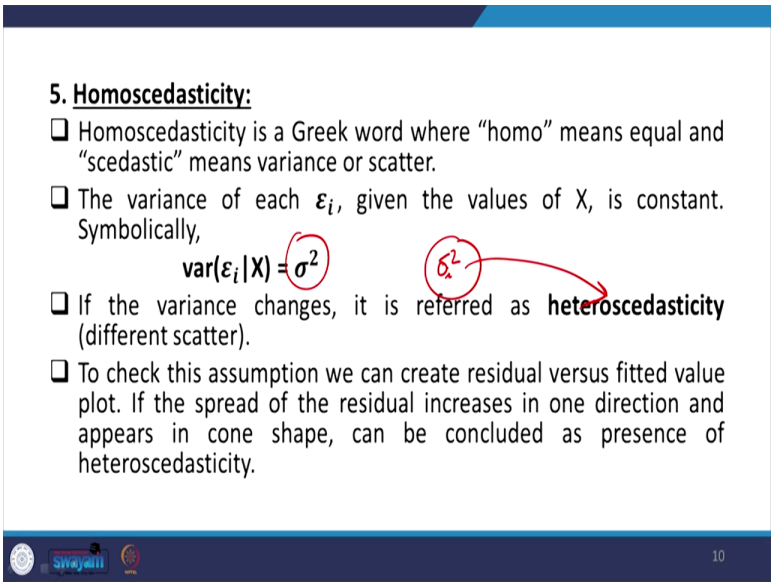
(Refer Slide Time: 09:20)



All the regressors are uncorrelated with the error term that is another one, assumption as required for the linear regression to be followed. All the regressors the explanatory variables should be also uncorrelated. So, the assumption is called erogeneity. If an independent variable is correlated with the error term, basically says that regressors are uncorrelated with the error term and there are some difference to it as well, uncorrelated with each other as well. We are going to discuss that and with error term. Basically, the correlation between the X_i with the epsilon should equal to 0.

If an independent variable is correlated with the error term, regressors can be used to predict the error term which violates the notion that the error term represents unpredictable random errors. So, it misled the prediction. Some of the possible reasons for this violation are simultaneity between the regressand and regressors omitted bias, omitted variable bias or measurement error in the regressions that result in error term and the explanatory variables and their relationship. So, the assumption of exogeneity is also important.

(Refer Slide Time: 10:42)



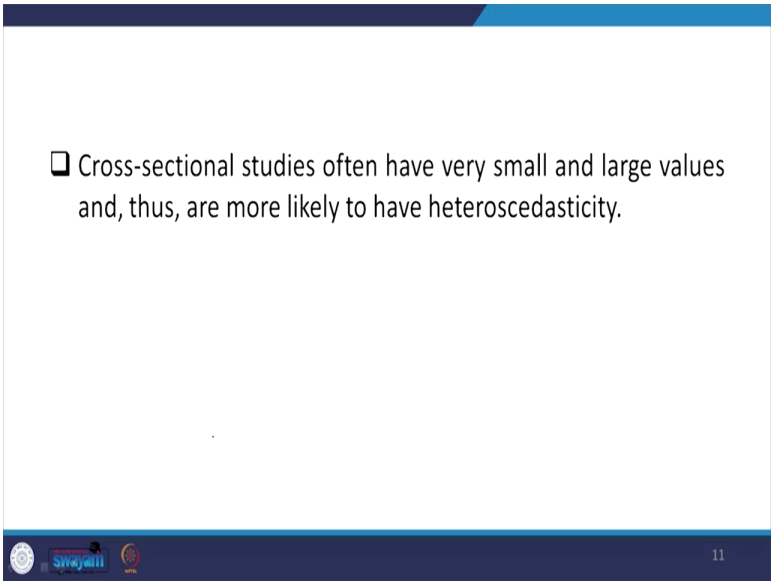
Coming to the issues of variance of the distribution error term, we are now going to emphasize the another assumption called homoscedasticity. The meaning of homoscedasticity derived from the Greek word with homo refers to equal and scedastic refers to variance or scatter.

The variance of each error term that is epsilon i given the value of X is constant. So, that means it is expected that the homogeneity or the homoscedasticity ensures that the variance of their items should be constant, should not be varying every time. If it is varying, then it is difficult to predict.

If the variance changes, if that is not sigma i, rather if it is equal to sigma i square then this is called heteroscedasticity. So, every error term correspond to a, and its variance correspond to a different standard deviation or variance led to the problems of heteroscedasticity and so , a different scatter information can be derived.

To take this assumption we create residual versus fitted value plot. If the spread of that residual increases in one direction and appears in cone shape can be concluded as presence of having heteroscedasticity. So, either it is rising in a cone shape that means is constantly changing or changing in certain direction that is called problems of heteroscedasticity.

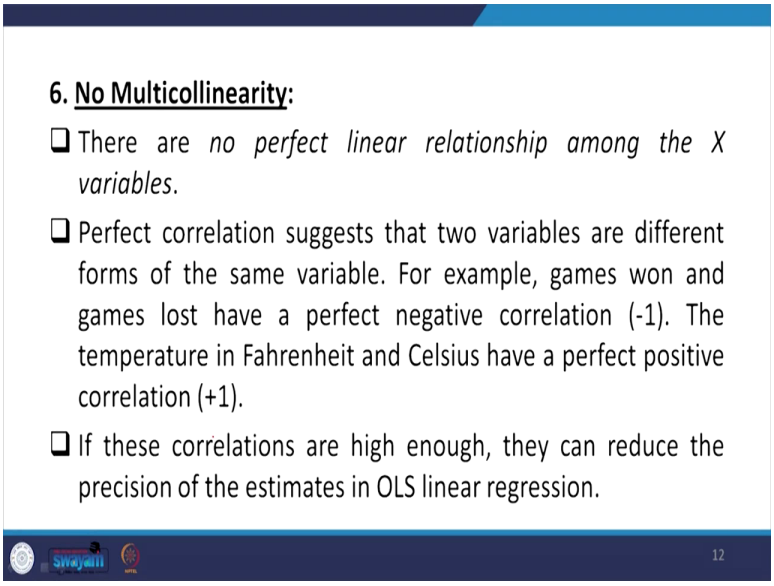
(Refer Slide Time: 12:47)



So, usually cross-sectional data or studies often have either very small or large values, so that amounts to heteroscedasticity. So, it has higher likelihood of having heteroscedasticity, because either some responses are having very small value or having very large values. The deviance between them is very high. So, heteroscedasticity is very likely.

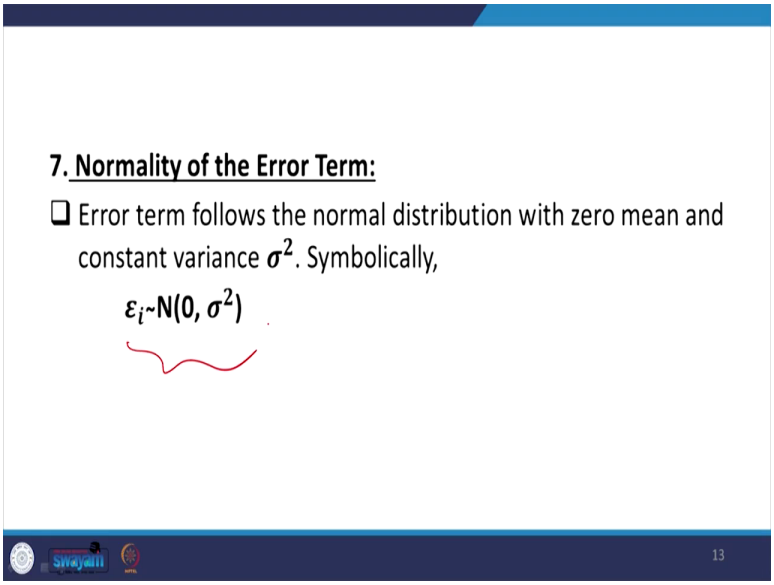
The sixth most important assumption is multicollinearity. There should be no multicollinearity assumption. So, in the previous one we said there should be homoscedasticity, no heteroscedasticity.

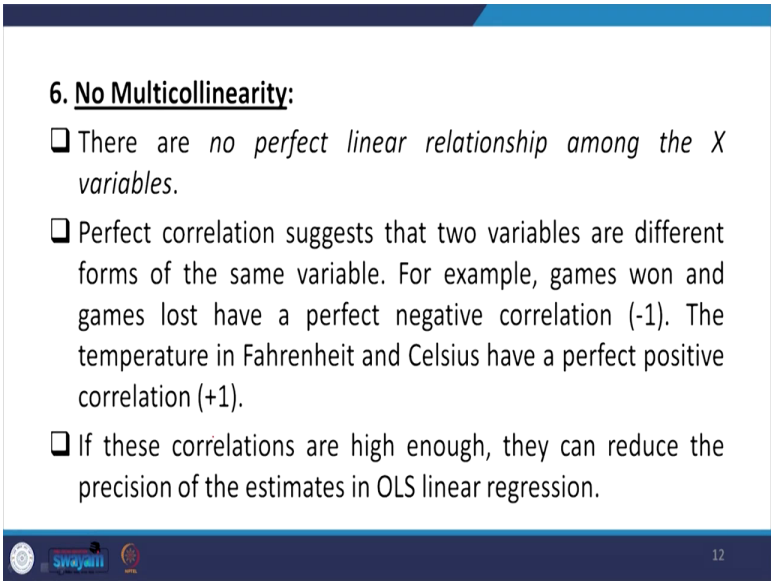
(Refer Slide Time: 13:34)



In case of multicollinearity we wanted to say that there are no perfect linear relationship among the variables or independent variables. So, perfect correlation suggests that two variables are different forms of the same variable. For example, games won or games lost have a perfect negative correlation with minus 1. The temperature, similarly, in terms of Fahrenheit or Celsius have a perfect positive correlation that is plus 1. If these correlations are high enough, then they can reduce the precision of the estimates of OLS regression, OLS linear regression estimations.

(Refer Slide Time: 14:16)

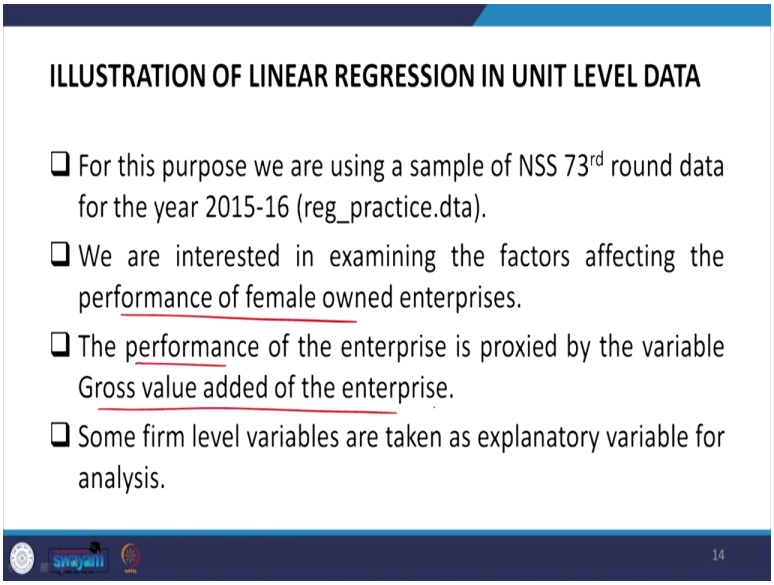




Coming to the seven most important assumption is normality of the error term. We are going to explain you what is multicollinearity in the data, what is heteroscedasticity in the data we are going to explain with the help of experiment with the real example or the data. So, next assumption is normality of the error term. The error term must follow a normal distribution. So, normal distribution having 0 mean and constant variance. We already assumed that variance must be constant.

And also we have assumed that the estimated value of the error term should be equal to 0. So, that way again it validates when both are followed so that means we will be eventually following a standard normal error distribution and that is explained with the help of this epsilon i distribution follows normality with 0 mean and variance sigma square.

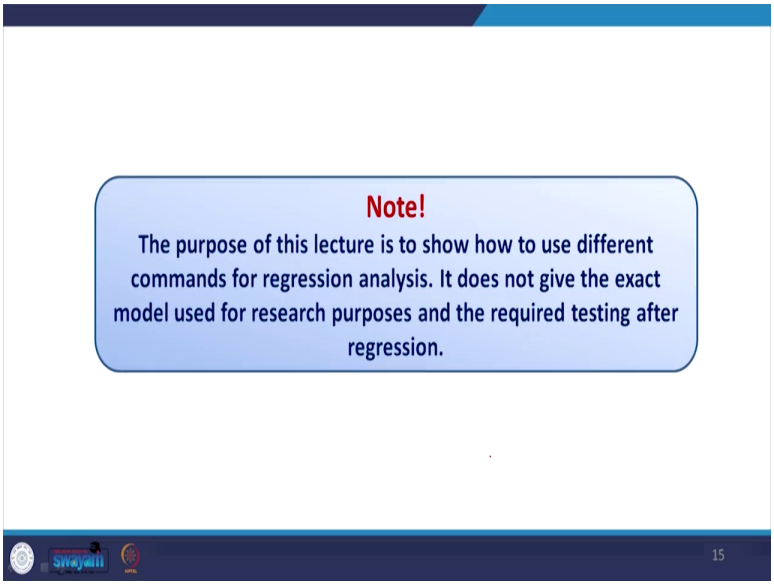
(Refer Slide Time: 15:22)



So, let us illustrate the linear regression in unit level data. Using the unit level data we are going to explain how we follow the important assumptions. So, for this purpose we are using 73rd data of National Sample Survey that was conducted in 2015-16. So, we are going to use a sample data for us that is with the name we are going to provide it for your practice that is with the name regression underscore practice data.

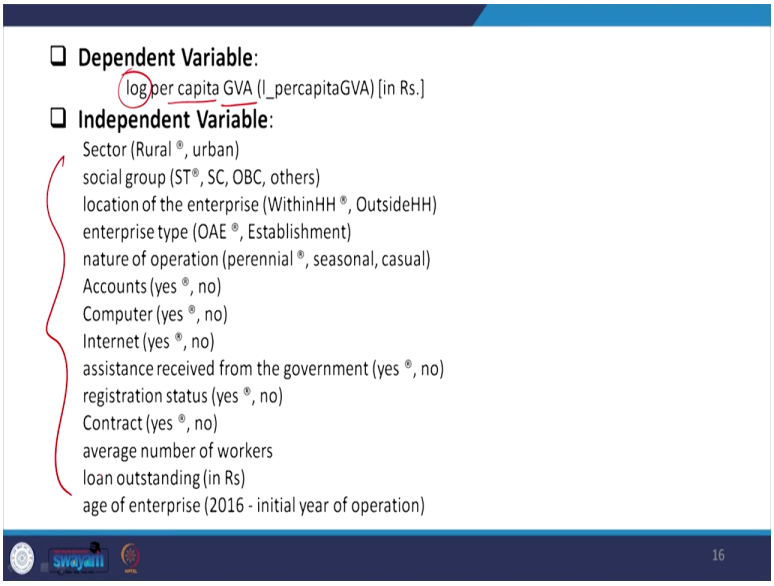
So, we are interested in examining the factors affecting the performance of female-owned enterprises. The performance of the enterprise is proxied by, here we are going to understand the performance of the female enterprises and that will be proxied by the gross value added that information is given in the data. So, there are some firm level variables are taken as explanatory variables and those will be used.

(Refer Slide Time: 16:25)



Please take a note; the purpose of this lecture is to somehow use different commands for regression analysis. It does not give the exact model used for research purposes and the required testing after regression. So, important aspects we are going to discuss that does not mean it gives the exact model. There are lots of experiment made in between for getting the right model and right result.

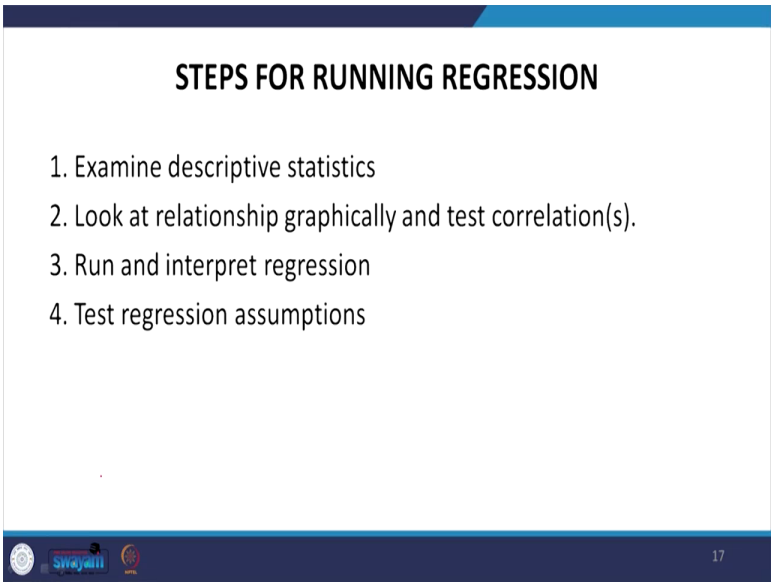
(Refer Slide Time: 16:49)



So, here are the details of our variable of dependent and independent. In the dependent, we said that we are going to use gross value added per capita terms and its log value we are going to use in order to normalize the dependent variable. I will tell you whether it is normalized or not through log transformation.

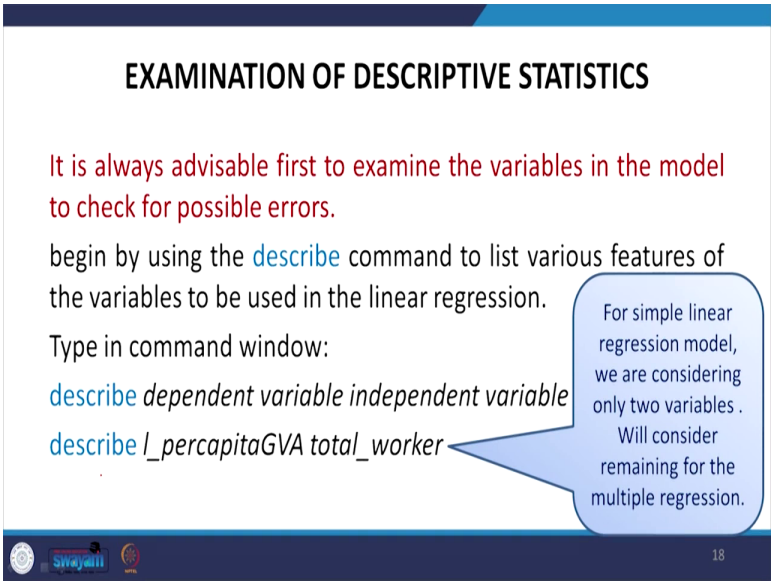
And independent variable we are going to use; sector that is rural and urban, social groups in terms of caste, then location of the enterprise, then enterprise type, own account enterprise or establishment. Similarly, we are also using computer use or loan outstanding of the enterprise. So, basically 2016 minus initial years of operation if you do it, we get that information for our explanation.

(Refer Slide Time: 17:47)



So, what are the steps involved for running the regression. First of all, we need to examine certain descriptives of the variables used in the model. So, we will understand the relationship graphically and also test their correlation and then we will run and interpret regression. So, then at the last we will test those assumptions, whether those assumptions we have taken, whether those have been proved or rejected are going to be understood from the session.

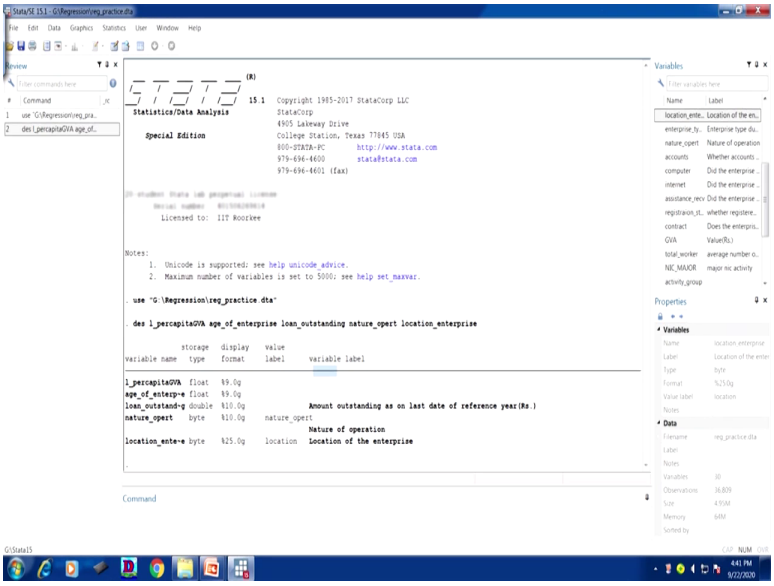
(Refer Slide Time: 18:31)

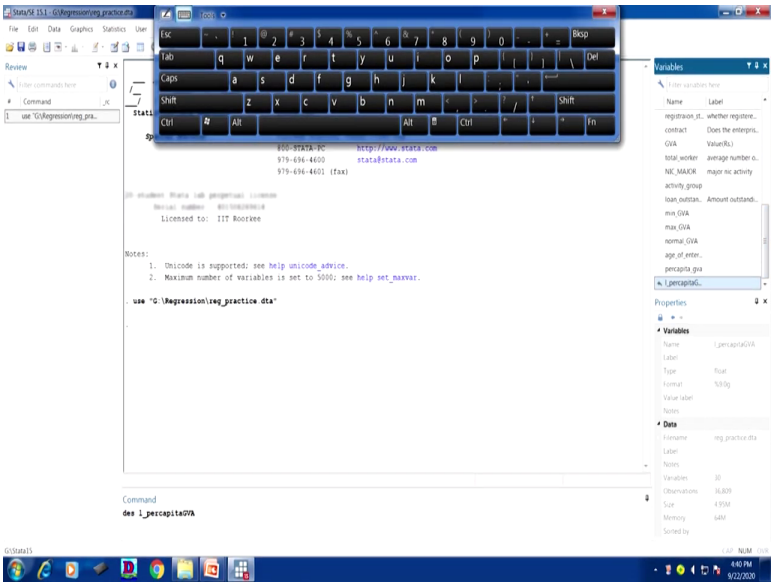


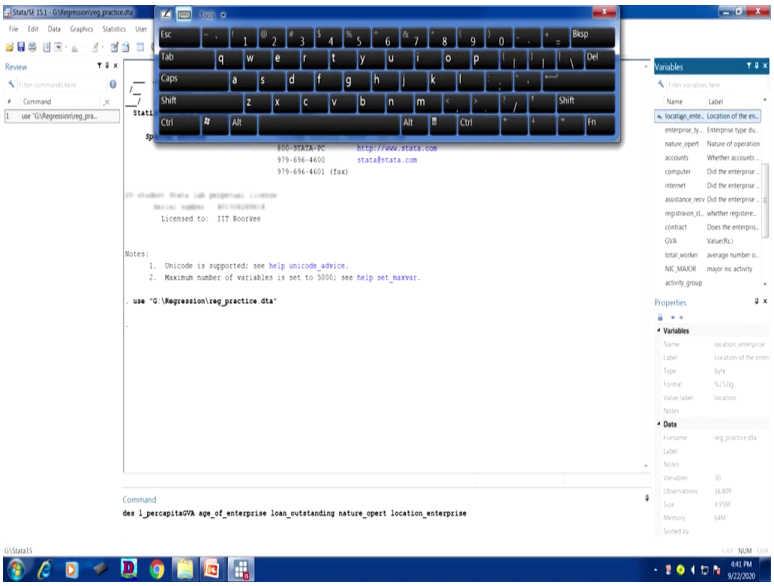
So, now we are going to examine the descriptive statistics as I just said. It is always advisable first to examine the variables in the model to check for possible errors. So, we will begin by using the describe command which we used earlier to list various features of the variables to be used in the linear regression. So we need to type in the command window as described of those variables; dependent as well as independent and that is going to give us the right result.

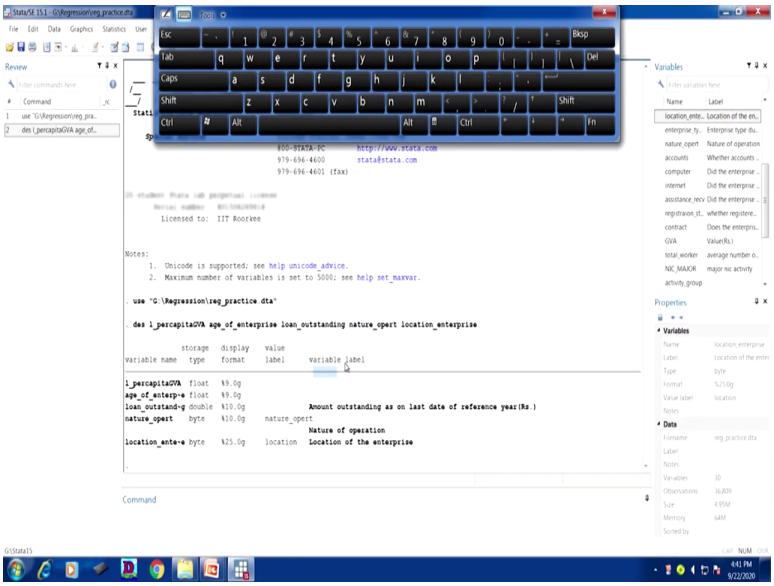
I think if you have forgotten, we are going to operate right now. For simple linear regression model we are considering only two variables. Otherwise, we will be using the multiple regression with multiple independent variables. So, I am going to switch the mode to the regression commands.

(Refer Slide Time: 19:24)









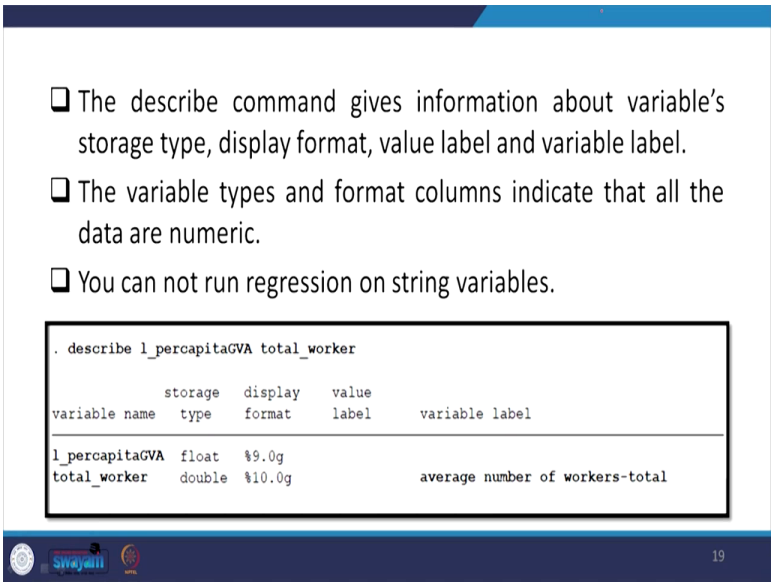
So, here is the window of the STATA and we are going to use the data which I said already. So our data is here. So we are going to use the practice data. This has already been opened. So what we will do, we will go by the variables which we have already shown to you. So, the log of the GVA we have already converted and I think it is at the bottom. So this is there and any sample variable we can use for our clarification or simply describe everything together you will also get except some other variables not interested for our explanation we should not do it.

For simplicity, I am simply copying the variable, otherwise if I type describe only d, so des then if you just click here this and any independent variable, this will give us some description, like

loan outstanding, like nature of operation, location etc. If I simply enter, it gives me the description of those variables.

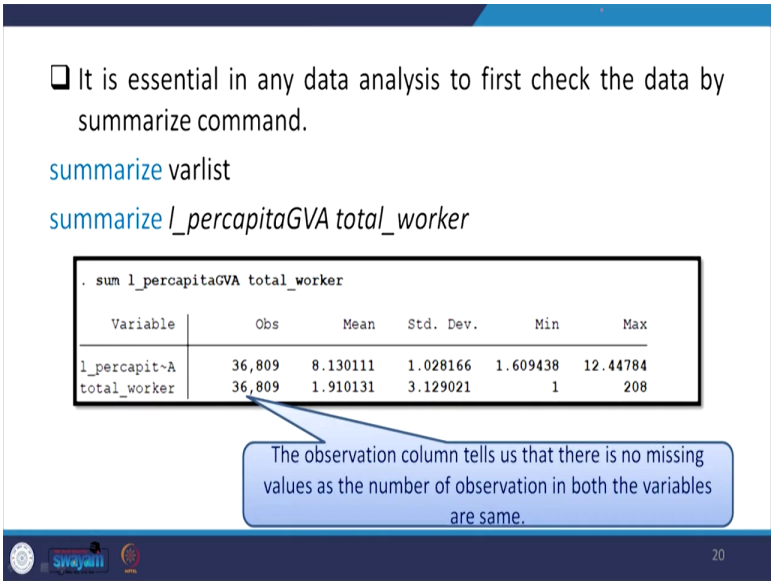
This also emphasizes whether the variable is in string format or in numeric format, those aspects we have already clarified earlier. If you have some difficulties, please follow our previous lectures for better clarification. It also gives variable labeling. So, variable labeling are also emphasized or also derived. Without wasting much time these are very simple and I am quite sure you can able to do it. So, let me proceed.

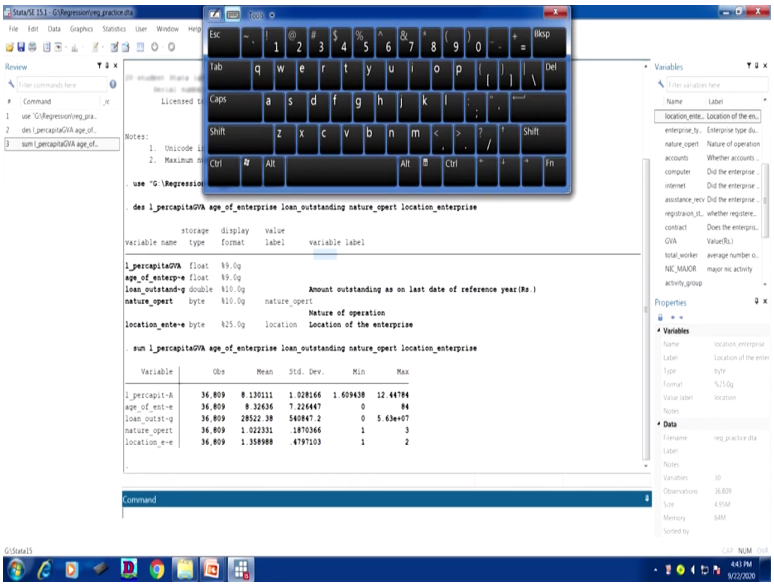
(Refer Slide Time: 21:43)



If I do like; describe these 2, 1 dependent variable and 1 total worker. I will get this type of picture, which I have just shown you. The describe command gives information about variable's storage type, display format and value label and variable label. The variable types and format columns indicate that all the data are numeric that we have already shown to you. So, you cannot run regression if the variable stored with string variable, you need to destring them and then you can go further operation.

(Refer Slide Time: 22:13)





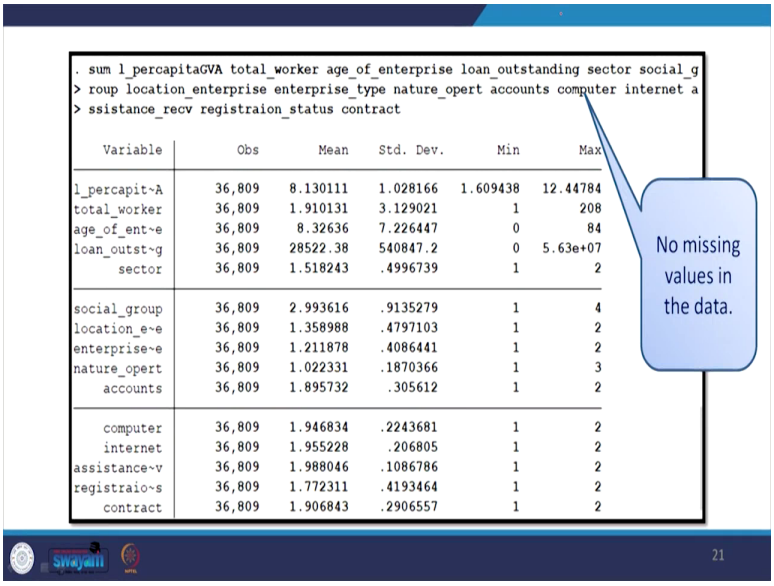
So it is essential in any data analysis to first check the data by summarize command. So, not just the describe, describe is not going to give all information. Likewise, what I did just now, if I just explain you with the same variable, just I wanted to find out that, instead of describe I will use summarize.

So, I will get certain information for sure and which are going to be useful further. In the summarize command we will be getting the observation details, the average value of it, the standard deviation, the minimum and maximum values of each of the variables which are useful

for us and usually these figures are expected everywhere in all your paper for publications. Wherever you communicate, they will certainly expect you to give the summarized table.

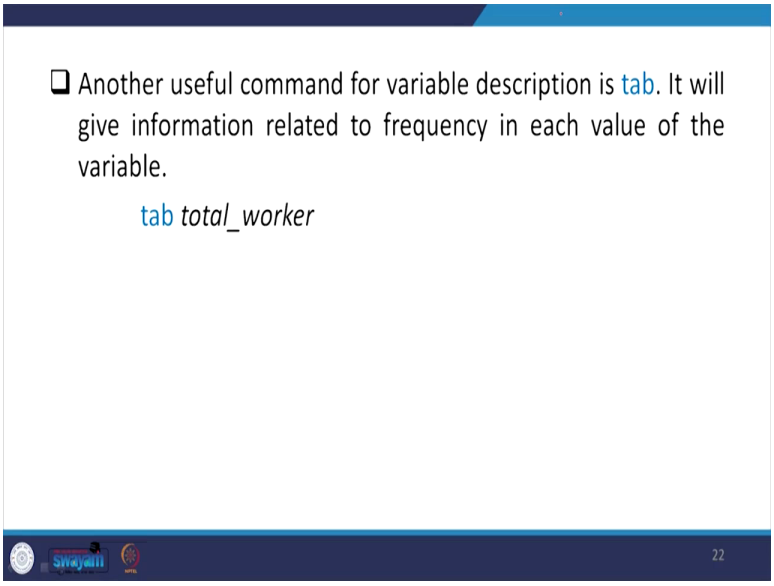
So, I am not going to derive the same result. You can experiment and find out. So, the observation column tells us that there are no missing values, because the number of values which are there in the data in this result also we do not have any missing value and how to deal with those aspects we have already mentioned earlier.

(Refer Slide Time: 23:37)



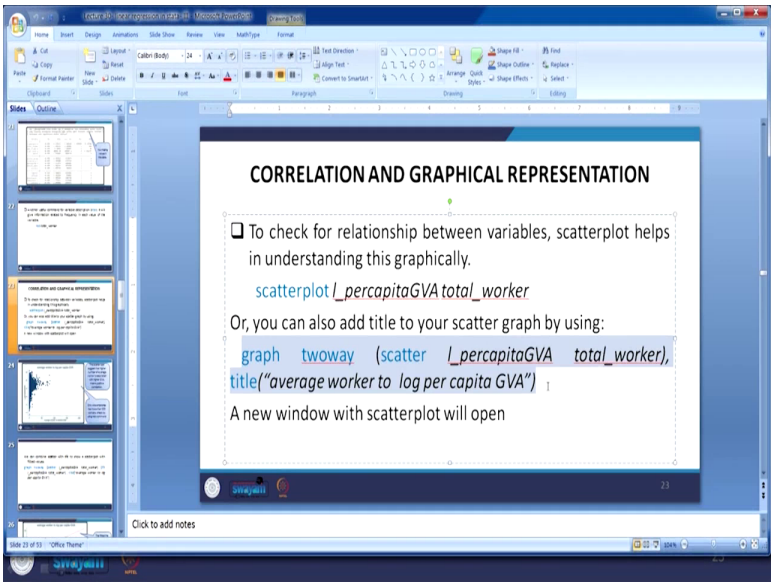
Similarly, if I just go by all the variables together, so I will get all such information and it clearly emphasizes that there is no missing value and all the observations are visible for us.

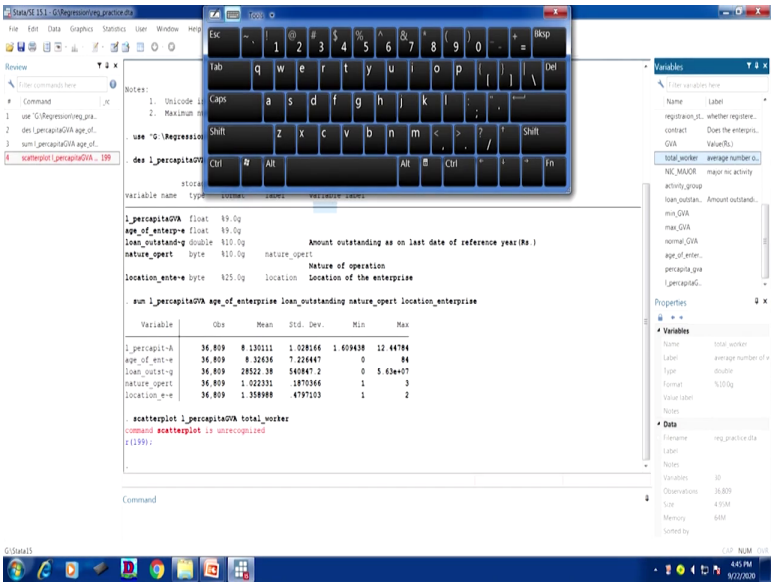
(Refer Slide Time: 23:56)

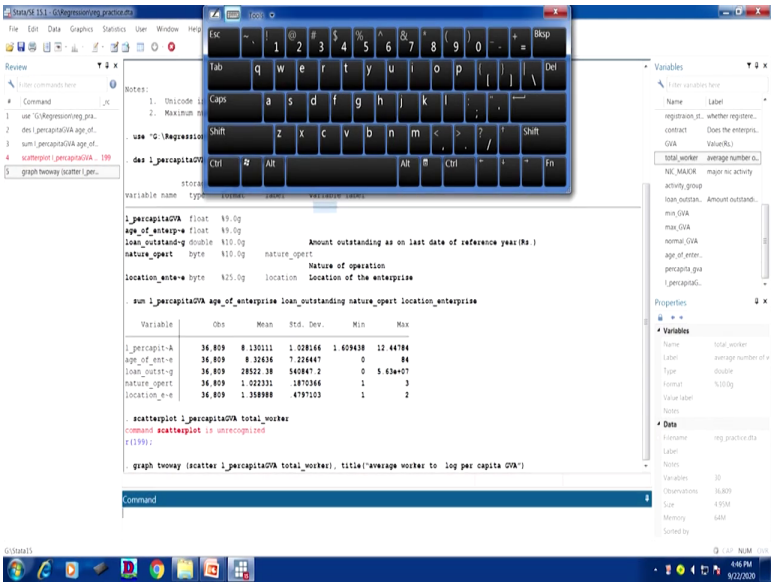


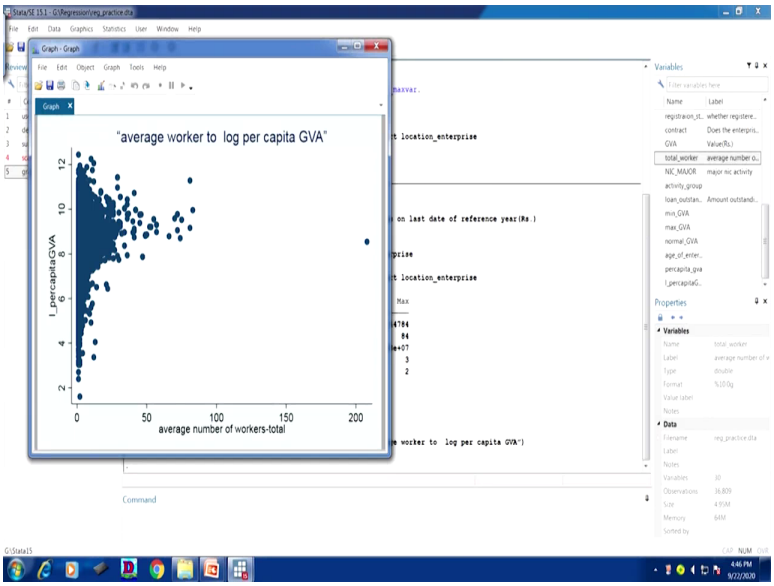
Another useful command for variable description is also tab. If I simply tap it, tab that particular variable it gives me the frequency distribution of that particular variable.

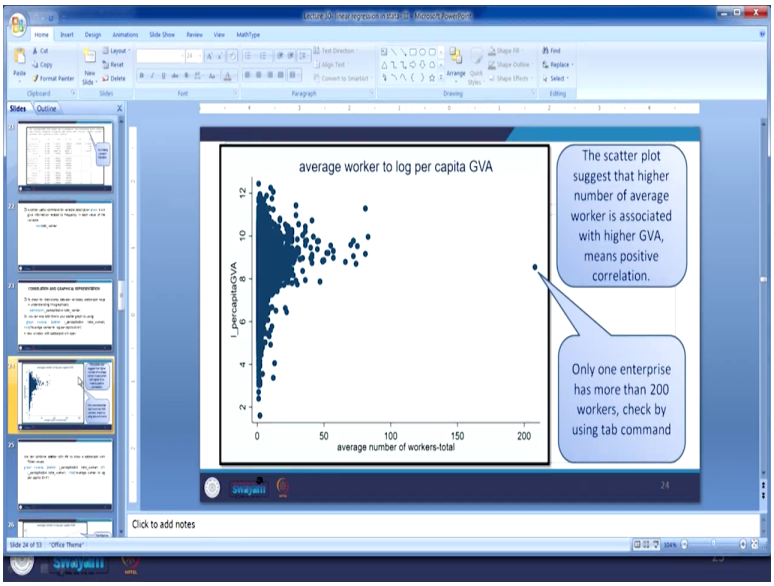
(Refer Slide Time: 24:17)




Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
1 use "G:\Regression\reg_pra...
2 des l_percapitaGVA age_of...
3 sum l_percapitaGVA age_of...
4 scatterplot l_percapitaGVA ... 199
Notes:
1. Unicode is supported
2. Maximum number of variables is set to 5000 but can be increased
. use "G:\Regression\reg_practice.dta"
. des l_percapitaGVA age_of_enterprise loan_outstanding nature_opert location_enterprise
storage display value
variable name type format label variable label
l_percapitaGVA float %9.0g
age_of_enterp~e float %9.0g
loan_outstand~g double %10.0g Amount outstanding as on last date of reference year(Rs.)
nature_opert byte %10.0g nature_opert Nature of operation
location_ente~e byte %25.0g location Location of the enterprise
. sum l_percapitaGVA age_of_enterprise loan_outstanding nature_opert location_enterprise
Variable Obs Mean Std. Dev. Min Max
l_percapit~A 36,809 8.130111 1.028166 1.609438 12.44784
age_of_ent~e 36,809 8.32636 7.226447 0 84
loan_outst~g 36,809 28522.38 540847.2 0 5.63e+07
nature_opert 36,809 1.022331 .1870366 1 3
location_e~e 36,809 1.358988 .4797103 1 2
. scatterplot l_percapitaGVA total_worker
command scatterplot is unrecognized
r(199);
Command
Variables
Properties
G:\Stata15
4:45 PM
9/22/2020


Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
1 use "G:\Regression\reg_pra...
2 des l_percapitaGVA age_of...
3 sum l_percapitaGVA age_of...
4 scatterplot l_percapitaGVA ... 199
5 graph twoway (scatter l_per...
Notes:
1. Unicode is supported
2. Maximum number of variables is set to 5000 but can be increased
. use "G:\Regression\reg_practice.dta"
. des l_percapitaGVA age_of_enterprise loan_outstanding nature_opert location_enterprise
storage display value
variable name type format label variable label
l_percapitaGVA float %9.0g
age_of_enterp~e float %9.0g
loan_outstand~g double %10.0g Amount outstanding as on last date of reference year(Rs.)
nature_opert byte %10.0g nature_opert Nature of operation
location_ente~e byte %25.0g location Location of the enterprise
. sum l_percapitaGVA age_of_enterprise loan_outstanding nature_opert location_enterprise
Variable Obs Mean Std. Dev. Min Max
l_percapit~A 36,809 8.130111 1.028166 1.609438 12.44784
age_of_ent~e 36,809 8.32636 7.226447 0 84
loan_outst~g 36,809 28522.38 540847.2 0 5.63e+07
nature_opert 36,809 1.022331 .1870366 1 3
location_e~e 36,809 1.358988 .4797103 1 2
. scatterplot l_percapitaGVA total_worker
command scatterplot is unrecognized
r(199);
. graph twoway (scatter l_percapitaGVA total_worker), title("average worker to log per capita GVA")
Command
Variables
Properties
G:\Stata15
4:46 PM
9/22/2020





So, let me proceed. We need to understand the correlation and graphical representation of the variables in order to understand some of the details very clearly. To check the relationship between variables scatterplot helps in understanding this graphically. These two variables of interest of us in case of biveriate model, so simply scatterplot and if I click these two variables, then I will probably get it and sometimes scatterplot may not, it might require installation. So, let us test.

So, the two variables of our interest is here and number of total workers, I think this. So, if I just do it, I told you already, it is saying unrecognized command. Then in that case what you need to do, you have to search scatterplot and accordingly you can derive it the way we guided earlier.

What I will do, instead of that I will quickly go to understand another approach of understanding the relationship through graphical representation of two-way table. I will simply do one thing to save our time. So what I will do, I will go to this and we will come to the particular slide and let us copy that you can test on your own. So, we will copy it and we will paste in the respective STATA window.

Once I paste it and I will go for entering the command it will certainly give us the right direction and the result is there to come. I think it is calculating, estimating, yes. We have got the graph. After getting this we get many information out of it. Look at the average number of workers.

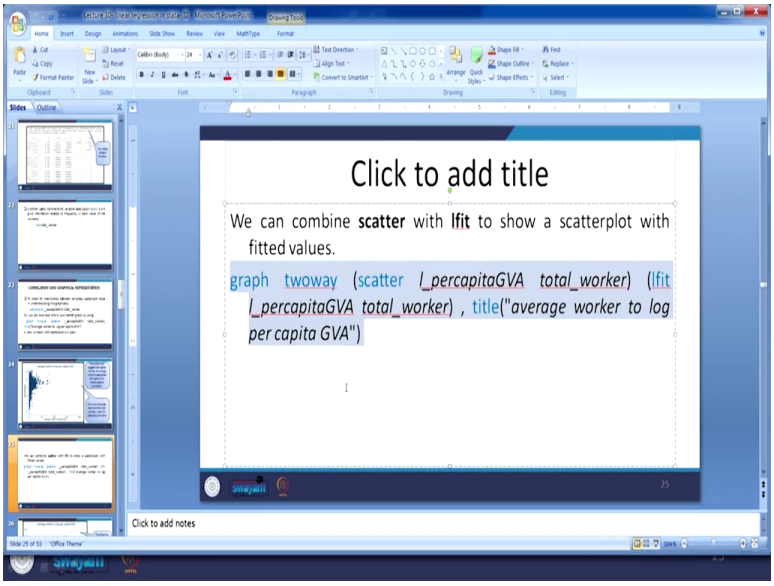
So, what we really submitted to the STATA window, two-way graph between per capita, log per capita GVA and total workers with a title that average worker to per capita GVA. So, the title if you mention, it comes at the above, average worker to per capita GVA.

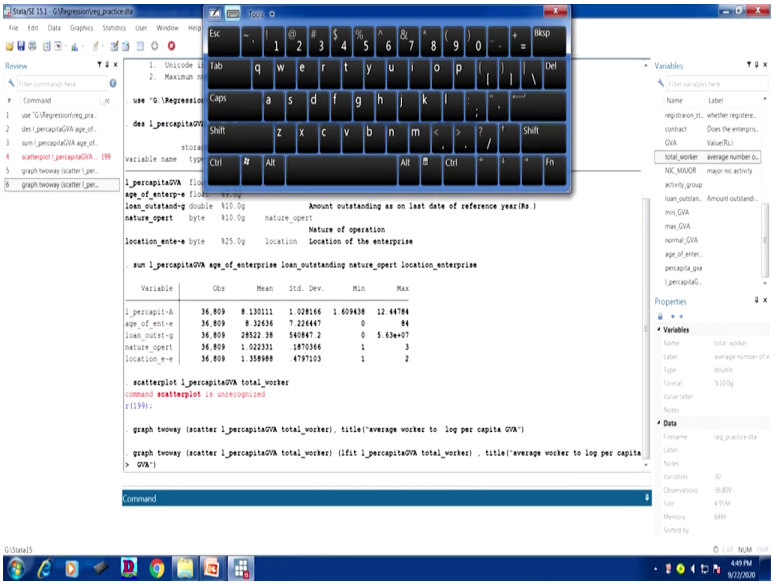
The per capita GVA in the vertical axis and average number of workers is there on the horizontal axis. A point is coming and highlighted here which represents 200 workers having log per capita income represents 200 workers. The, among the average workers, this is one outlier.

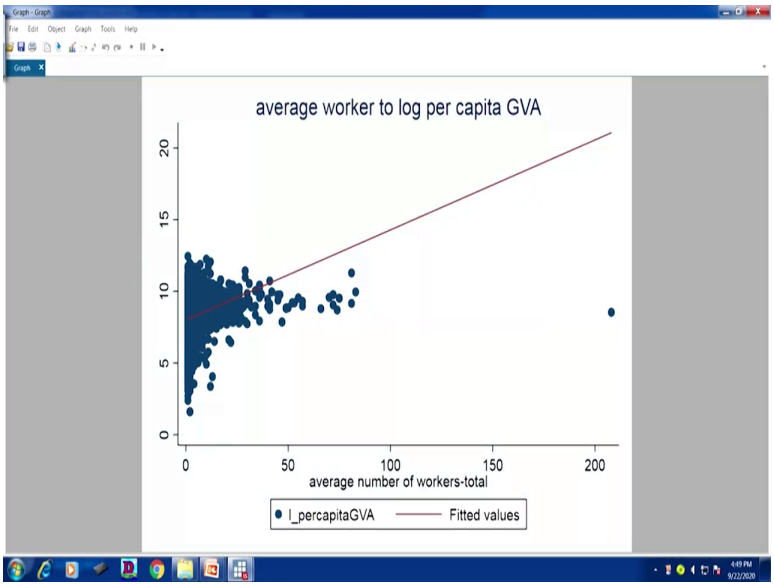
You can easily understand, this is a clear outlier and the outlier suggest that from the diagram if I can interpret like this, it is like only 1, that is 1 enterprise having 200 workers. One enterprise has more than 200 workers. It is more than 200 workers and that plotted against the log per capita GVA.

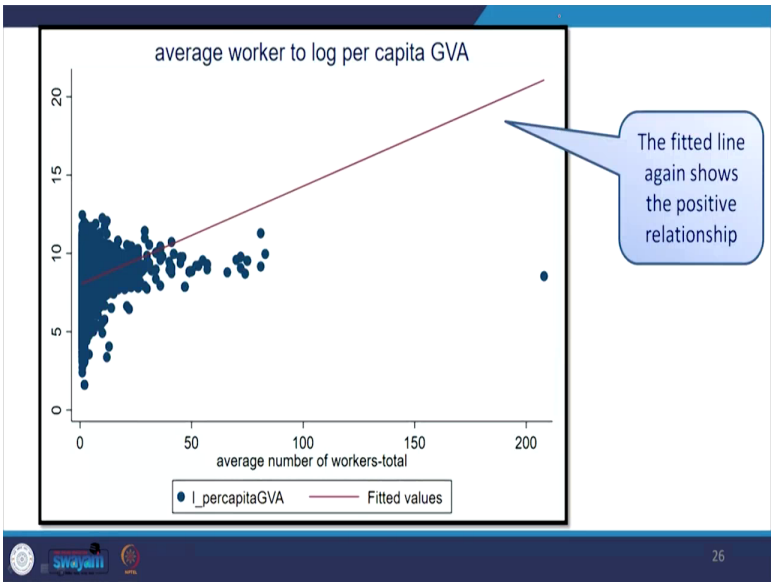
And this plot also suggests that higher the number of average worker higher the GVA that is roughly highlighting the positive relationship, but how to understand it unless we do not have a trend line. But somewhere higher amount of GVA per capita, GVA is associated with increasing trends of average number of workers except in one case.

(Refer Slide Time: 28:42)





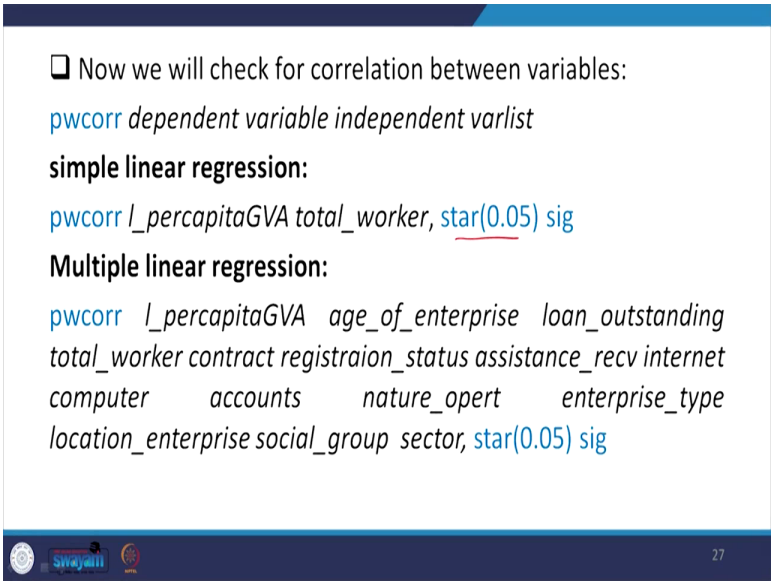


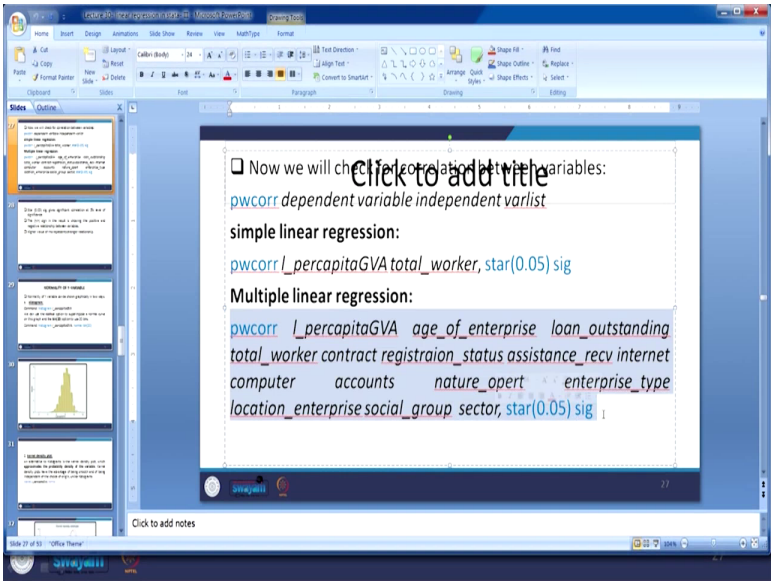


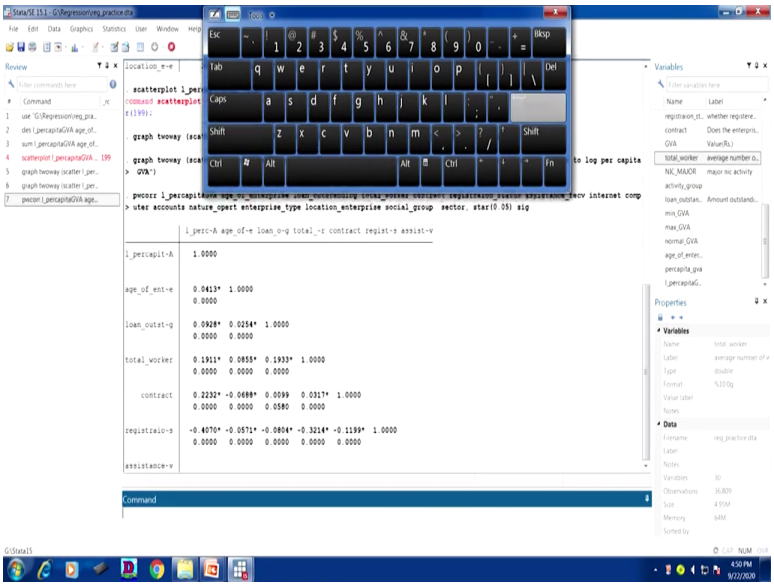
So for that, we will take the help of fitted line, whether it is rising or not, we have to take the help of a fitted line. So, let us copy this then we will operate this and then in the STATA page if you operate you will find out the result. From this it gives me a fitted line. Not just the graph, it also gives an upward trend line.

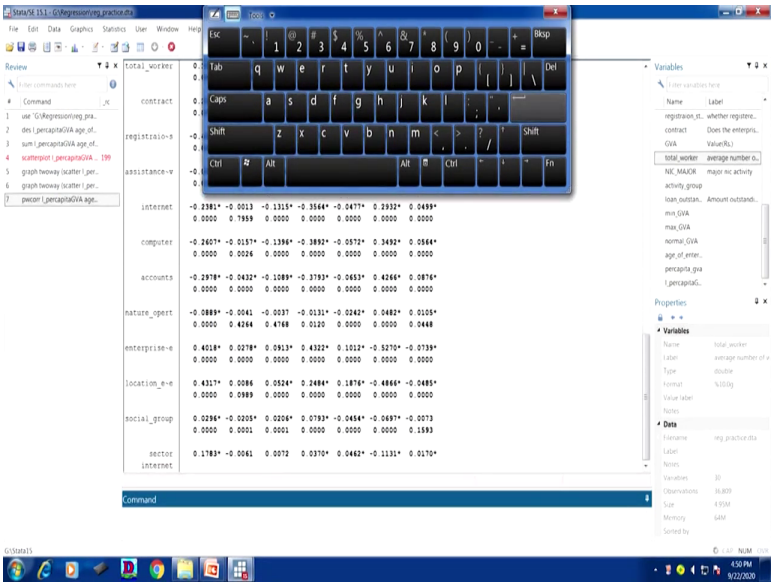
So this red line highlights fitted values. This shows that there exists a positive relationship. So, we are not spending much time on it and what it indicates is more important for our explanation and this is what we have explained. The fitted line shows the positive relationship as I just mentioned.

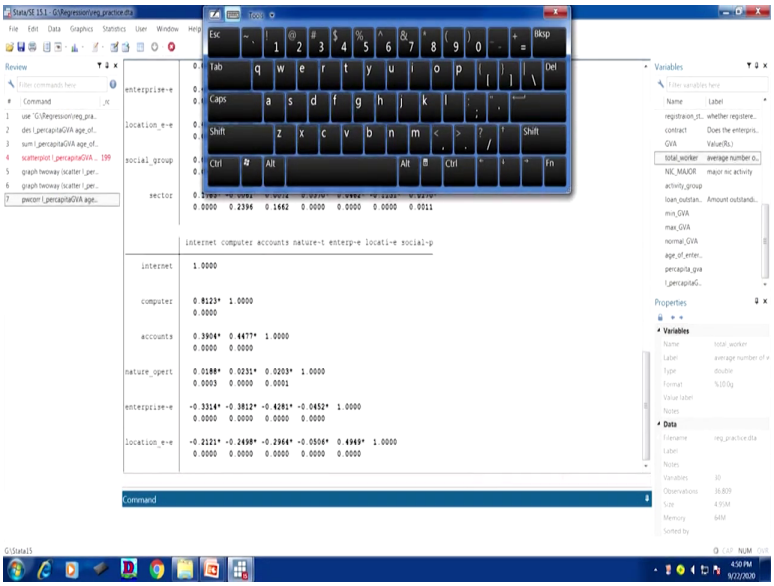
(Refer Slide Time: 29:41)

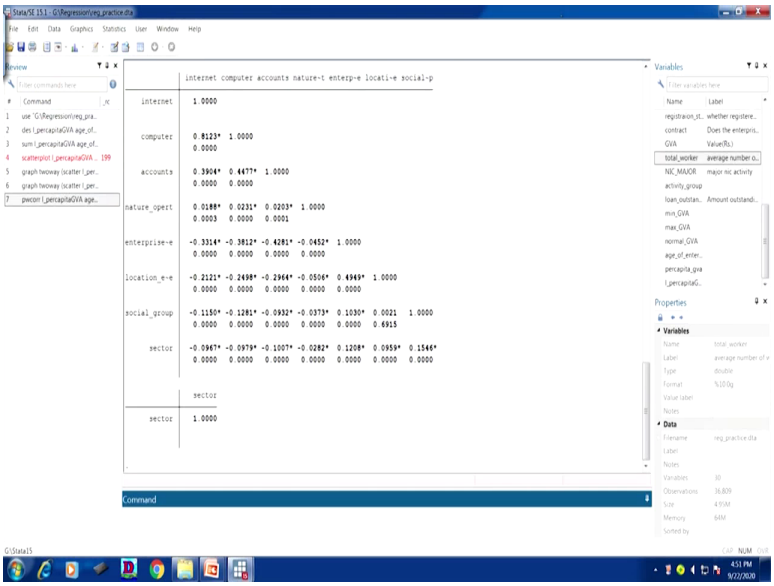


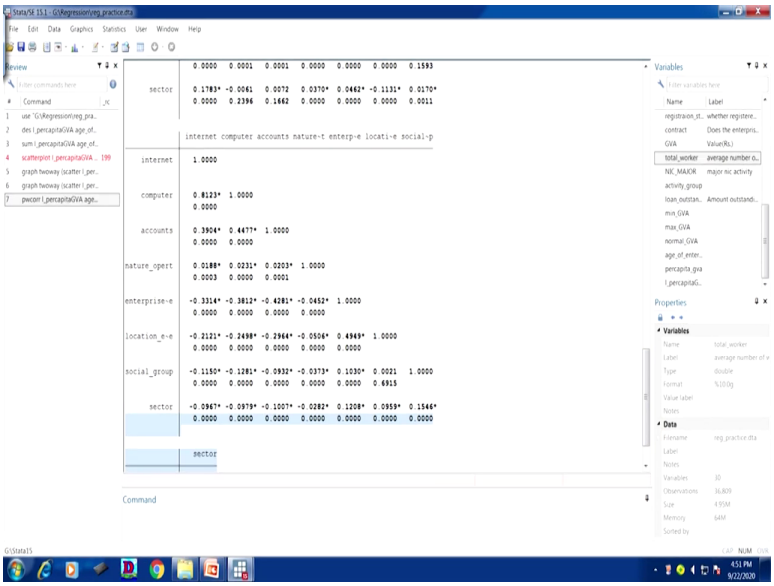



Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
1 use "G:\Regression\reg_pra...
2 des l_percapitaGVA age_of...
3 sum l_percapitaGVA age_of...
4 scatterplot l_percapitaGVA... 199
5 graph twoway (scatter l_per...
6 graph twoway (scatter l_per...
7 pwcorr l_percapitaGVA age...
Variables
Properties
Command
G:\Stata15
4:50 PM
9/22/2020


Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
1 use "G:\Regression\reg_pra...
2 des l_percapitaGVA age_of...
3 sum l_percapitaGVA age_of...
4 scatterplot l_percapitaGVA... 199
5 graph twoway (scatter l_per...
6 graph twoway (scatter l_per...
7 pwcorr l_percapitaGVA age...
Variables
Properties
Command
G:\Stata15
4:50 PM
9/22/2020


Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
Command
G:\Stata15
4:50 PM
9/27/2020


Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
Command
G:\Stata15
4:51 PM
9/27/2020

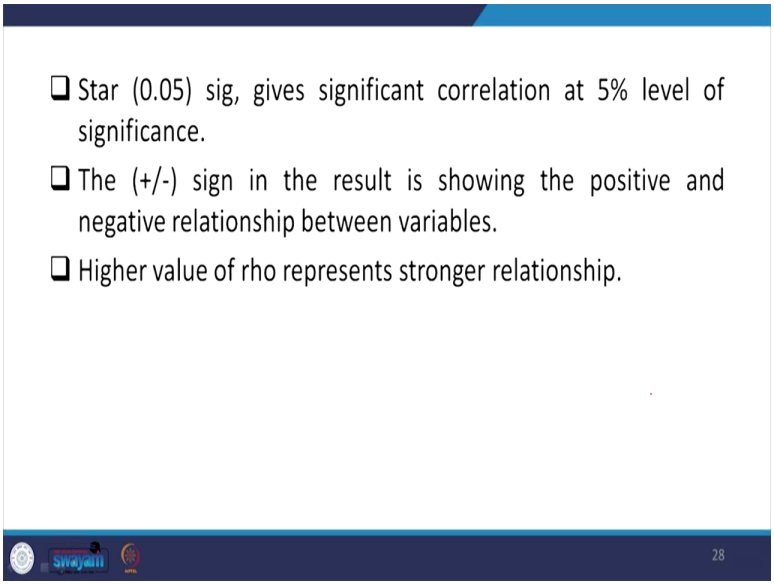


We will check for the correlation between the variables of our interest. Correlation can be checked with their significance level as well. Simple correlation, if I type this dependent and independent variable, we can get that. But just correlation may not have any interpretation unless we have any forms of significance level.

So, we will attach the significance level to it as well. Let us go by that and experiment this from the data. So, this is the 1 percent. Yes, all the variables are selected and we are operating through the regression window. Once I enter it, we derive the results. All the correlation value its 5 percent significance level has also been interpreted.

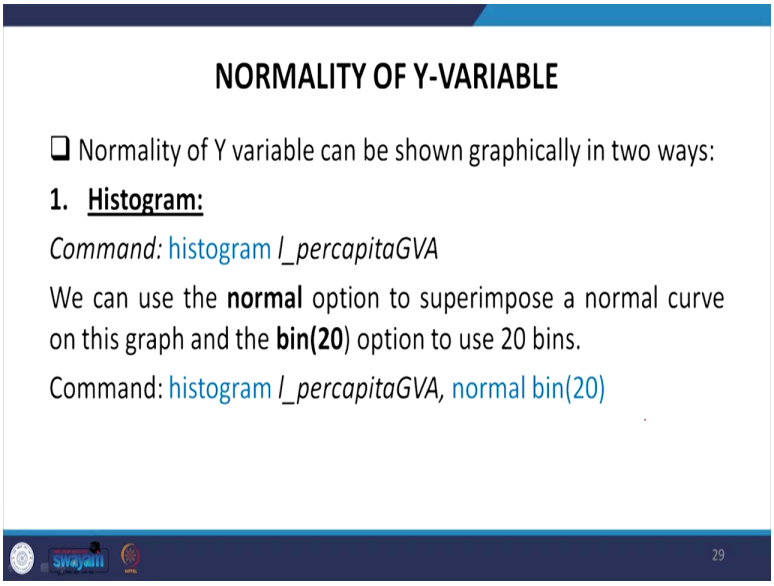
You need to understand very clearly that the significance level here we are mentioning is 5 percent you can change it to your number at other indicative level as well. Which variables are positively linked, which are negatively linked are clearly highlighted. So, there are two window to here all the variables and its correlation is highlighted. From the correlation we can establish whether the variables are linked to each other or not. So, this is what we said.

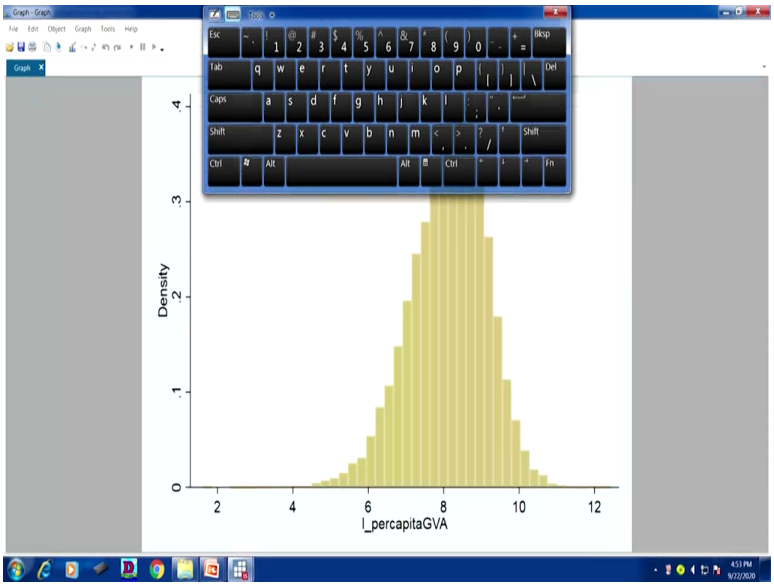
(Refer Slide Time: 31:19)

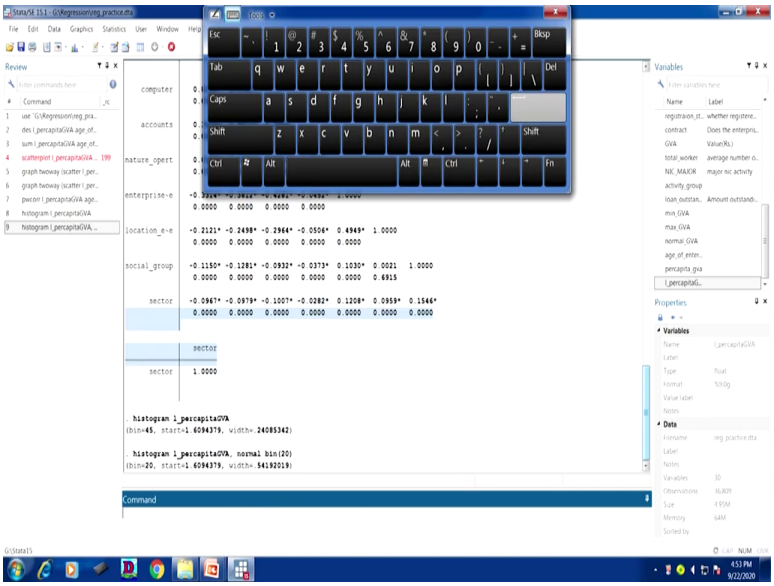


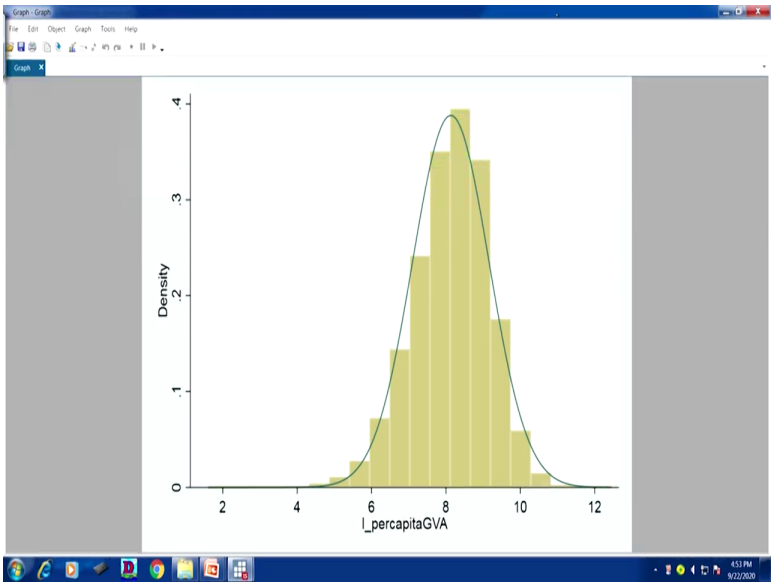
The plus and minus sign regarding their relationship has already been mentioned. Somewhere we get positive. Somewhere we get negative. So, higher values of row represent stronger relationship. The correlation if it is having higher value close to one generally that indicates better correlation, otherwise there is a weak correlation, but the significance level identifies even if there exist weak correlation but those variables are important for analysis if significance level gives you the direction.

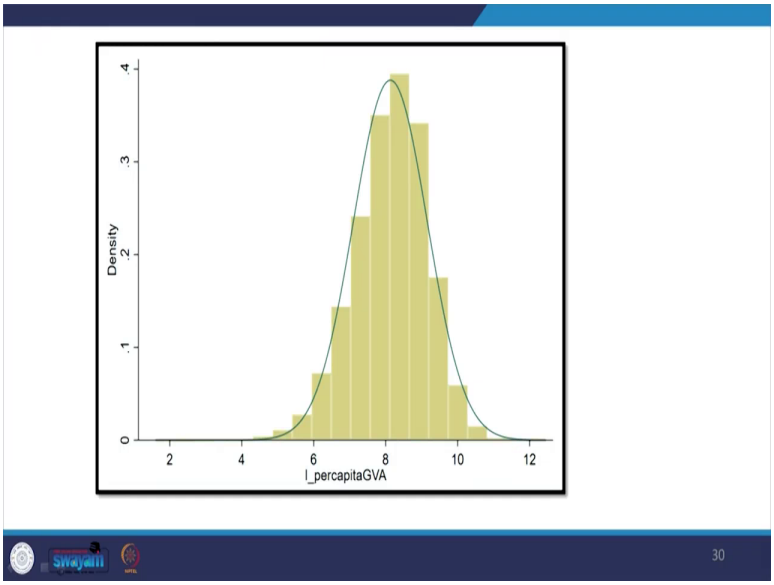
(Refer Slide Time: 32:02)






Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
computer
accounts
nature_opert
enterprise-e
location_e-e
social_group
sector
sector
sector 1.0000
. histogram l_percapitaGVA
(bin=45, start=1.6094379, width=.24085342)
. histogram l_percapitaGVA, normal bin(20)
(bin=20, start=1.6094379, width=.54192019)
Command
4:53 PM
9/22/2020


Graph - Graph
File Edit Object Graph Tools Help
Graph
Density
0
.1
.2
.3
.4
2
4
6
8
10
12
l_percapitaGVA
4:53 PM
9/22/2020

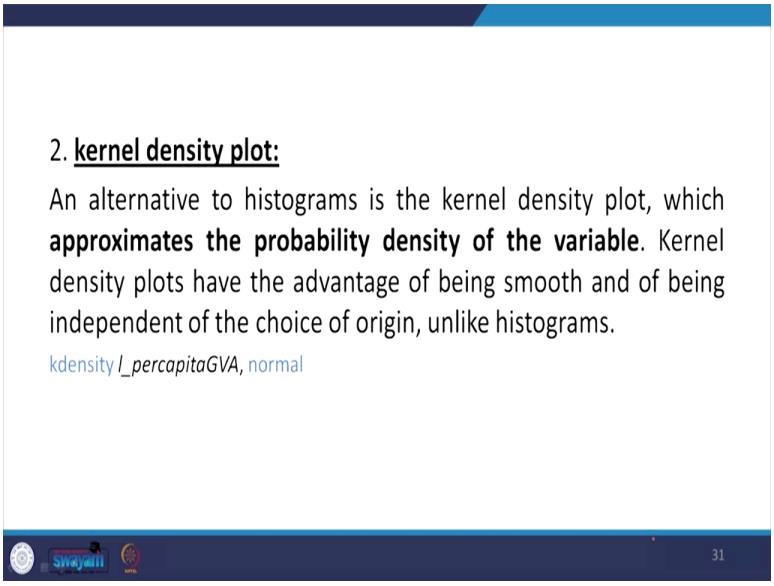


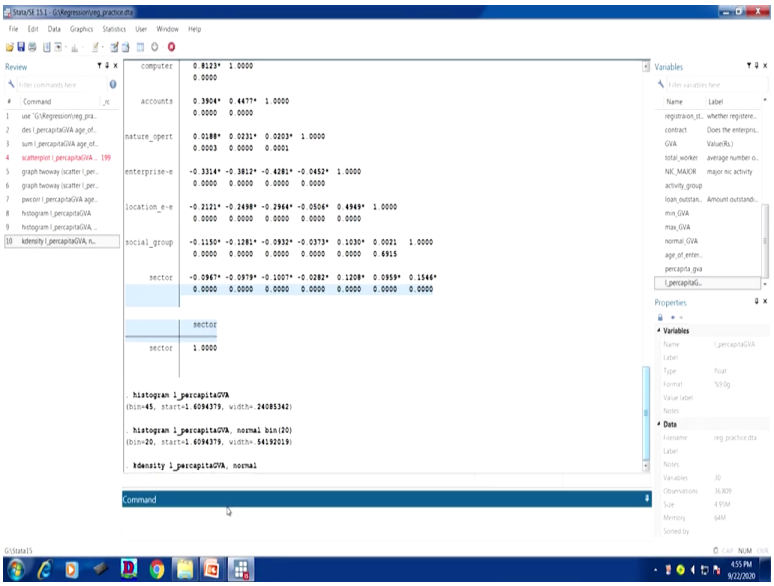
Coming to the explanation of normality of the Y variable, the dependent variable and that we can identify through the STATA and one of the approaches of measuring the normality is through histogram. So, histogram and the variable name if I just enter, we will get that. So, simply histogram and the variable name and this is the one and if you enter it derives the histogram with different bins.

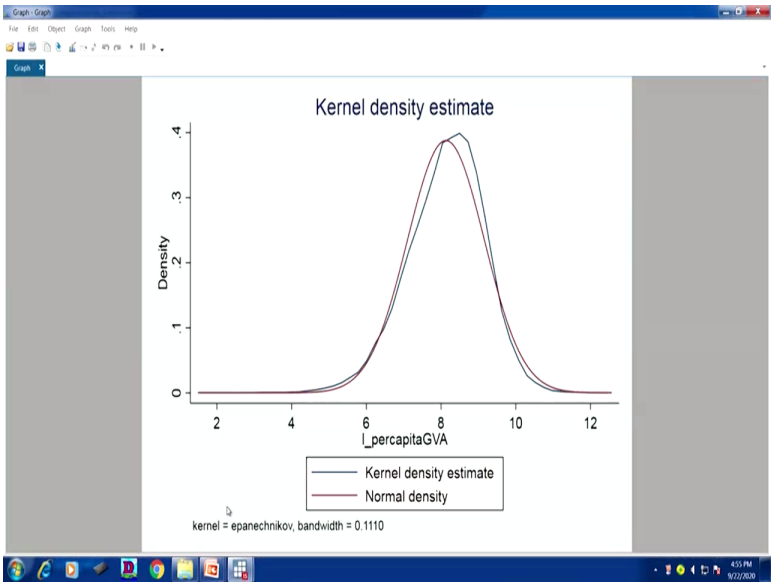
Different bins might be difficult to read at a go and might be problematic. We need to restrict it to its indicative level. Let we are going to control with a 20 number of bins. But if you want to plot the normality of that particular histogram you need to add with a comma normal bin 20.

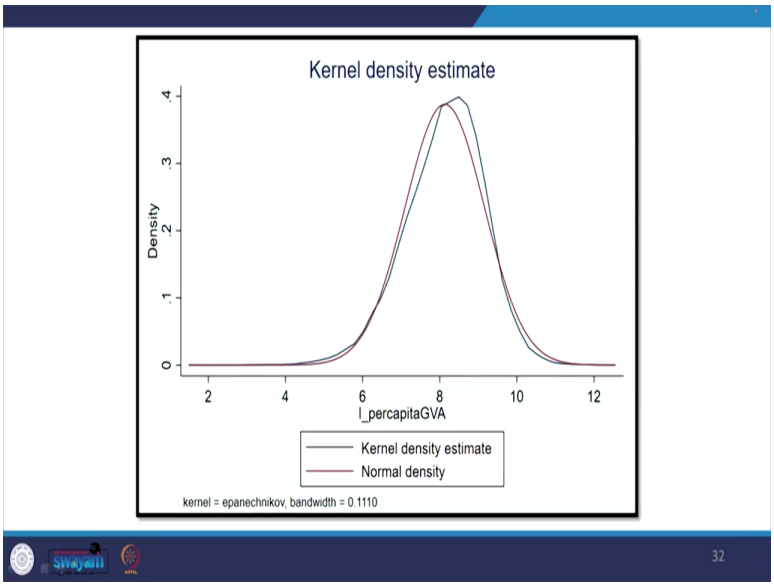
If I just simply do that then with the same command, the normal then bin 20. We will get the normal plot of that particular variable that is the dependent variable we are supposed to take is the log per capita GVA. It seems that the dependent variable is approximating to a normal distribution and one of the approaches has made it a near normality because of the log transformation of the GVA. The original GVA is generally not having normally distributed. So, let us move on to understand some other details. So, I have to close this then only it will work. So, here we are trying to explain the normality diagram that we have done it.

(Refer Slide Time: 34:17)





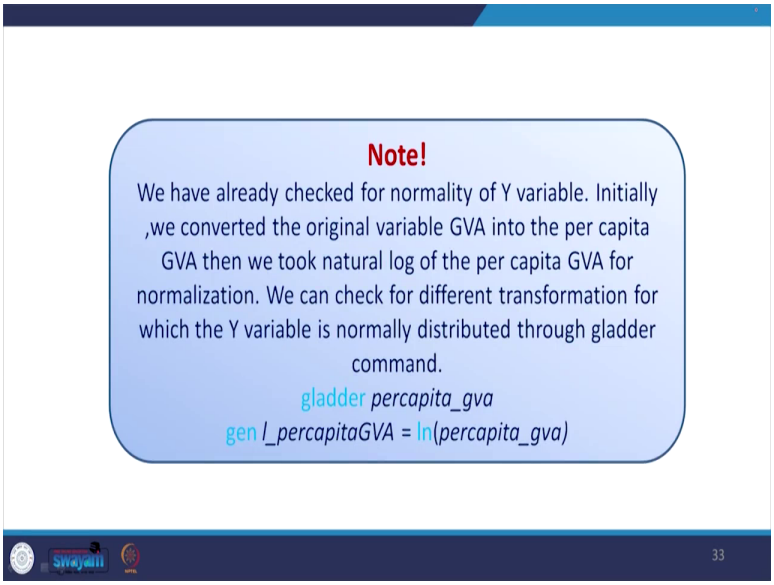




I am going to discuss about another approach of understanding the details. So, the kernel density plot, this is an alternative to histograms and which approximate the probability density of the variable. The kernel density plots have the advantage of being smooth and being independent of the choice of origin unlike the histogram. So, we can experiment and we can find out.

So with the same command, only kdensity, we are writing kdensity, the same Y dependent variable then normal. So that gives us a smooth diagram. The kernel density estimate is presented against a standard normal distribution. So, it is almost approximating to the distribution. So, this is what we have shown.

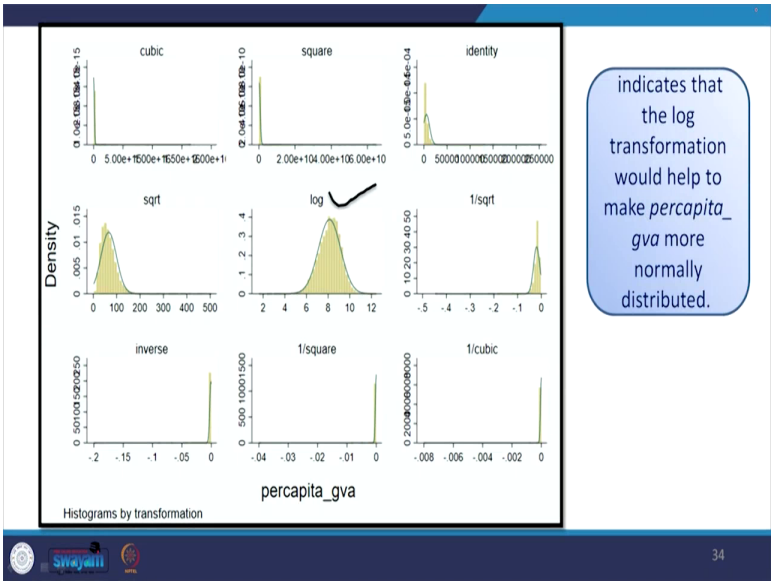
(Refer Slide Time: 35:28)



I wanted to mention that there is a command. We need to note that we have already checked for normality of the Y variable. So, initially we converted the original variable that is GVA into a per capita GVA then we took the natural log of it and we can check for different transformation of which the Y variable is normally distributed through a gladder command.

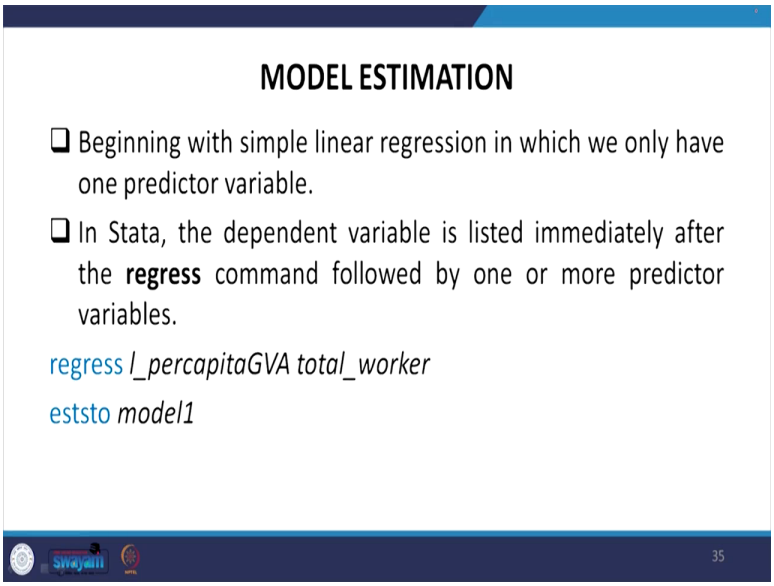
Gladder command establishes, whether particular transformation is most fitted. Most appropriate for converting to a normality. So, let us go by the gladder and the per capita GVA first then we can generate in terms of.

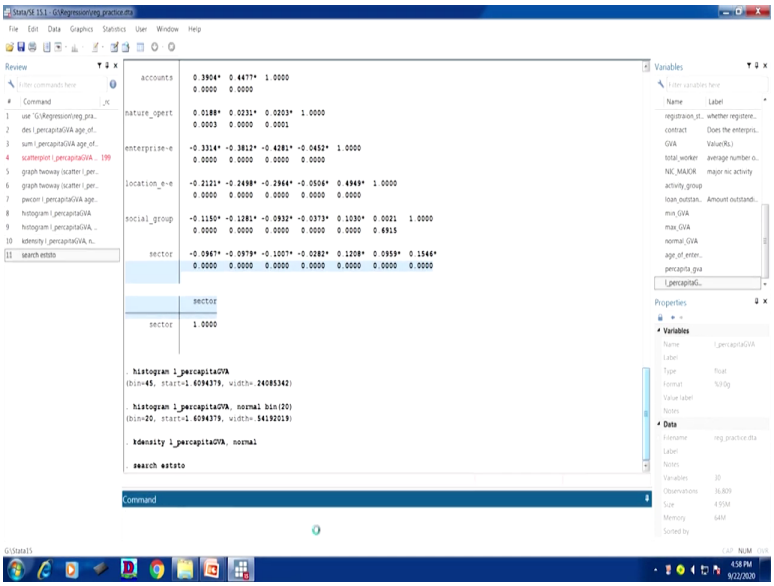
(Refer Slide Time: 36:16)

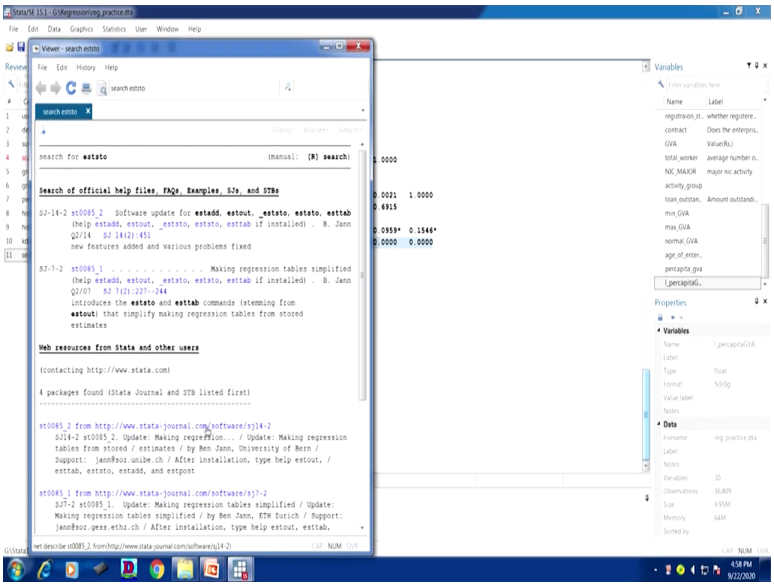


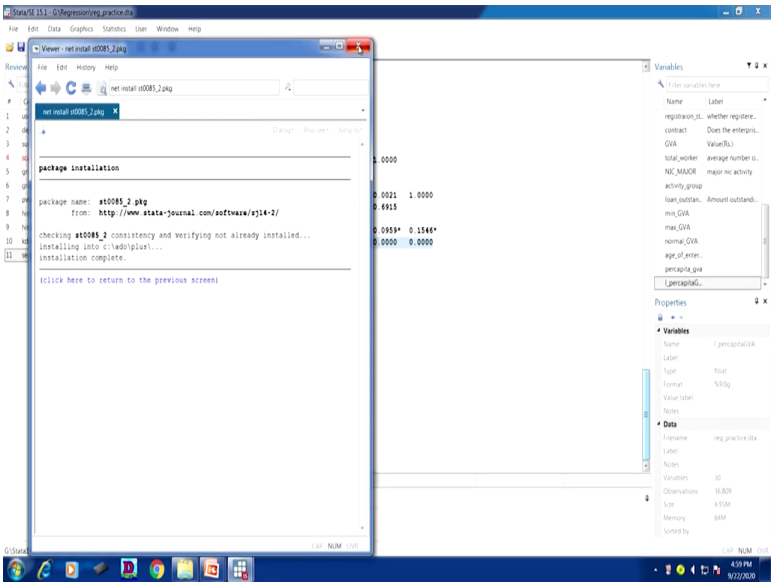
So, it looks like this. If you go for the same command, you will find this type of results. So, out of all those things, so many different results popped up and now it seems that the log transformation of this particular per capita GVA is approximating to a normal distribution. So, I would not spend much time. We have many information to be explored.

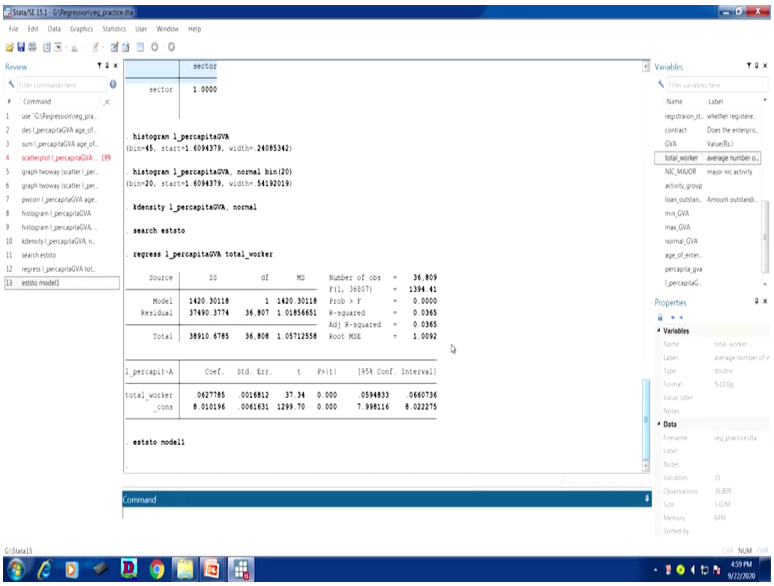
(Refer Slide Time: 36:54)




Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
accounts 0.3904* 0.4477* 1.0000
nature_opert 0.0188* 0.0231* 0.0203* 1.0000
enterprise-e -0.3314* -0.3812* -0.4281* -0.0452* 1.0000
location_e-e -0.2121* -0.2498* -0.2964* -0.0506* 0.4949* 1.0000
social_group -0.1150* -0.1281* -0.0932* -0.0373* 0.1030* 0.0021 1.0000
sector -0.0967* -0.0979* -0.1007* -0.0282* 0.1208* 0.0959* 0.1546*
sector
sector 1.0000
. histogram l_percapitaGVA
(bin=45, start=1.6094379, width=.24085342)
. histogram l_percapitaGVA, normal bin(20)
(bin=20, start=1.6094379, width=.54192019)
. kdensity l_percapitaGVA, normal
. search eststo
Command
G:\Stata15
4:58 PM
9/22/2020


Stata/SE 15.1 - G:\Regression\reg_practice.dta
Viewer - search eststo
search eststo
search for estsato (manual: [R] search)
Search of official help files, FAQs, Examples, SJs, and STBs
SJ-14-2 st0085_2 Software update for estadd, estout, _eststo, eststo, esttab
(help estadd, estout, _eststo, eststo, esttab if installed) . B. Jann
Q2/14 SJ 14(2):451
new features added and various problems fixed
SJ-7-2 st0085_1 . . . . . Making regression tables simplified
(help estadd, estout, _eststo, eststo, esttab if installed) . B. Jann
Q2/07 SJ 7(2):227--244
introduces the estato and esttab commands (stemming from
estout) that simplify making regression tables from stored
estimates
Web resources from Stata and other users
(contacting http://www.stata.com)
4 packages found (Stata Journal and STB listed first)
st0085_2 from http://www.stata-journal.com/software/sj14-2
SJ14-2 st0085_2. Update: Making regression... / Update: Making regression
tables from stored / estimates / by Ben Jann, University of Bern /
Support: jann@soz.unibe.ch / After installation, type help estout, /
esttab, eststo, estadd, and estpost
st0085_1 from http://www.stata-journal.com/software/sj7-2
SJ7-2 st0085_1. Update: Making regression tables simplified / Update:
Making regression tables simplified / by Ben Jann, ETH Zurich / Support:
jann@soz.gess.ethz.ch / After installation, type help estout, esttab,
net describe st0085_2 from http://www.stata-journal.com/software/sj14-2
Variables
Properties
G:\Stata15
4:58 PM
9/22/2020





Let us come to the model estimation. So, regarding Y variable and its normality, we have already explained. Let us understand the model estimation. Beginning with simple linear regression in which we only have a one predator variable. In STATA, the dependent variable is listed immediately after the regress command. It is followed by one or more predictor variables. So, simply we will write down regress, its dependent and independent variable.

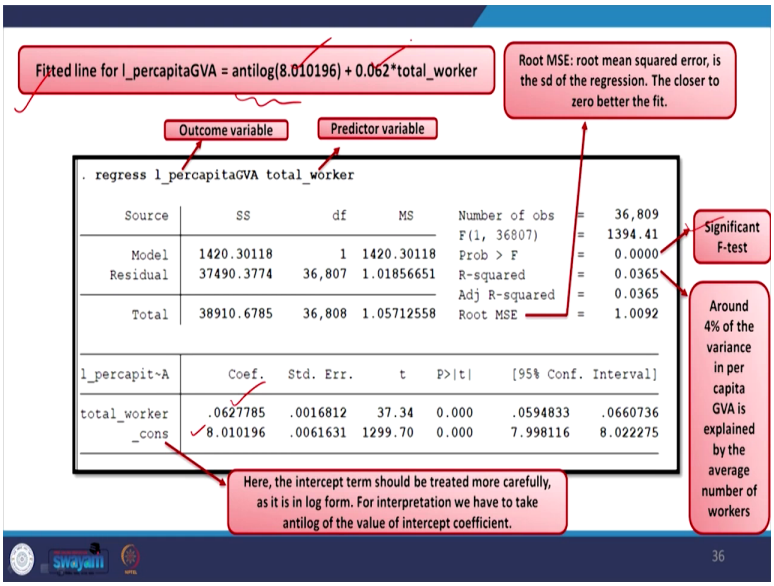
One of the very important suggestions to you that please try to run another regression command to it after regress so that you can compare all the models, eststo command of that model 1, similarly, in the model 2, I am going to show you in model 2. So, let us see, if it is not there, we

can install it and accordingly we will move further. So, let us explain this two together first. So, in that case it will be of eststo. So the first suggestions we need to install. So, after installing only it will work and it gives the right window for comparing the models. So, it is being installed and then only we can able to operate.

The regression command takes the variables. I think it has been installed. We are going to start with the reg. So it will be reg then dependent variable and independent variable. So, this is the dependent variable we wanted to discuss then number of workers. By entering it we derive the results.

Followed by that you have to go by eststo then let us write down as model 1. This has already saved. Basically, eststo saved the results of the model 1 and then at the end we can able to compare. This is the regression result with one independent variable. Similarly, you can go for many independent variables, if I mentioned and the results will display correctly.

(Refer Slide Time: 39:17)



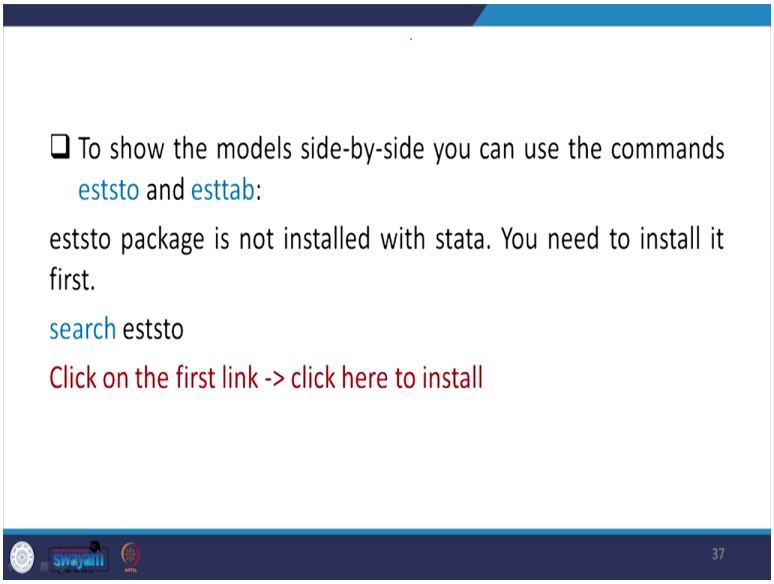
So, here is the result and what is important to be mentioned, everything is given here, that we start with the command here as outcome variable is here, then the predictor variable is here highlighted very clearly. This result is also gives the significance level, the F-test value and it suggests that the R square value is around 4 percent. Then adjusted R square is value is also

given. We have already clarified in the previous lectures that it says that around 4 percent of the variance in per capita GVA is explained by the average number of workers.

The root mean square error is the standard deviation of the regression. The closer it is to 0 it is the better fit for the model. So, less mean square basically less standard deviation so better is the outcome and better is the model. Last one in this particular result to interpret is that the intercept or term that is constant, term should be treated more carefully as it is in log form, because our dependent variable is in logarithmic transmission.

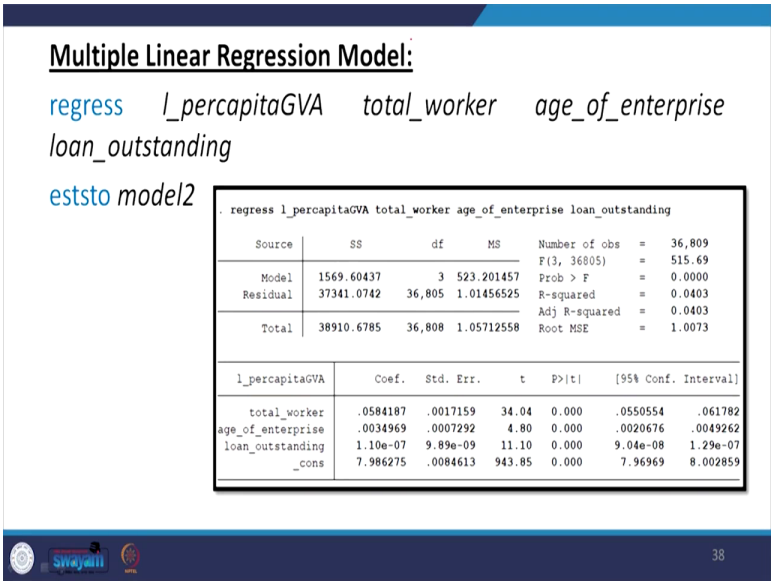
So, in order to get that for the interpretation we have to take antilog of that coefficient. If you do not take antilog I think the right interpretation is not there. Therefore, the fitted line is written at the top as this. So, the fitted line suggests that you have to take the antilog for the interpretation. The coefficient is of interest is that the coefficient we have derived this 8.01 is mentioned here and the total worker at the beta coefficient is given here. So, if you just fit those things, we can estimate and estimate the equation correctly.

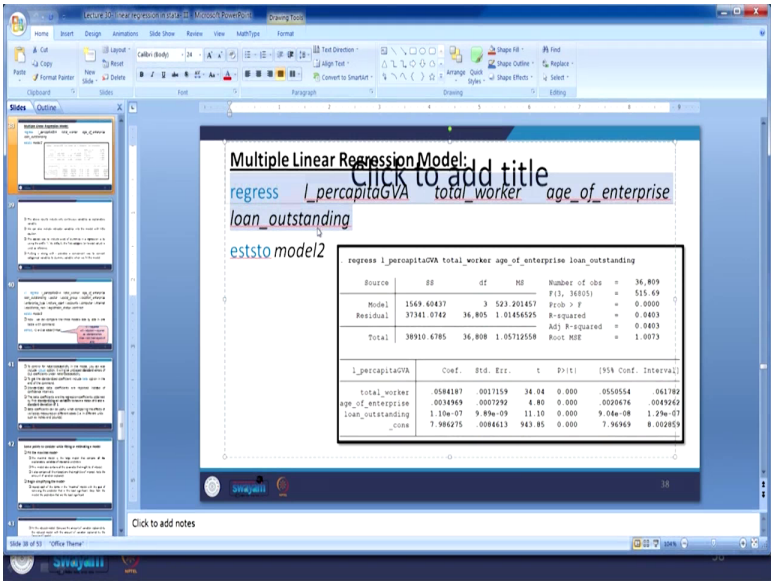
(Refer Slide Time: 41:32)

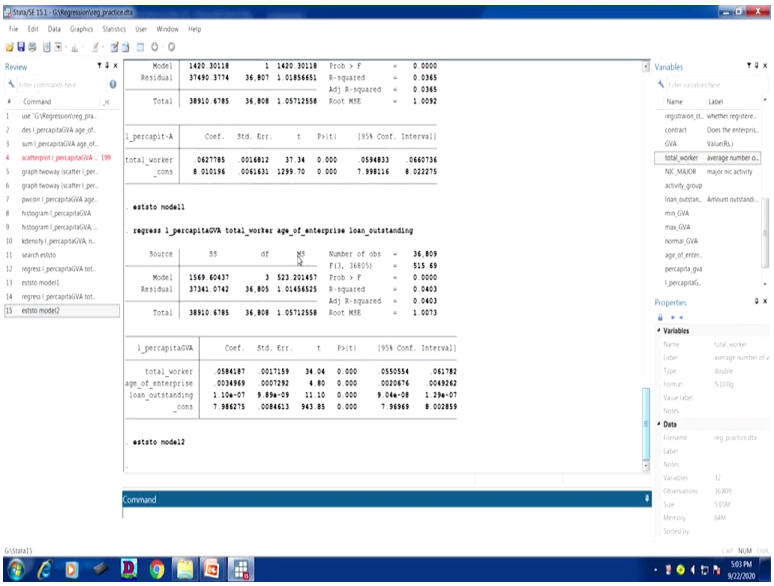


So, and also we have entered the eststo command and at the end we will operate with esttab to find out the result for comparison.

(Refer Slide Time: 41:42)

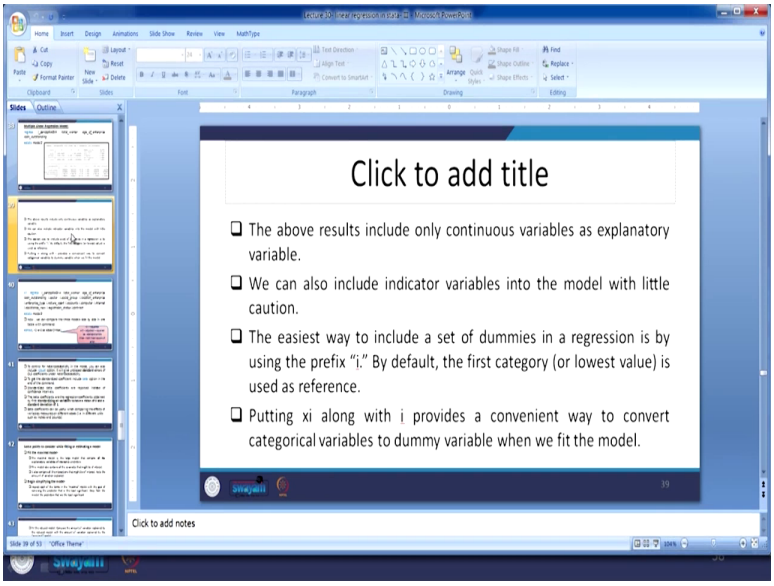






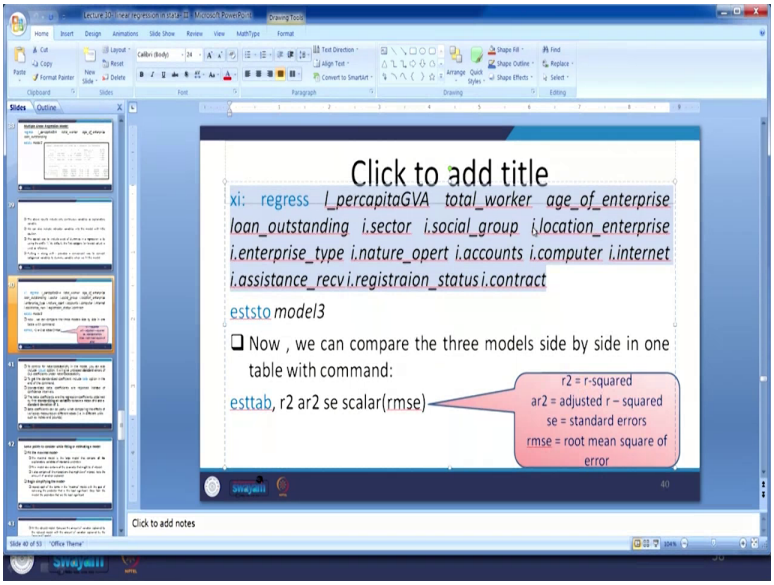
So, let us understand another approach for another model also. So, regress then we will copy this or we will operate these variables and I wanted to show that how eststo command works. So, it is here. So let us go to that window. Simply we will copy it and operate this through our STATA and STATA gives the result. Then you can save with eststo model 2. Then this has already saved. Now I am not interpreting much on this. It is more or less similar. So we are going to discuss another aspect. It is more important.

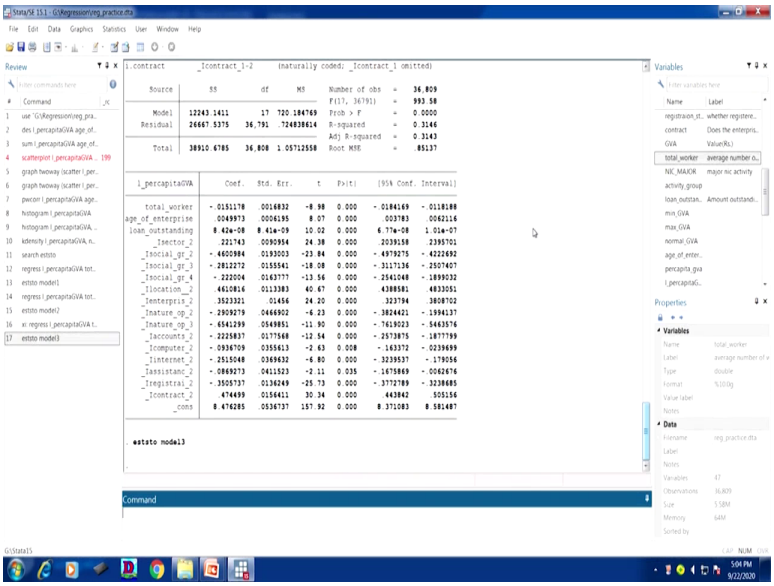
(Refer Slide Time: 42:33)

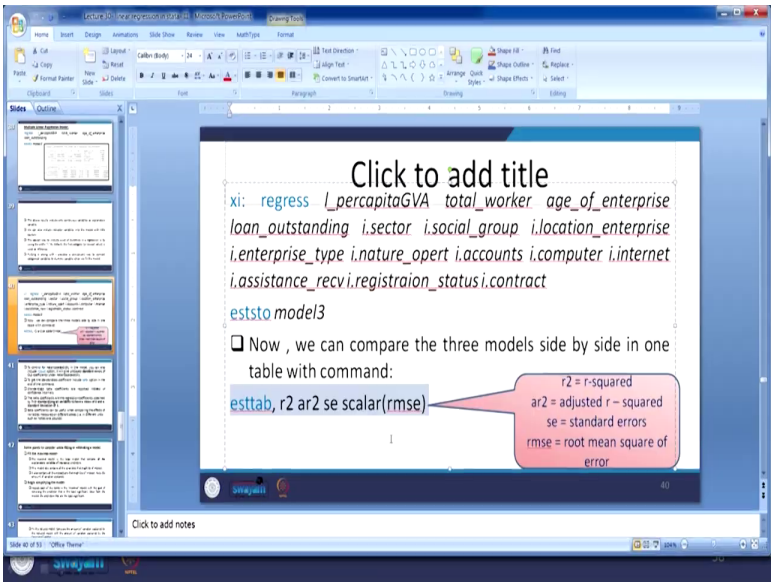


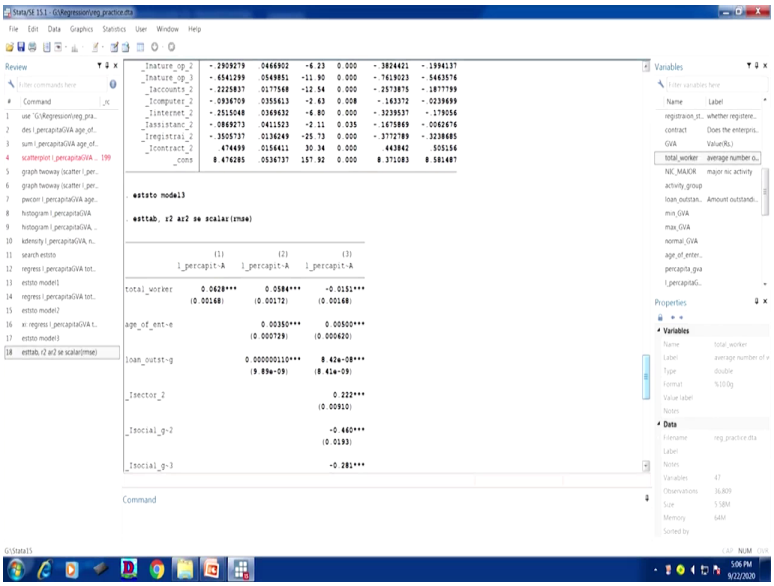
If there are so many variables, there are some dummies in the regression command, in order to emphasize the dummies just averaging regression usually does one thing, it simply averages the variables. So, what so far in both the models we have taken the averages. But if you bifurcate by the dummies, we can include a prefix command with i. So, by default the first category is considered as a reference term. We are going to show it right now and for that we need to operate with a xi command. So, let us go by that command to operate.

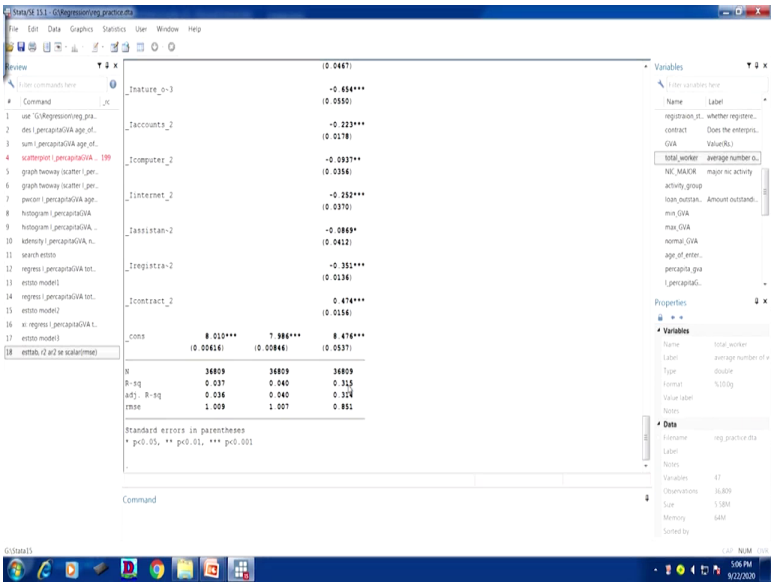
(Refer Slide Time: 43:20)




. i.contract _Icontract_1-2 (naturally coded; _Icontract_1 omitted)
Source SS df MS Number of obs = 36,809
F(17, 36791) = 993.58
Model 12243.1411 17 720.184769 Prob > F = 0.0000
Residual 26667.5375 36,791 .724838614 R-squared = 0.3146
Adj R-squared = 0.3143
Total 38910.6785 36,808 1.05712558 Root MSE = .85137
l_percapitaGVA Coef. Std. Err. t P>|t| [95% Conf. Interval]
total_worker -.0151178 .0016832 -8.98 0.000 -.0184169 -.0118188
age_of_enterprise .0049973 .0006195 8.07 0.000 .003783 .0062116
loan_outstanding 8.42e-08 8.41e-09 10.02 0.000 6.77e-08 1.01e-07
_Isector_2 .221743 .0092934 24.38 0.000 .2039158 .2395701
_Isocial_gr_2 -.4605984 .0193003 -23.84 0.000 -.4979275 -.4232692
_Isocial_gr_3 -.2812272 .0155541 -18.08 0.000 -.3117136 -.2507407
_Isocial_gr_4 -.222004 .0163777 -13.56 0.000 -.2541048 -.1899032
_Ilocation__2 .4610816 .0113383 40.67 0.000 .4388581 .4833051
_Ienterpris_2 .3523321 .01456 24.20 0.000 .323794 .3808702
_Inature_op_2 -.2909279 .0466902 -6.23 0.000 -.3824421 -.1994137
_Inature_op_3 -.6541299 .0549851 -11.90 0.000 -.7619023 -.5463576
_Iaccounts_2 -.2225837 .0177568 -12.54 0.000 -.2573875 -.1877799
_Icomputer_2 -.0936709 .0355613 -2.63 0.008 -.163372 -.0239699
_Iinternet_2 -.2515048 .0369632 -6.80 0.000 -.3239537 -.179056
_Iassistanc_2 -.0869273 .0411523 -2.11 0.035 -.1675869 -.0062676
_Iregistrai_2 -.3505737 .0136249 -25.73 0.000 -.3772789 -.3238685
_Icontract_2 .474499 .0156411 30.34 0.000 .443842 .505156
_cons 8.476285 .0536737 157.92 0.000 8.371083 8.581487
. eststo model3
Command
Variables
Properties
Data
Review


Click to add title
xi: regress l_percapitaGVA total_worker age_of_enterprise loan_outstanding i.sector i.social_group i.location_enterprise i.enterprise_type i.nature_opert i.accounts i.computer i.internet i.assistance_recv i.registraion_status i.contract
eststo model3
Now , we can compare the three models side by side in one table with command:
esttab, r2 ar2 se scalar(rmse)
r2 = r-squared
ar2 = adjusted r – squared
se = standard errors
rmse = root mean square of error
Click to add notes





So, xi command is there, xi then regress. Regress and where we have dummy variable we have taken i dot. You mark this carefully and the i dot command is going to give us the correct result. we have got the result and the dummy variables, also we will save with eststo model 3. I just wanted to mention that you have already seen these basic results like F-test, its significance label, its R square, adjusted R square. I just wanted to say, when our sector is a dummy it has by default considered 1 as the first reference category. As compared to one, sector 1 that is rural area, rural has been considered as the reference category or the base category.

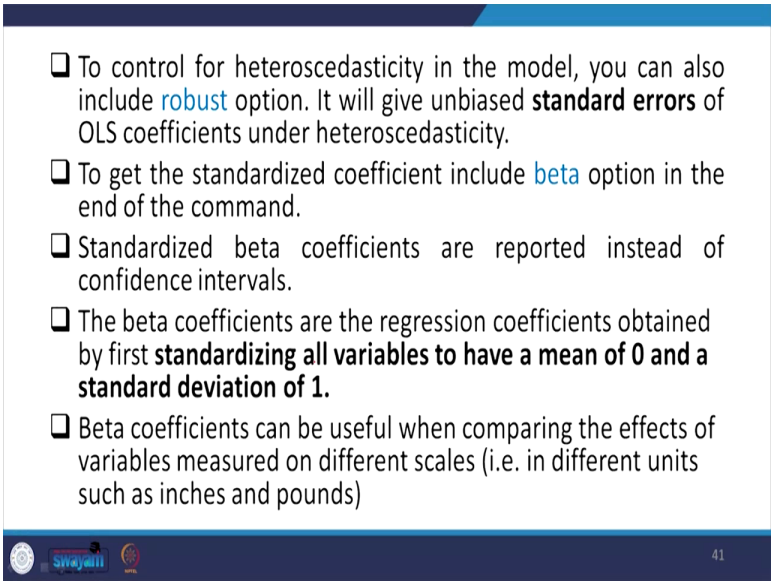
As compared to rural, the urban coefficient is 0.221743 and that is significant in our model. Similarly, other social categories then 1 is considered to be the base category, then 2, 3, 4 are considered as the other categories. So, interpretation is made accordingly. At the end, after discussing all those 3 regression models we are there to compare. We are now going to understand the eststo command and its results.

So, we will just take this there est command then we will paste it and then we can also find out the result. Look at all the models so far we have taken is presented here. Its model 1 is presented here, model 2, model 3 and in model 1, if you remember, we have taken only 1 variable then in model 2 there are 3 independent variables and in model 3 all other variables have been taken.

Very important for the paper, in most of the papers, they do require comparison. You can find out that the R square are also compared. The adjusted R square is also compared. Look at in the third model adjusted R square or the R square value is more than 31 percent as compared to other 2.

So, we may consider that the third model is more fitted and that too the root mean square error is also less than that of the other two. So the third model, since we have considered more variables, more diversity is there that explains the model correctly and accordingly that also reduces the R square value, so that reduces the mean, root mean square error that is the variance error term has been reduced. But the R square value is increasing. So, let us move for some other clarifications. We have almost dealt with most of the important aspects, but some other interpretations are still left.

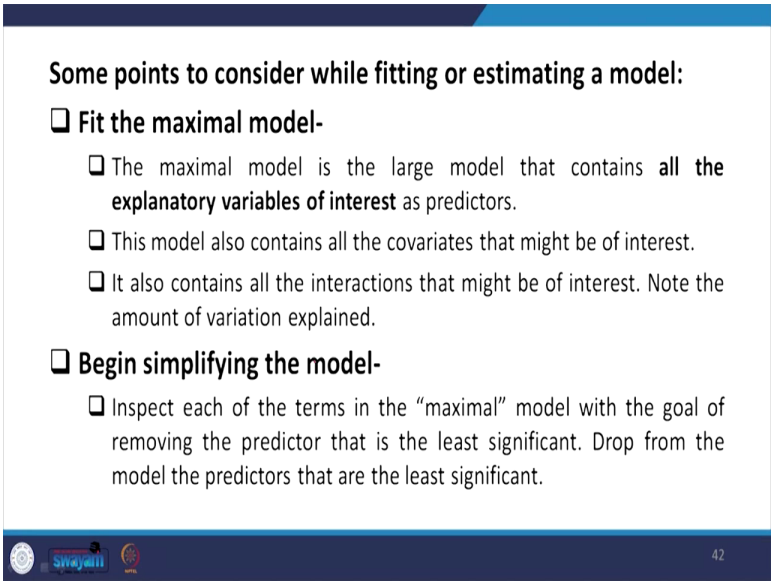
(Refer Slide Time: 46:46)



To control for heteroscedasticity in the model we can also include, I think we have already discussed the aspect of heteroscedasticity, we can also include robust options. So, robust option gives us the unbiased standard errors of OLS coefficients and that is discussed under heteroscedasticity.

So, to get the standardized coefficient include beta options in the end of the command and standardized beta coefficients are reported instead of confidence interval. The beta coefficients are the regression coefficients obtained by first standardizing all variables to have a mean of 0 and a standard deviation of 1. That we have already discussed earlier. Beta coefficient can be useful when comparing the effects of variables measured on different scales or in different units such as inches or pounds.

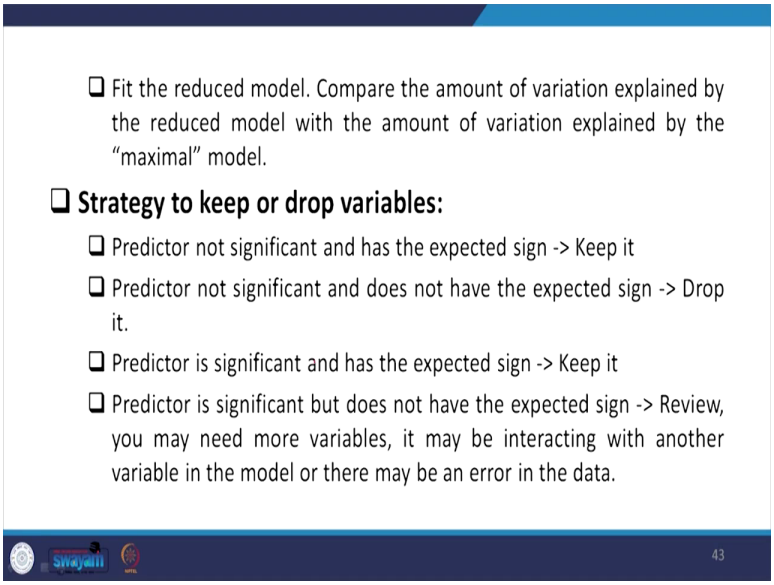
(Refer Slide Time: 47:52)



So, there are some points for consideration in order to understand the fitting or eststo model. There are two majorly we are discussing at this moment. First is the fitting the maximal model. The maximal model is the large in a model that contains all the explanatory variables of interest. This model also contains all the covariates that might be of interest. This contains the interactions that might be of interest. Note that the amount of variation is also explained.

Begin simplifying the model through inspect each of the terms in the maximal model with the goal of removing the predictor that is the least significant and drop from the model of the predictor that are least significant. Some of the variables we will drop to find out the best fit.

(Refer Slide Time: 48:30)

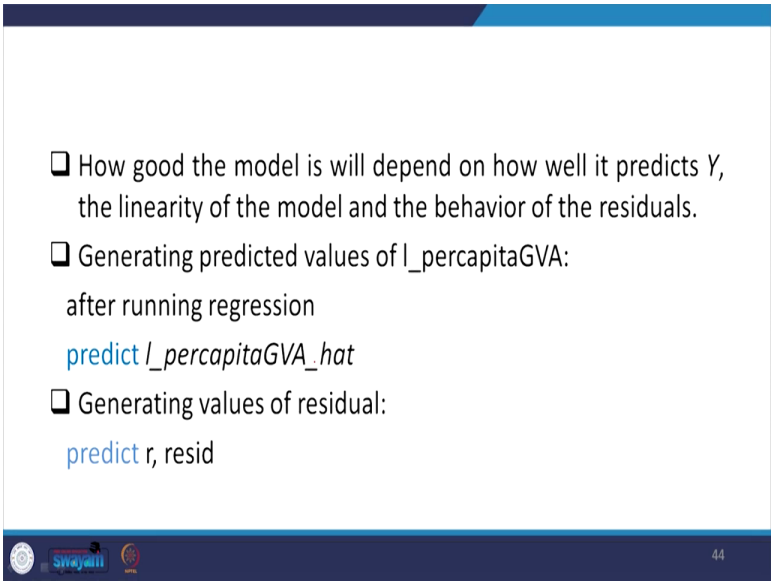


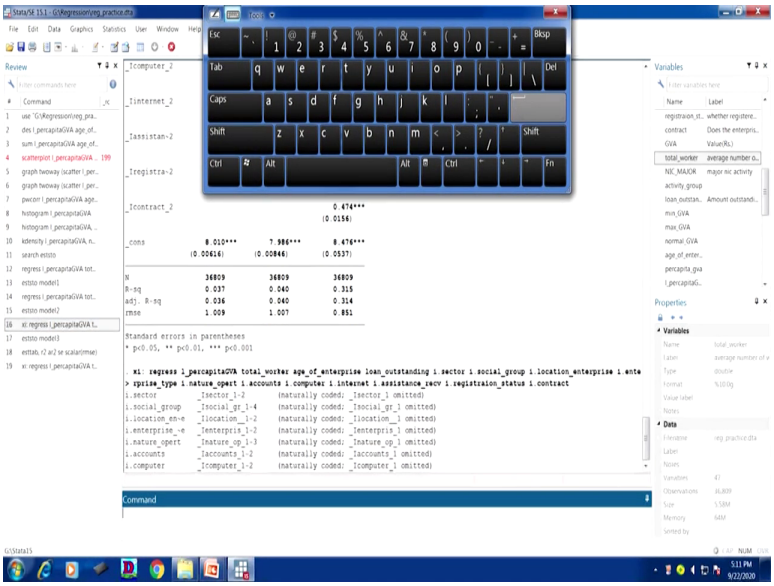
So, fit the reduced model compare the amount of variation explained by the reduced model with the amount of variation explained by the maximal model. We are going to discuss in our slide. So, strategy to keep or drop variable is important. So, the predictor not significant and has the expected sign then we need to keep it.

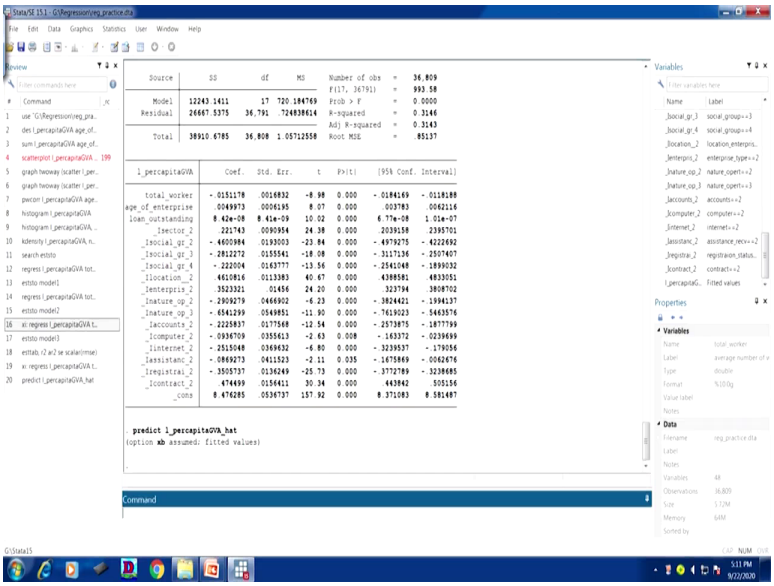
So, keep command we already mentioned. The predictor which has better expected sign we need to keep it. So, keep and that variable will keep, if you write down keep and variable name, just write the variable name, it will keep the variable and it will drop other variables.

When you are doubly sure that this is not important and does not have the expected sign then drop that variable. So, predictor is significant and has the expected sign then keep it, we have already said and then predictor is significant but does not have the expected sign then review it. You may need more variables in order to have better result and it may be interacting with another variable in the model or there may be an error in the data.

(Refer Slide Time: 49:46)



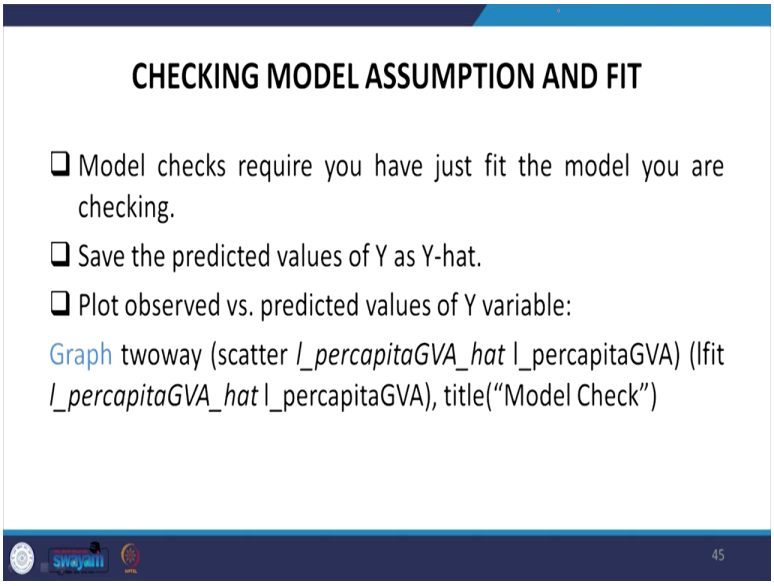


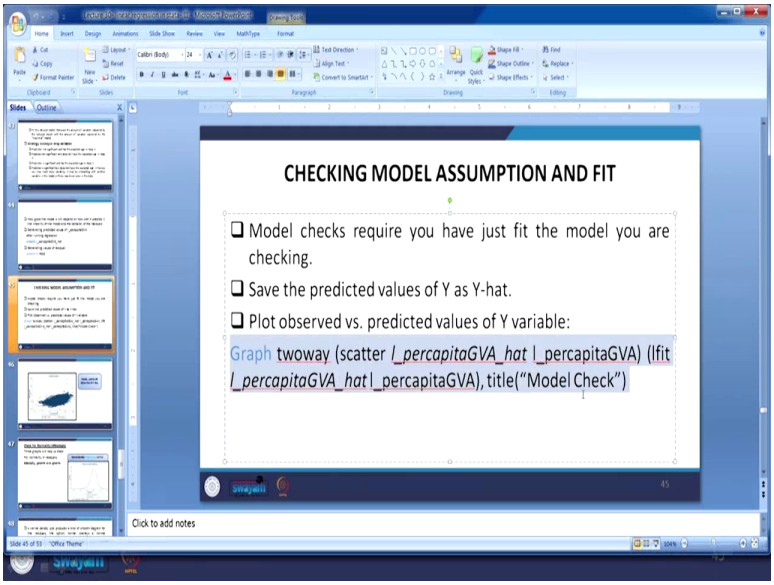


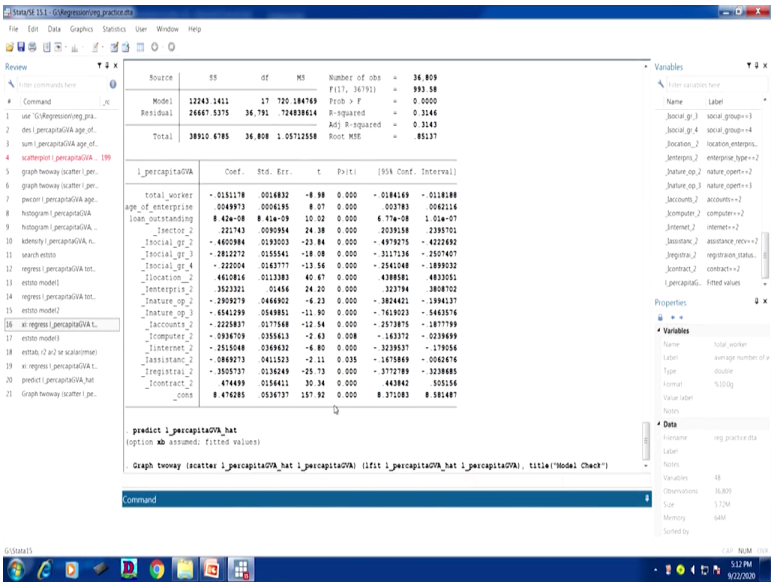
How good is the model will depend on how well it predicts the Y, the linearity of the model and the behavior of the residuals is important. Generating predicted values of the variable of interest after running the regression is very very important. So, generally we go for predict then the variable name, if you do it, once again we drawn the regression. Let it be this is the regression.

So, simply if you enter it, then we get the regression result then followed by predict, simply predict and if you take a name of that variable, since we are predicting the one better to write down with a hat. With a name underscore hat and that gives a predicted variable. So, I think this is the one, hat has already come here with the name hat. So, generating values of the residual, how to do it, with the same approach the way we have done it, is simply predict r, resid, it gives the residual details.

(Refer Slide Time: 51:06)

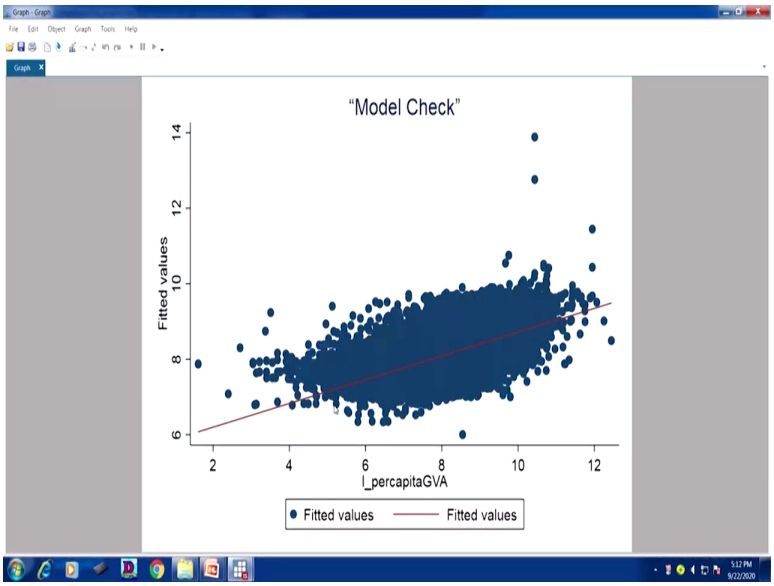


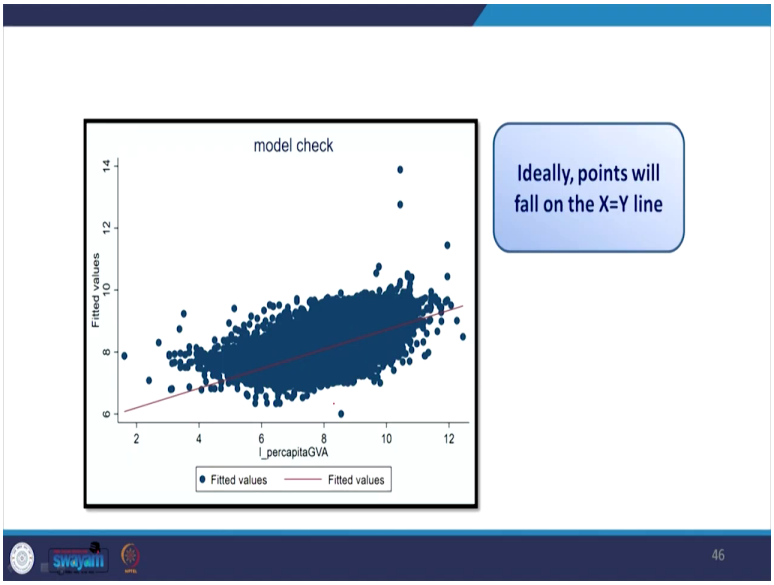




Source	SS	df	MS	Number of obs	=	36,809
				F(17, 36791)	=	993.58
Model	12243.1411	17	720.184769	Prob > F	=	0.0000
Residual	26667.5375	36,791	.724838614	R-squared	=	0.3146
				Adj R-squared	=	0.3143
Total	38910.6785	36,808	1.05712558	Root MSE	=	.85137

l_percapitaGVA	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
total_worker	-.0151178	.0016832	-8.98	0.000	-.0184169	-.0118188
age_of_enterprise	.0049973	.0006195	8.07	0.000	.0037883	.0062116
loan_outstanding	8.42e-08	8.41e-09	10.02	0.000	6.77e-08	1.01e-07
_Isector_2	.221743	.0090954	24.38	0.000	.2039158	.2395701
_Isocial_gr_2	-.4605984	.0193053	-23.84	0.000	-.4979275	-.4232692
_Isocial_gr_3	-.2812272	.0155541	-18.08	0.000	-.3117136	-.2507407
_Isocial_gr_4	-.222004	.0163777	-13.56	0.000	-.2541048	-.1899032
_Ilocation__2	.4610816	.0113383	40.67	0.000	.4388581	.4833051
_Ienterpris_2	.3523321	.01456	24.20	0.000	.323794	.3808702
_Inature_op_2	-.2909279	.0466902	-6.23	0.000	-.3824421	-.1994137
_Inature_op_3	-.6541299	.0549851	-11.90	0.000	-.7619023	-.5463576
_Iaccounts_2	-.2225837	.0177568	-12.54	0.000	-.2573875	-.1877799
_Icomputer_2	-.0936709	.0355613	-2.63	0.008	-.163372	-.0239699
_Iinternet_2	-.2515048	.0369632	-6.80	0.000	-.3239537	-.179056
_Iassistanc_2	-.0869273	.0411523	-2.11	0.035	-.1675869	-.0062676
_Iregistrai_2	-.3505737	.0136249	-25.73	0.000	-.3772789	-.3238685
_Icontract_2	.474499	.0156411	30.34	0.000	.443842	.505156
_cons	8.476285	.0536737	157.92	0.000	8.371083	8.581487

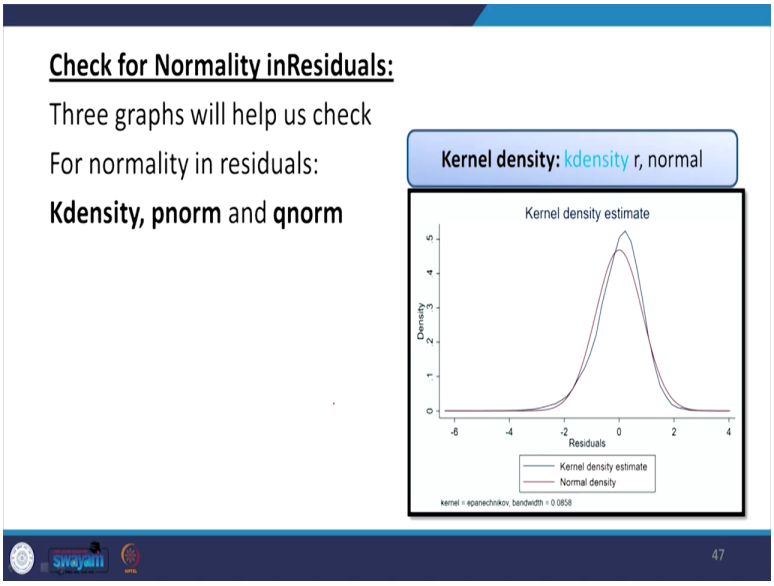


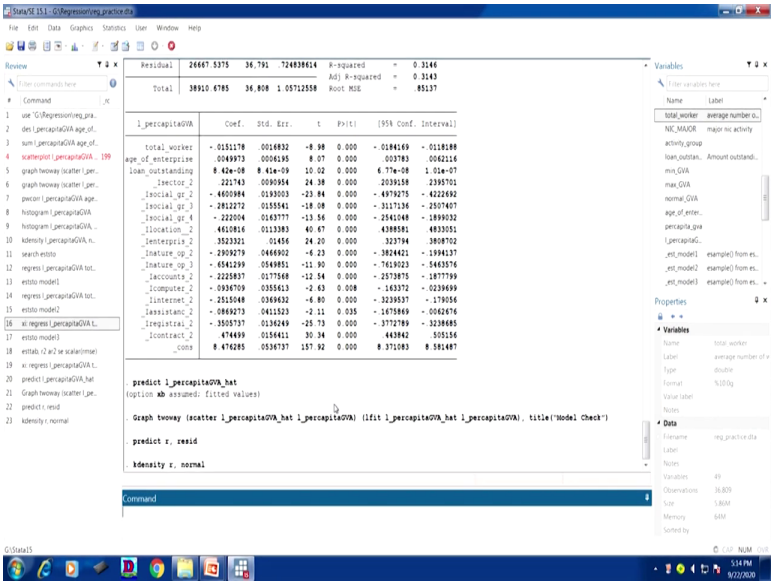


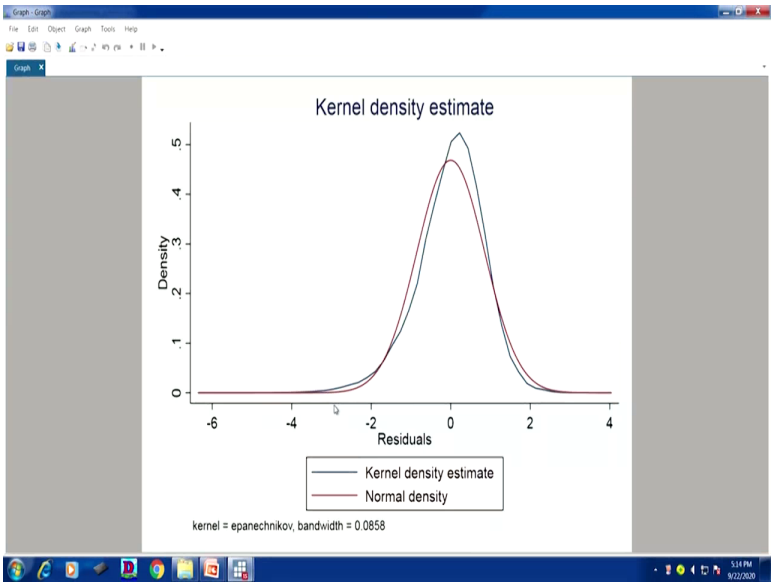
So, checking the model assumption and fit. Like the model checks require you have, checks require you have just fit the model you are checking. Save the predicted values of Y and Y-hat. We have just mentioned. Then plot observed and predicted values of Y variable. So, if you do it, then it will give, if you then copying that and operate it or if the right word to be, this not the one. after predicting we can get the fitted values and the predicted values of our model that has already been explained to you.

So what I will do, I will now explain. So that we can predict and this will look like this and ideally points will fall on the X and Y line. More falling, more points are assembled nearby that point or on that line that model is expected to be better fit.

(Refer Slide Time: 52:28)

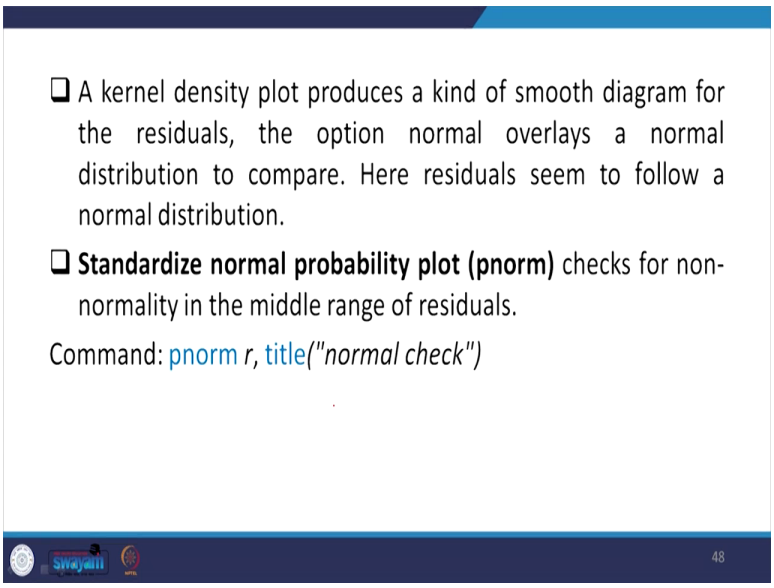


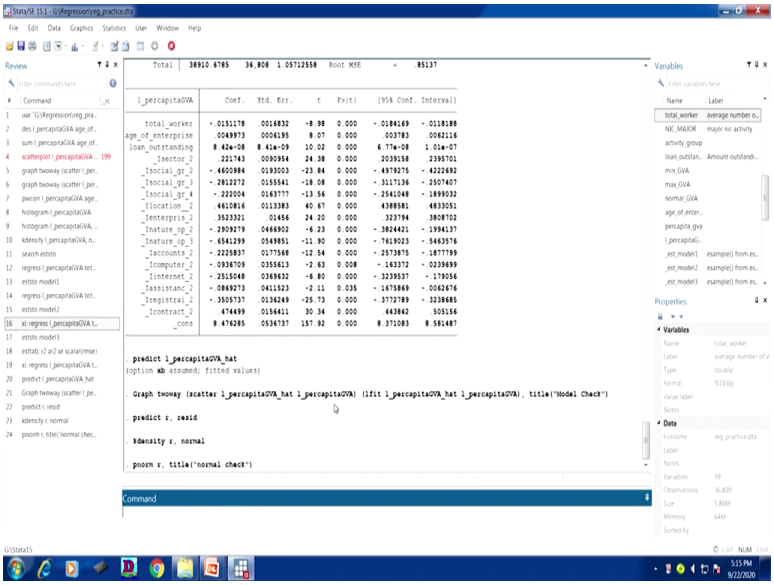


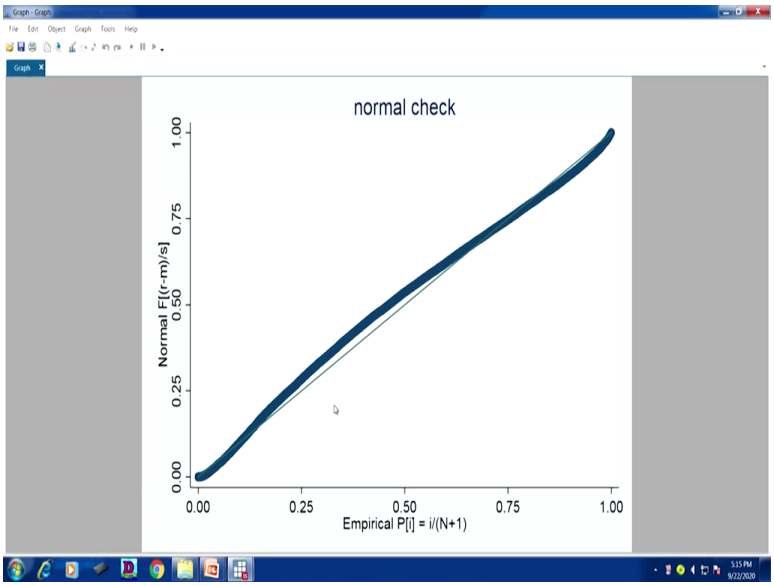


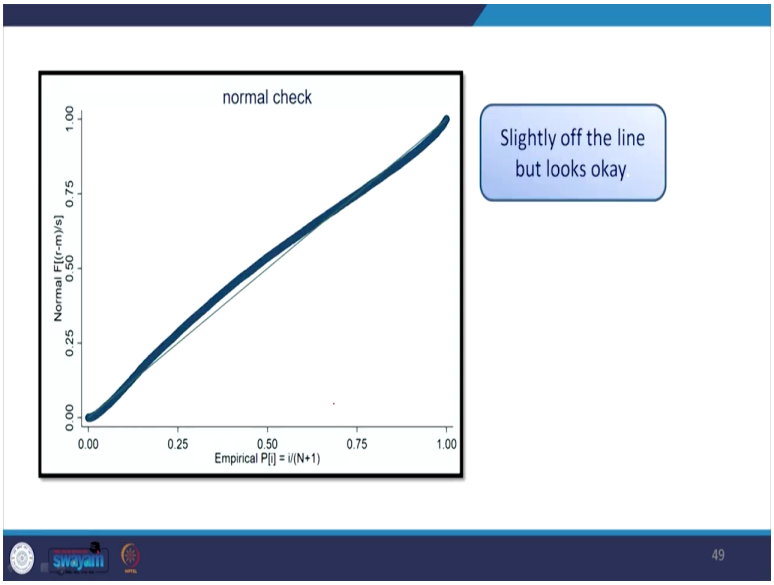
Similarly, we can check the normality in residuals. So, three graphs will help us to check the normality. The kdensity, then p normal provide normal distribution or quintile normal distribution gives the normality of the distribution. From there kdensity are normal if I just do it then probably we can able to do it. I think we need to have regression first, predicted values of the residual term, then we can find out the kdensity and the normality of that diagram. This is what we wanted to show here.

(Refer Slide Time: 53:37)




Stata/SE 15.1 - G:\Regression\reg_practice.dta
File Edit Data Graphics Statistics User Window Help
Review
Variables
Properties
Command
. predict l_percapitaGVA_hat
(option xb assumed; fitted values)
. Graph twoway (scatter l_percapitaGVA_hat l_percapitaGVA) (lfit l_percapitaGVA_hat l_percapitaGVA), title("Model Check")
. predict r, resid
. kdensity r, normal
. pnorm r, title("normal check")
Command
5:15 PM
9/22/2020

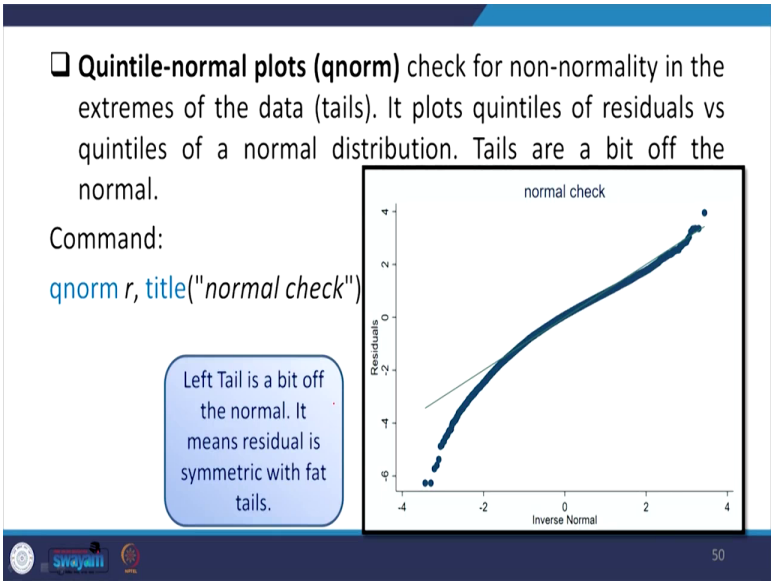

Graph - Graph
File Edit Object Graph Tools Help
normal check
Normal F[(r-m)/s]
Empirical P[i] = i/(N+1)
0.00
0.25
0.50
0.75
1.00
5:15 PM
9/22/2020



So, where is the one. This is the diagram we wanted to refer. It is approximately overlapping with the standard normal distribution. So, we can assume that our distribution is better having a normal density function.

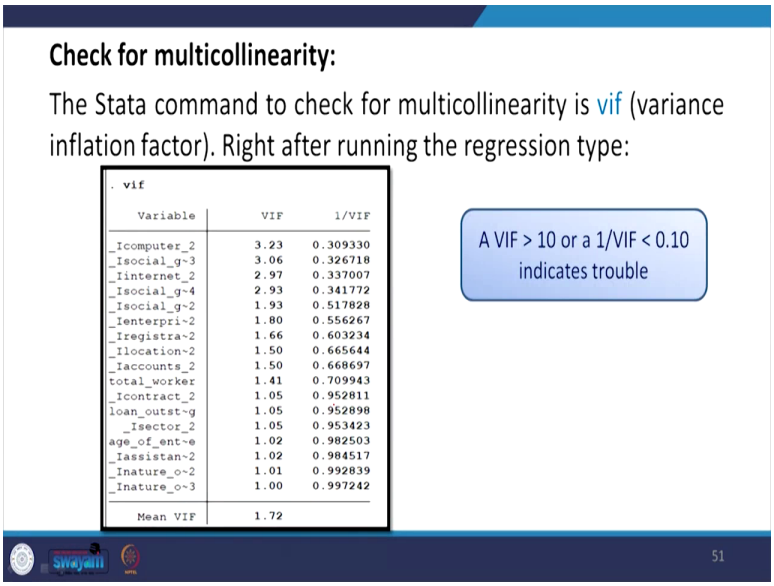
So, what is important here to note that the standardized normal probability plot that is pnorm checks for non-normality in the middle range of the residuals. So, for that you simply go by the command pnorm r and if you go by a title like this we can explain it correctly. So, it gives us the plot and it is almost overlapping with the standard the trend line. So, that will be very useful for interpretation as well. If it is completely deviating from it, then it might be problematic. So slightly off the line, but looks perfectly fine.

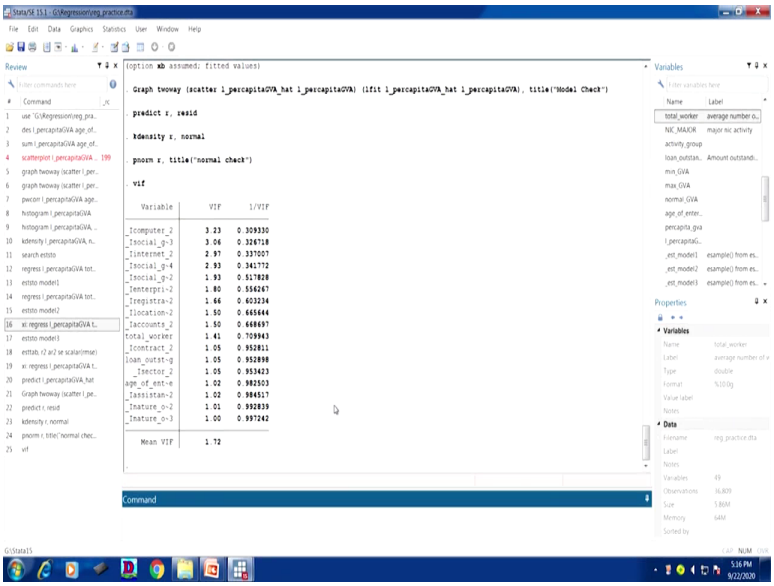
(Refer Slide Time: 54:49)



Similarly, quintile normal plots can also be derived with qnorm r, then if a title and you go and check on your own and you will certainly find out. It plots quintiles of residuals versus quintiles of a normal distribution. Tails are a bit off the normal. But still it is largely overlapping so the distribution looks better.

(Refer Slide Time: 55:13)



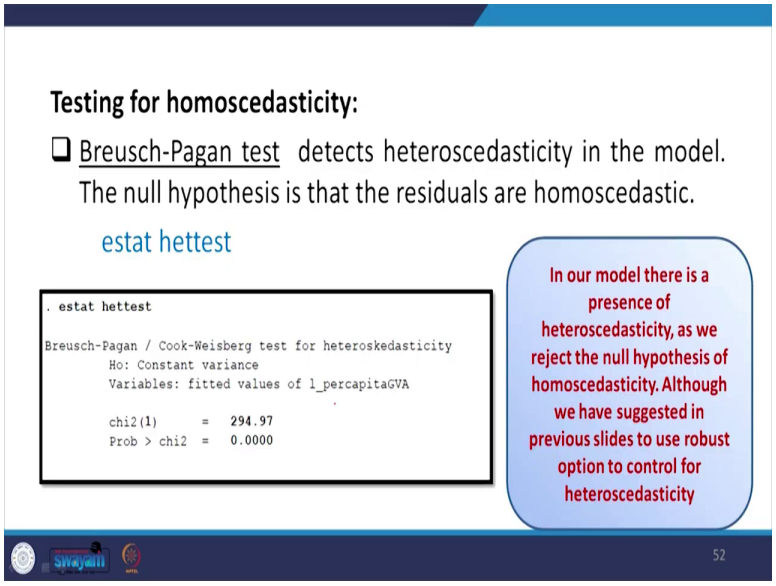


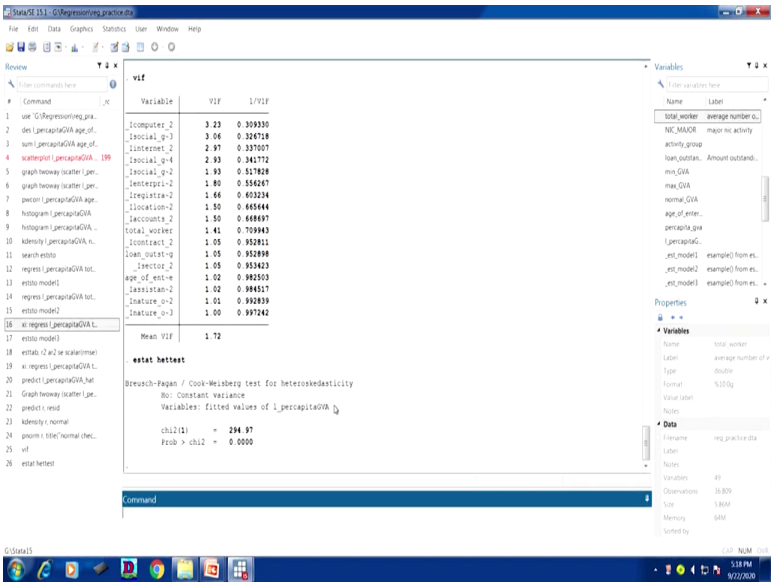
Another couple of things are there to find out. So, let me just stick to the last important guidance for you after having the result, regression result we need to understand whether there exist multicollinearity, whether there exist relationship between the variables or not, whether there exist perfect linearity between the independent variables or not, for that we need to check the VIF function. So, VIF is going to give us. So, VIF if it is less than 10 then it is perfectly fine. So, if it exceeds 10 then that is problematic.

So, if VIF is exceeding 10 or the inverse VIF that is variance inflation factor if it is less than 0.10 inverse that is 1 upon VIF that is inverse of it, if it is less than 0.10, it indicates a trouble that means there exists multicollinearity and we need to check accordingly.

So finally, the last one to be guided in this session is testing of homoscedasticity. The Breush-Pagan test detects heteroscedasticity in the model. The heteroscedasticity we told you that the variance is having sigma i square or the variance of the error term varies. But our assumption is that there should be homoscedasticity.

(Refer Slide Time: 56:53)





So, the estat hettest is going to give us, this is like it. So, estat so it has given us the interpretation, for interpretation our assumption is that the null hypothesis, there should be constant variance or homoscedasticity. But we have derived that our result is significant that means we are rejecting the null hypothesis that means our data is heteroscedasticity.

And if there occurs heteroscedasticity the model requires further check. In our model there is a presence of heteroscedasticity, so we reject the null hypothesis of the homoscedasticity, although we have suggested in previous slides to use robust options to control the heteroscedasticity. So,

that is all the guidance so far. In this week we have covered everything about linear regression model. We discussed so many things. I think if you go carefully and operate on your own from the slide with the sample data you will enjoy in reading the slides. Thank you so much.