Exploring Survey Data on Health Care

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

Lecture - 26
Regression Models of Quantitative Healthcare Variables

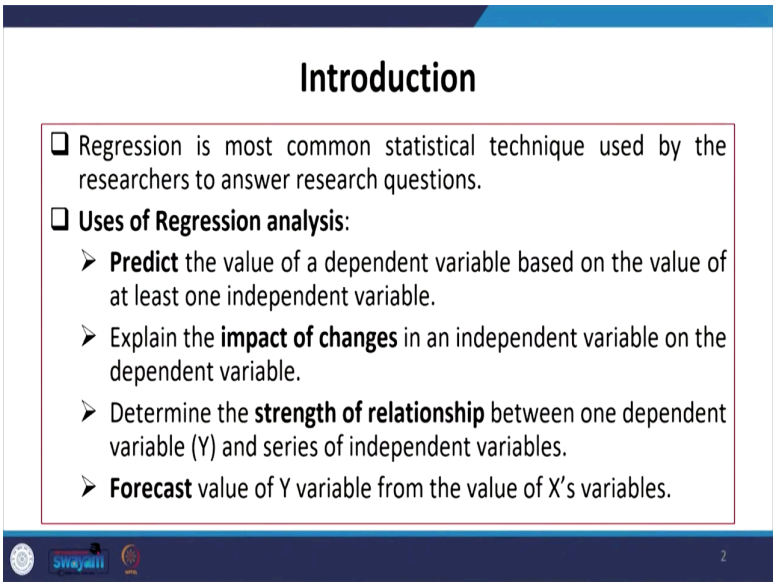
Welcome friends to this NPTEL MOOC module on Exploring Health Care Survey Data. We are on the particular week explaining n influential statistics in health care. The first lecture is on understanding regression models specifically on the quantitative and qualitative variables that are used in models. In this particular lecture, we will be emphasizing only the quantitative variables, in our next lecture will be on qualitative variables.

Again, you might be having some sort of problem with whether it is a quantitative dependent variable or quantitative independent variable or qualitative dependent variable, or qualitative independent variable. We will give all sorts of clarifications to you so that your understanding in this regard will be much crystallized.

Without discussing the background of it let us move on and clarify what is all about. And starting from the very first aspect of regression as we all know that this is quite common, in different statistical techniques as researcher used to answer say several questions.

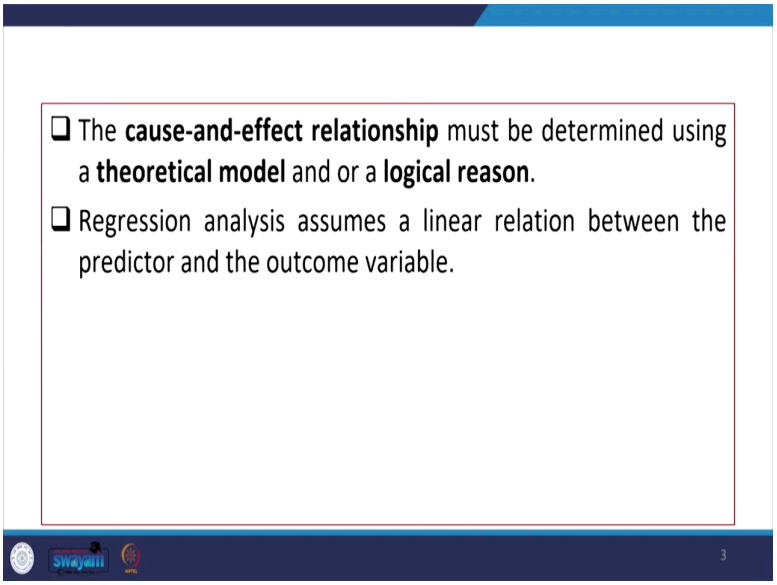
The regression results and their analysis have various uses, especially since it gives certain predictions about the value of the dependent variable based on at least one independent variable.

(Refer Slide Time: 02:07)



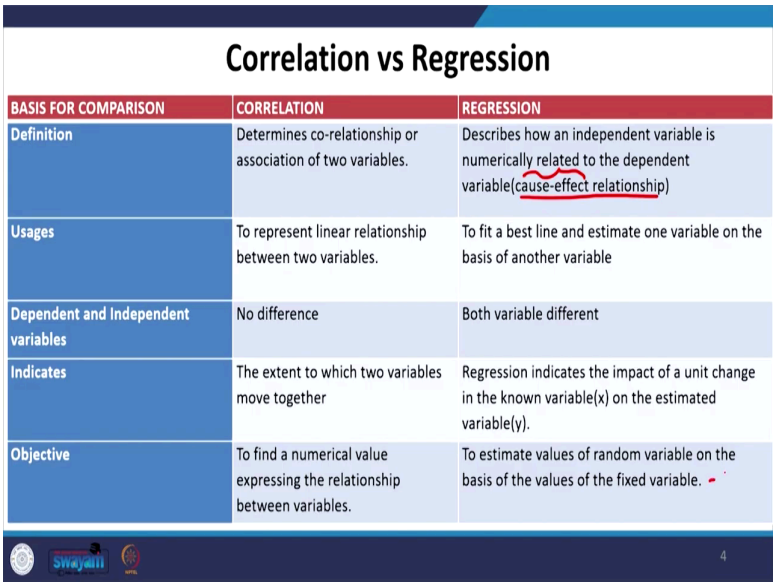
This explains the impact of changes in an independent variable on the dependent variable. This helps in determining the strengths of the relationship between these two variables X and Y, also it helps in forecasting the value of Y based on the control variables or X variables.

(Refer Slide Time: 02:28)



This gives cause and effect relations though the causal influences cannot be drawn, certain logical understanding through the theoretical model can be mapped through the regression coefficients. And regression analysis assumes a linear relationship between the predictor and outcome variable.

(Refer Slide Time: 02:46)



We need to first establish the relationship between correlation and coefficient by their definition usage which kind of variable they consider what are the indicators of the objective etc. Starting with the definition of correlation we know that there exists a co-relationship or association between two variables. Whereas, in the case of regression this describes how an independent variable is numerically related to the dependent variable.

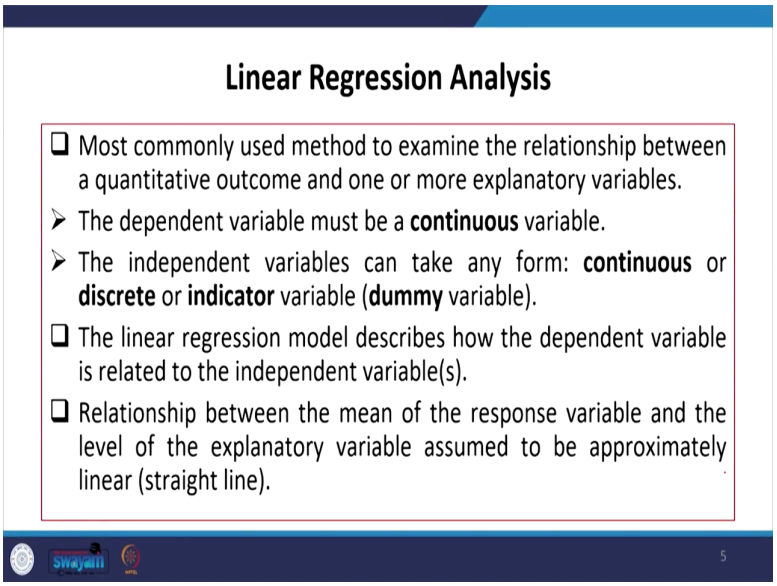
With a certain logical relationship between these two. Now, you might be confused about whether it gives cause and effect, in reality, it does not give perfect cause and relations or effect. It gives some understanding related to why it has happened and how it has happened, how it is relevant? But the exact causal conclusions are not derived through regression.

But it is for sure how much it is affecting, what is the quantification of it can be derived from the regression. Correlation and its uses are like it gives you linear relationship whereas, a perfect fit line or trend line (best-fit trend line) could be derived, based on the variables and their relationship with another variable.

Then in correlation, the perfect trend line is not derived. Regarding dependent and independent variables, correlation does not differentiate these two. Whereas, in the case of regression we have to differentiate then only we can find out the impact of one on another one. Containing the indicators of correlation and regression the extent to which two variables move together is called correlation.

Whereas in the case of regression, this indicates the impact of a unit change in the known variable on the estimated variable. The objective of correlation is to find a numerical value expressing the relationship between the variables whereas, in the case of the regression analysis this gives estimates and their values of random variable on the basis of values of the fixed variable.

(Refer Slide Time: 05:14)

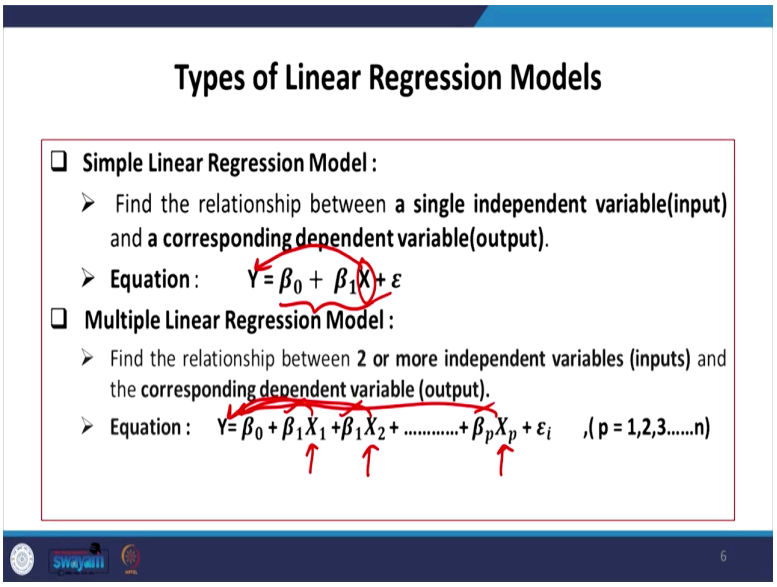


We are now attaching information about linear regressions and their analysis of how these are read. This is the most commonly used method to understand the outcome variable and the explanatory variable. Then the dependent variable must be a continuous one. If it is not a continuous one, then we will discuss some forms of non-linearity or some transformation to a linear model.

That we will discuss in our successive models. At this moment we are saying that the dependent variable must be continuous. The independent variable can take any form and may be continuous-discrete indicator type like in dummy variables etc. The linear regression model describes how the dependent variable is related to the independent variable.

The relationship between the mean of the response variable and the level of the explanatory variable is assumed to be approximately linear. So, that is why it is called linear regression analysis.

(Refer Slide Time: 06:18)



There are different types of linear regression models; one is called simple linear regression another is called multiple linear regression model. Simple where we also call it as bivariate regression model other one is called multivariate regression model. Variate means the extent of variability due to the number of factors including a number of variable factors including, independent factors included.

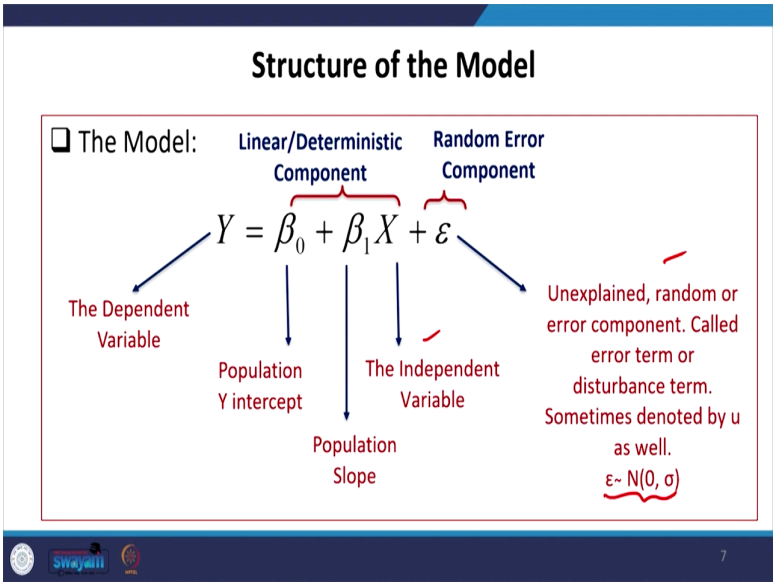
At this moment there are only two factors in fact of X and Y. But here there are so many factors. So, that is why even if we are saying X is in fact a vector of information of control variables. So, $\beta_1 X$ the equation which we have mentioned here is a kind of vector of dependent and independent variables. And this equation is best explained in Greene's book of Econometrics.

Where the matrix method how is explained, how different vectors are explained in the shaping the equation like this are well explained you can have a look and find out the exact reasoning behind writing on $\beta_0 + \beta_1 X + \varepsilon$.

Epsilon is the error term. We know that the X cannot explain everything about Y. There must be some other component called some fixed component as explained by β_0. And there will be still some unexplained aspects captured through the error term.

In the case of multiple linear regression, we have to discuss so many other independent variables along with the first one and we can be able to estimate each of their coefficients differently and that will be helpful in interpreting the result correctly.

(Refer Slide Time: 08:34)



Let us understand the structure of the model. As we know that Y represents the dependent variable, and β_0 is our intercept or the fixed coefficient, which is independent of the change in the covariates. And likewise, we can give an example of consumption as a function of income.

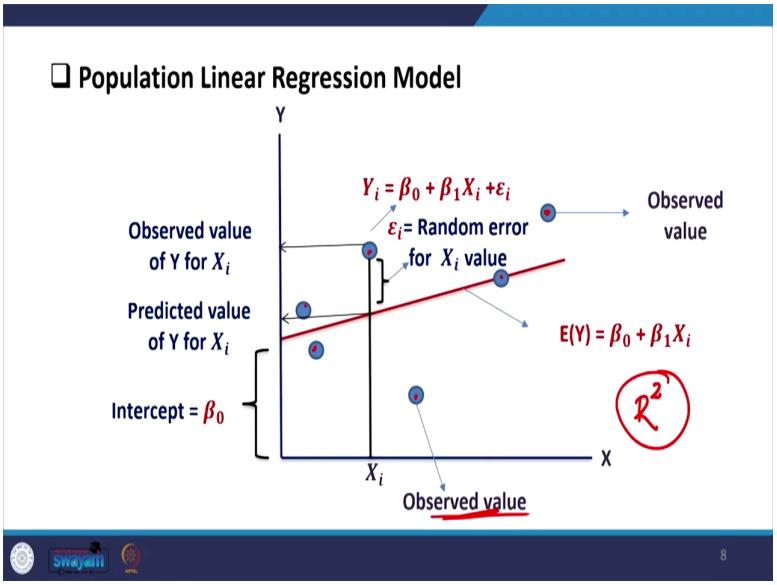
And other factors if I write down C as a function of income when income tends to 0, C may not be tending towards 0, there would be some positive consumption even if income boils downs to 0 for a particular consumer. The particular consumer or household may not get income in that particular month where they are supposed to consume something by borrowing or by other means of getting the food.

Therefore, consumption there must be a fixed component. Similarly, almost all the variables in the world that we are trying to project must follow this kind of structure. Other details we have already highlighted, and I am sure this will help you. Epsilon is called the random error component, which is also called the unexplained or random or error component.

This term is called error term and or also called disturbance term. Sometimes denoted by U and that is in fact distributed normally with 0 and 1 standard deviation. This shows the

normality of the distribution, that all sorts of projections in the model can be made. We will explain some of these things in this particular lecture.

(Refer Slide Time: 10:31)



Population linear regression model how is explained in this diagram you can see a trend line it is fitted with that is simply called the expected Y value is equal to β_0 plus β_1Xi. And β_0 is the intercept and any dots you could see in this particular diagram, dots like here some of the dots are deviating from the trend line.

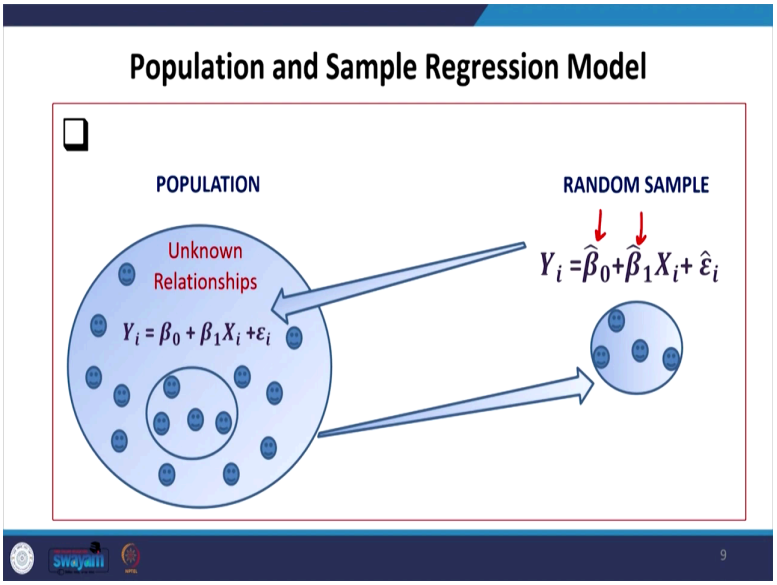
And the trend line is explained in the expected value given they are control variables. Now our purpose here is to identify the extent of error, and the differences of each value from the standard line or from the trend line, or from the fitted line. So, it is explained here, and these are also all dots that explain the exact information, or they are also called observed value, we have also highlighted that.

How the observed value is actually deviating from the predicted value of Y for Xi? if our observed value is almost on the trend line or on the predicted line; that means, our estimation is actually expected to be very good in that case. Your R square is expected to be quite good, so it will be close to 1.

But in most the cases like in survey data R square value as we already explained earlier in our previous lectures that it is usually lesser. You need not worry much by we have given all reasons behind R square and its range which one could read how it should be interpreted.

Though this is what we already explained in our previous weeks' lecture. You have to follow and find out the differences.

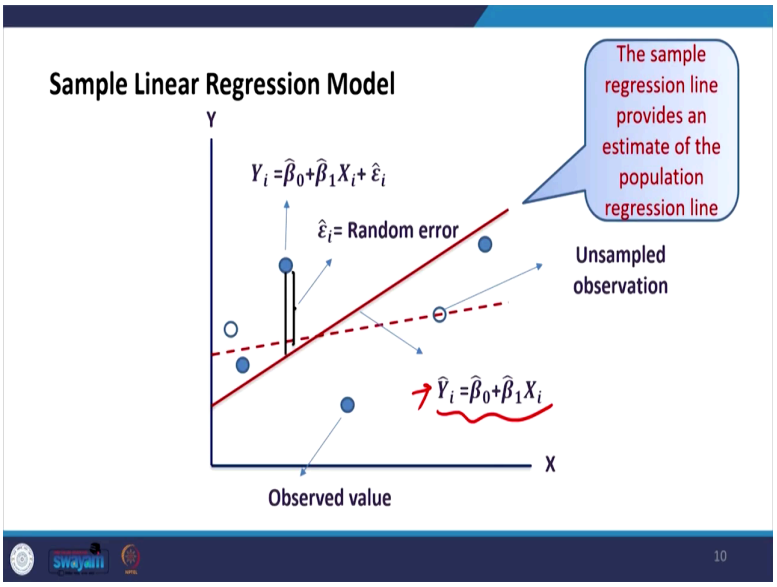
(Refer Slide Time: 12:39)



Now here we are presenting a difference between the population and sample regression model. A population where we have written the equation Yi is equal to $\beta_0+\beta_1$Xi+ε from here a random sample is taken. That random table is going to give a sample series of data.

That series we can estimate, and here is our estimated β values, and that is going to project the population β values and if it is correctly projecting then our sample is in fact giving the right result in terms of predicting the population information. This is what we usually follow in our sampling design and the rest of the details are very much clarified from the diagram itself.

(Refer Slide Time: 13:38)



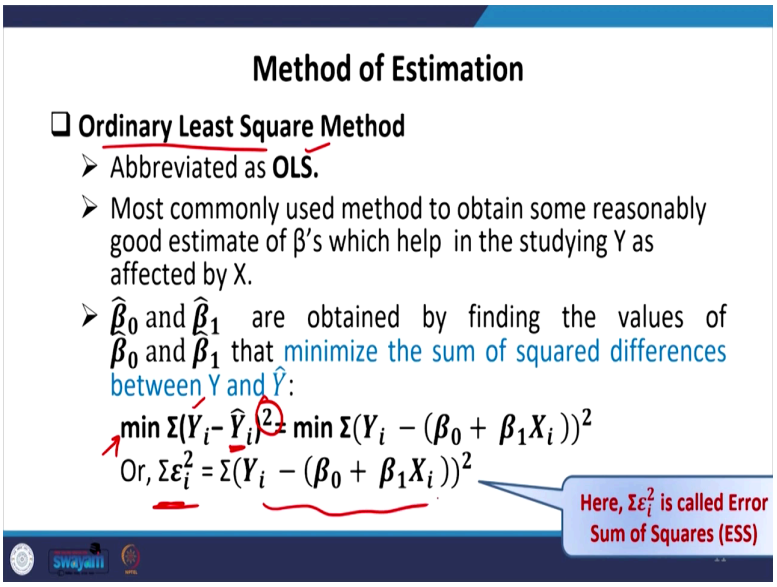
Now, we are explaining sample linear regression we have already explained the population linear regression in this chart and this table here is sample linear regression. Therefore, the population now you can see how on sampled observations, we have we are just highlighting sample. The sample linear regression line provides an estimate of the population regression line and these are all going with the estimated value.

These are in fact called estimated values, one that is written; that means, they are the estimated values based on the sample data we have collected. So, those hat is in fact useful in predicting the population values. So, like we take so many varieties of samples those have different distributions.

In each case, we will establish a number of tests, whether the sample is actually catering to the or in fact representing the population. So, whichever is estimating quite well, then our regression model in fact predicts the population very correctly.

Sometimes we take the number of a different sets of samples if each of the estimated values are not varying much; that means, our estimation is in fact robust. Those techniques we do. I am not emphasizing much here. Now, coming to the method of estimation there are different methods of estimation sometimes use maximum likelihood estimation as well, but at this hour we are explaining about ordinary least square method OLS.

(Refer Slide Time: 15:39)



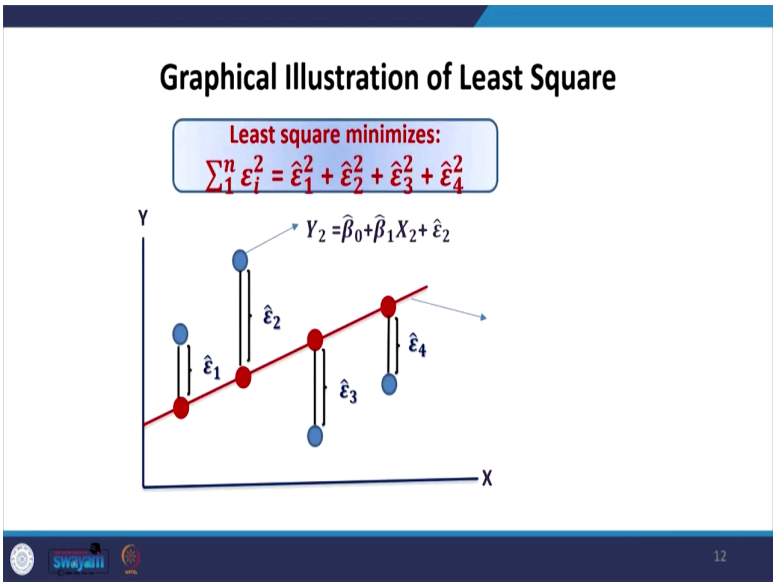
One of the methods of estimation through OLS; OLS is everywhere deviated and is used frequently by researchers. This is in fact as I already said the most commonly used method. And estimated values represented with they are hat beta β_0 or β_1 hat. They are determined based on the error least square error.

So, now you can see the estimated value is presented here and this is in fact each of the observed values, the observed value minus the estimated value the difference is in fact called an error. So, that is Ei ε_i. ε_i and its square are taken wherever we can minimize the errors. So, that eventually indicates our estimation is going to be fitting with the trend line. Therefore, minimizing that error is important.

Now, if you will not have taken it might be the case that the minimum is fine, but if you do not take this square, it is very difficult to minimize the error. Because sometimes some values of this might be positive some, if it is not having square term then the some might be negative, the average might boil down to be zero. So, then in that case estimating or minimizing the value nearest to the estimated value is in fact very difficult.

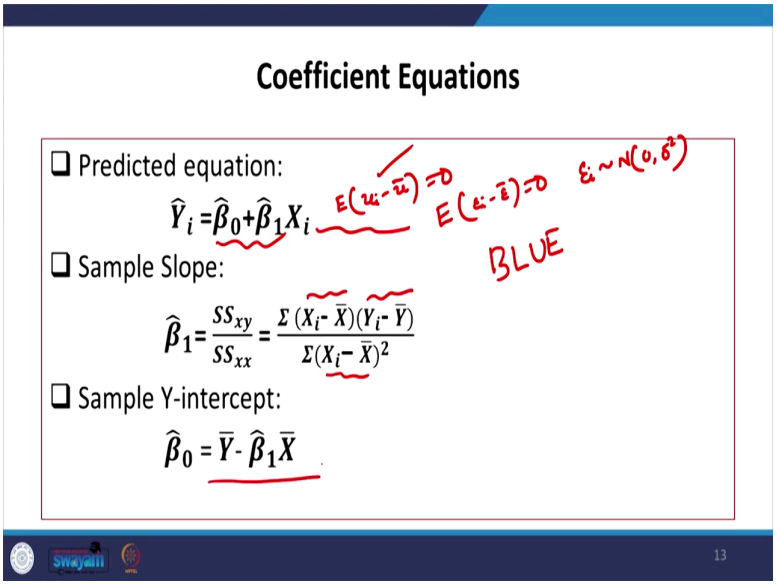
Squaring is going to give us non-negative values. Then we have options to minimize it as close to the 0-value as possible. That is why it is called squaring the error term. That is why it is called the least square. We are trying to understand which one is in fact giving us the least error. So, that is why this is called the sum of ε_i^2 and then these are estimated like this.

(Refer Slide Time: 17:58)



The graphical illustration is given in this particular picture. There are different ε_i ε_{i1} hat ε_{i2} hat ε_{i3} if all these are in fact estimated and derived with the lowest levels, minimized level with their square term is called least square error.

(Refer Slide Time: 18:19)



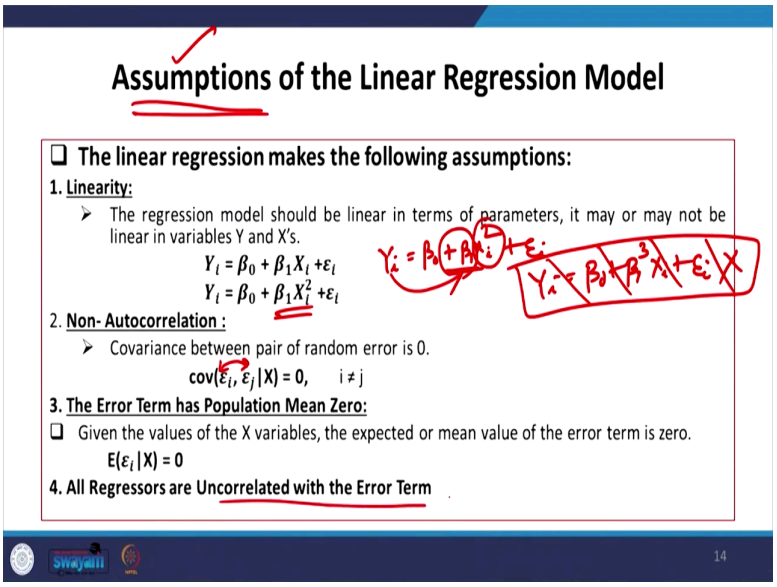
The next one is coefficient equations. What do we mean by this coefficient? First is the estimated value of ε_i basically is estimated. So, the estimated value of the error term is missing here because the expected value of ui minus u bar is equal to 0 and then you can say

that ε_i minus ε bar is equal to 0 because of the assumption. The assumption is that our distribution is normal; since we have already said this is distributed N with $(0, \sigma^2)$.

Therefore, this is what we all already said. When this assumption is fulfilled; that means, the error term is estimated to be 0. Whatever is left is in fact the estimated value in terms of β_0 and β_1 hat. now, based on this we can estimate finally, the value of β_1 hat and β_0 hat. you can just see further on β, how β values fulfilling BLUE properties or not, whether best linear unbiased estimation matters or not.

So, this is estimated as β_1 hat is derived as the sum of the square relationship between X and Y that is the covariance divided by the variance of X. So, covariance basically is the relationship between Y and X divided by it is the sum of squares of it is own variance. Then once the β hat is determined then we can estimate the β_0 hat as well.

(Refer Slide Time: 20:08)



The assumptions of the linear regression model are very important, and you must take note of them. Each of it is very important to validate your model of OLS whether it is in fact applicable or can be applied in many contexts or not. So, the first assumption is that the linear regression makes the following assumptions. The first one is called linearity the relationship between the dependent and independent variable must be linear.

Linear by its coefficient not by its variables and linear by its parameter not by its variables. So, variables could be like this, Yi if I said, so it should be if it is β_0 then β_1 it could be of Xi

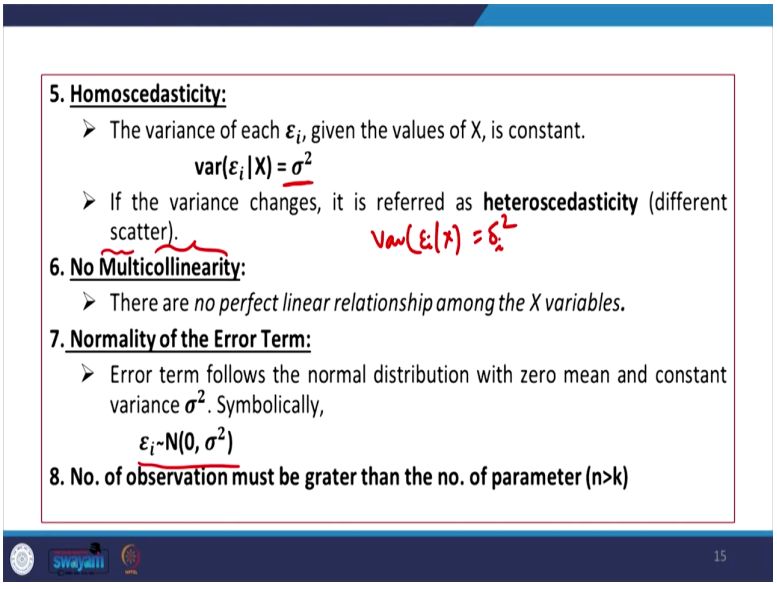
square and ε if I take this as a modified equation because linearity stills hold with the parameters.

We are estimating, this is explained here already and if it is Yi is equal to β_0 plus β_1^3 then Xi and ε this is in fact not a linear model. This is rejected at this moment, we are not estimating a non-linear model. The next one is called non-autocorrelation. Autocorrelation the word itself says it is having a covariance between pairs of random error. So, error inferior period over the time if there has been some continuity of the data. The errors are actually having a certain path of progress.

In one period there are set of errors, in the second period there are again sets of distribution of error. There are those errors, if errors i and j, given the control variables are X if it is equal to 0. That means there is no covariance. So that means, the error terms are considered to be and have no autocorrelation, if you are having autocorrelation this will be obviously non-zero.

So, the error term has a population mean of 0, the error term should be normally distributed as we already said, especially its population mean i.e., for the entire population. The expected value of the error term given its control variable should be equal 0, all regressors are correlated with the term with the error terms. So, the error term should be uncorrelated this is what we wanted to emphasize here.

(Refer Slide Time: 23:14)



The next assumption is on homoscedasticity, which simply says that your variance of the error term which we have been emphasizing those variances must be constant that must not be different every time. So, it should be constant this is what is explained with sigma square.

Once it is constant; that means, we are assuming we have assured ourselves that our distribution is not going to be changed much. So, our prediction is going to be better. So, as against homoscedasticity, in that case, your variance is going to be explained as sigma i square.

Therefore, i stands for some variability and it is different in different scatter plots. if you take several samples out of your population your variance is going to be different every time. If it is not varying much near about the same value; that means, that the data is in fact following OLS assumption or is more or less normal.

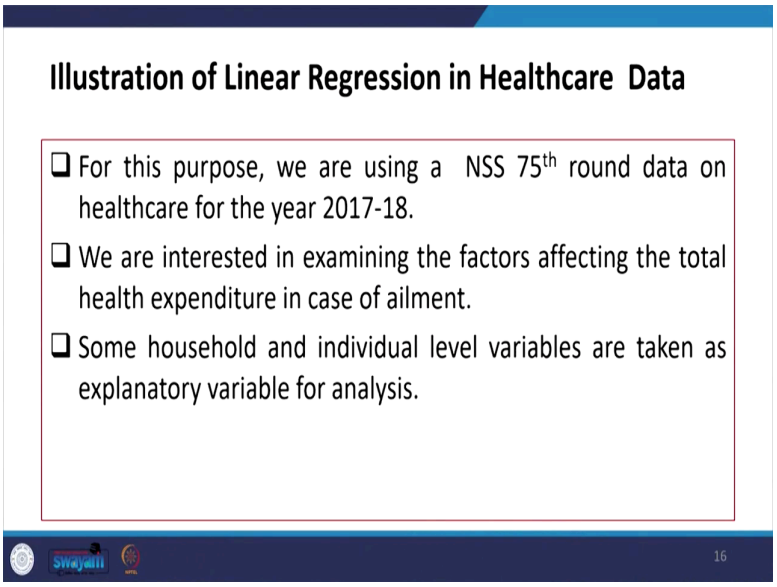
What about multicollinearity? Multicollinearity is basically the word that says collinearity. Therefore, how linear collinear; that means, more than one variable the explanatory variable. If there are multiple linearity relationships; that means, we simply say multicollinearity. In other words it says that there should not be any perfect linear relationship among X variables.

They are the among the explanatory variable if there exist some linear relationship or perfect linear relationship, then we are actually violating. OLS is not going to be an efficient because from one β value you wanted to predict something about the Y, but in fact that β value is not independent, it is explained by another independent variable.

So, your model is not correctly estimated. Therefore, we have to assume the model with no multicollinearity. Next is about the normality of the error term where the error term is in fact distributed like this that I have just explained. So, it should have 0 mean and constant variance and the next assumption is on a number of observations must be greater than the number of parameters to be estimated.

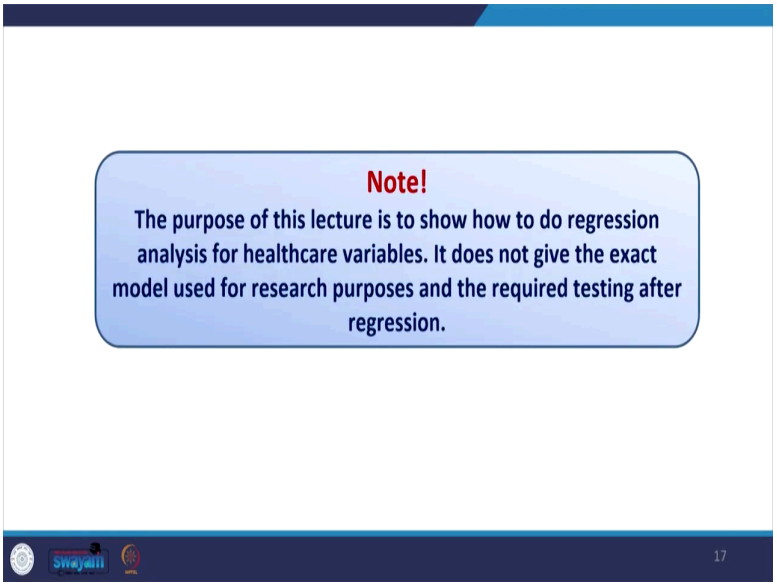
Like beta 1, beta 2, beta 3, beta naught etcetera number of parameters must be less than that of the observations we have considered. Otherwise, the wireless estimation cannot able to predict, predict correctly. Next to guidance is on illustration of linear regression in health care data.

(Refer Slide Time: 26:25)



For this purpose, we are using NSS 75th round data on health care for the year 2017-18. We are interested in examining the factors affecting the total health expenditure in the case of element or element persons. Some households and individual-level variables are taken as an explanatory variables for the analysis.

(Refer Slide Time: 26:45)

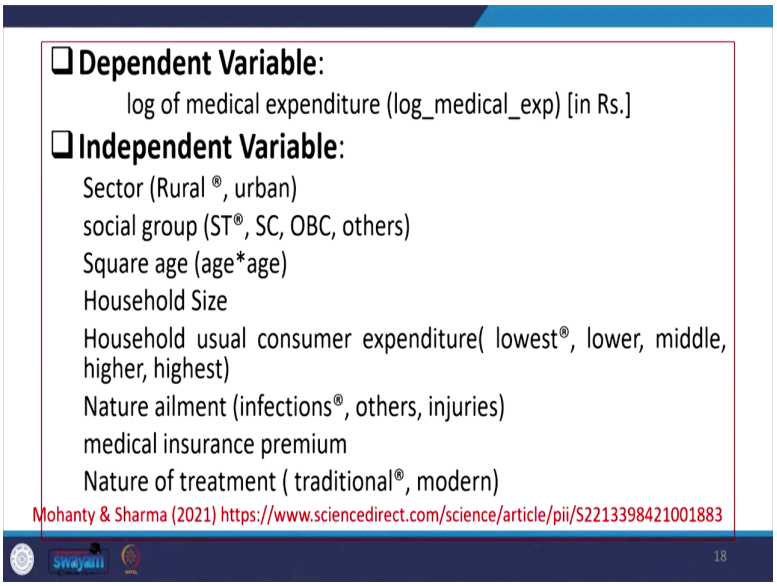


One note is here for you that the purpose of this lecture is to show how to do regression analysis using health care variables, it does not give the exact model used for research

purposes and the required testing after regression. Therefore, we are just giving some sample set of explanation and practical results that you may not directly use for writing papers.

Therefore, you need to understand the concept and develop your own model. Now here the sample variables we have filtered from the 75th round of the national sample survey. The dependent variable we have considered for using the ordinary least square regression model is medical expenditure.

(Refer Slide Time: 27:34)



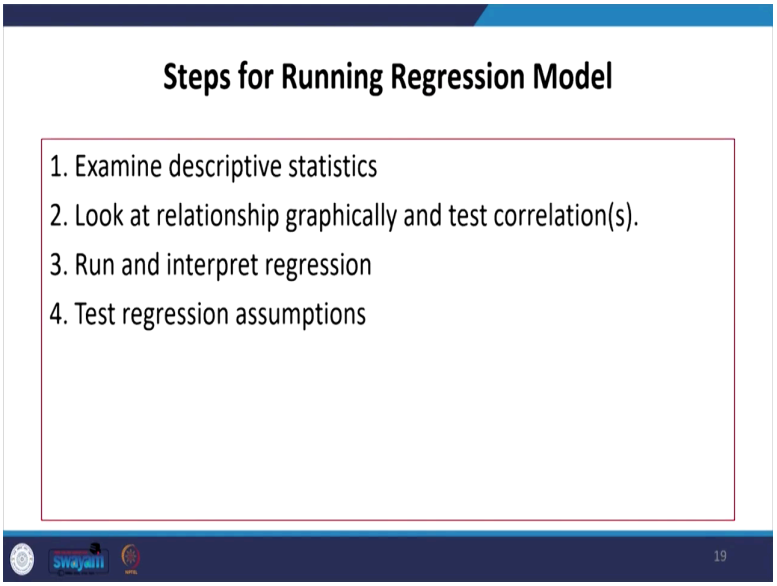
And medical expenditure we have transformed to a log medical expenditure the log transformation is taken, to make it more rationalizing. Because usually, expenditure as a variable is more skewed. Therefore, expenditure is overstated whereas, income is usually understated.

Therefore, when expenditure is overstated a transformation log transformation is going to nullify and simplify the data to make a better distribution. Similarly some of the relevant independent variables we have noted down like sectored dummy rural with any so urban categories then social groups, age square we have taken age into age, household size, household usual consumer expenditure and nature of the element, medical insurance in their premium, nature of treatment etcetera.

Even you can refer to some of these variables in one of the models we published this is our paper published this year just recently in the journal on the repository of a science direct. So,

clinical epidemiology and global health journal. You can just follow and I have highlighted this on the screen just click on that link and I am sure you will get the exact paper for your reference.

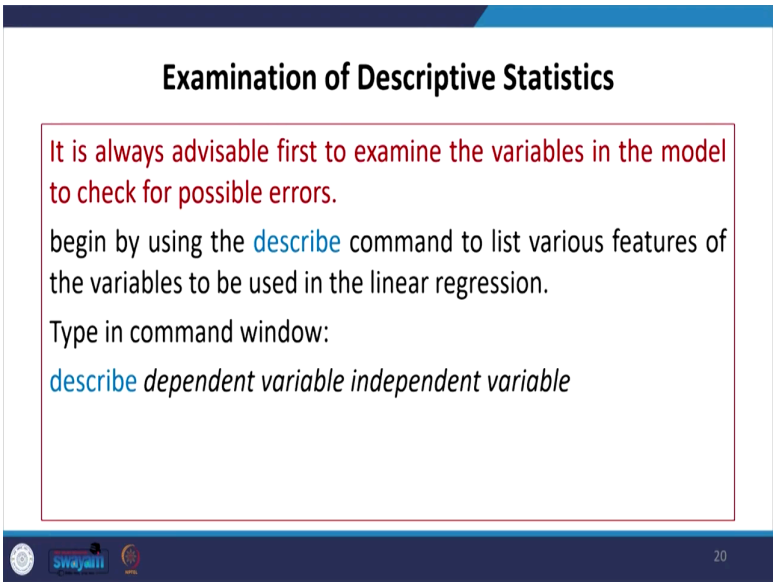
(Refer Slide Time: 29:01)



Further this will also explain about the non-linear qualitative dependent variable models in the next class and you will understand things better. Now, we are explaining the steps for running the regression model. Steps are here like we will examine first of all they are statistics the basics descriptive statistics then we will understand them graphically and also understand their correlation.

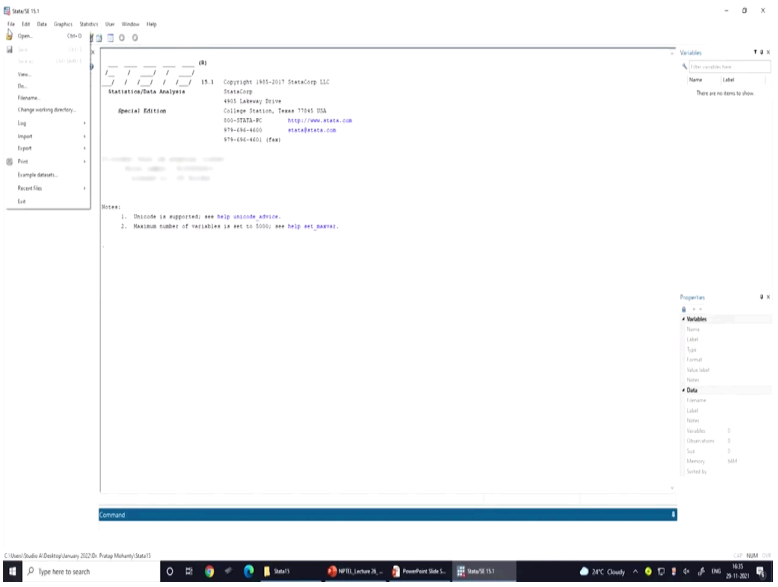
Then we will draw and interpret some regression results, then regression assumptions to be tested whether the regression was actually drawn correctly or not. These four important steps all must be mandatorily followed to write an article research paper. Examination of descriptive statistics first is on descriptive statistics. I will simultaneously open the database in front of you and then operate it.

(Refer Slide Time: 30:14)



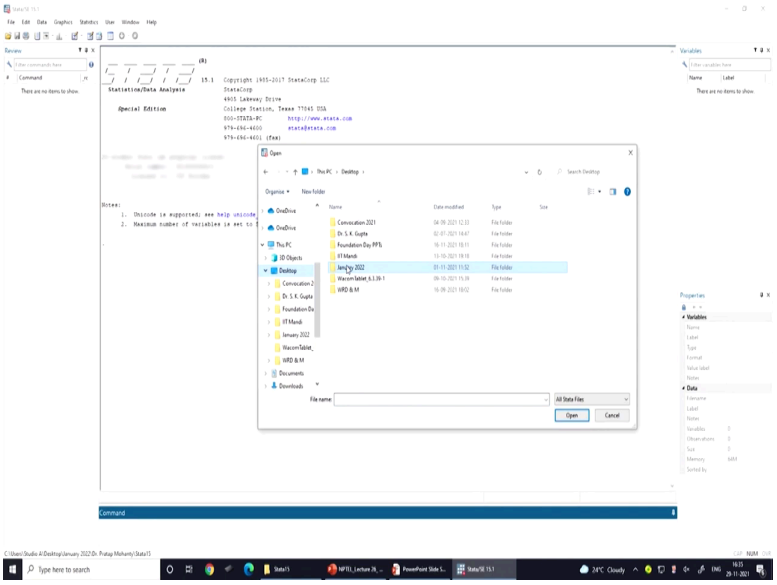
The rest of the result I will leave at the end I will simply leave to you and I am sure you going to enjoy it, but let me now open a database in front of you.

(Refer Slide Time: 30:22)

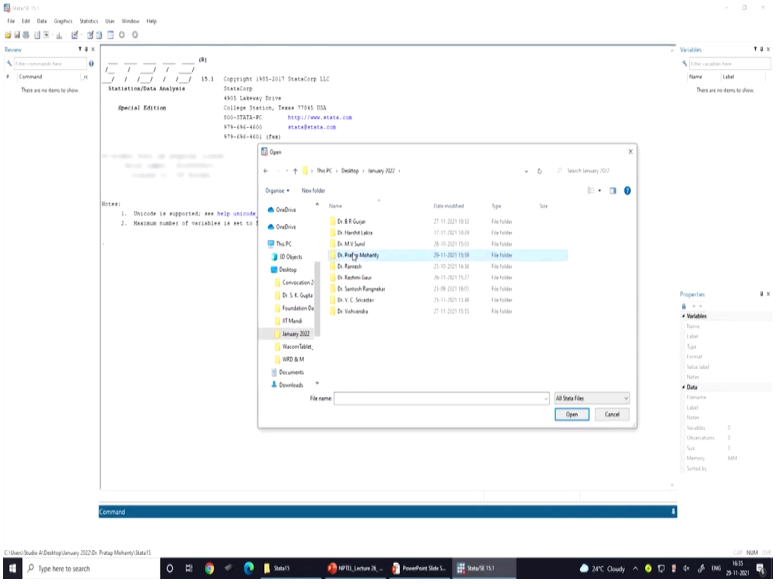


I am just going to open the sample data.

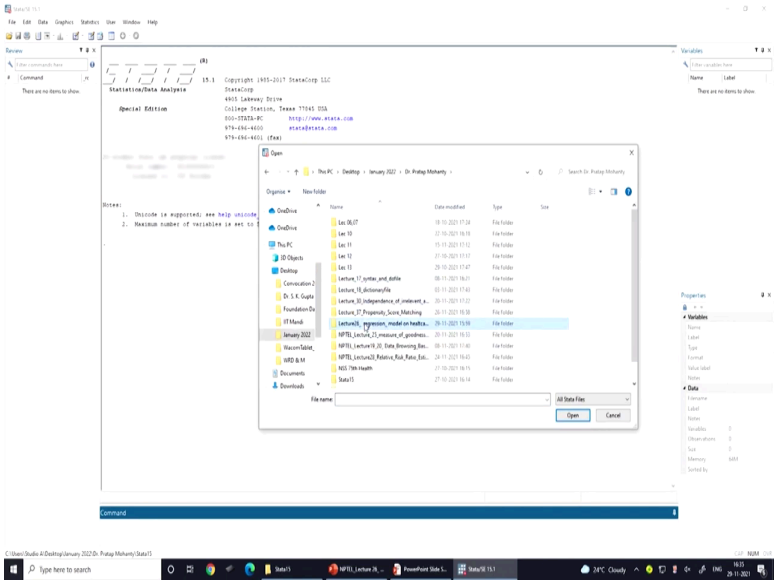
(Refer Slide Time: 30:24)



(Refer Slide Time: 30:28)

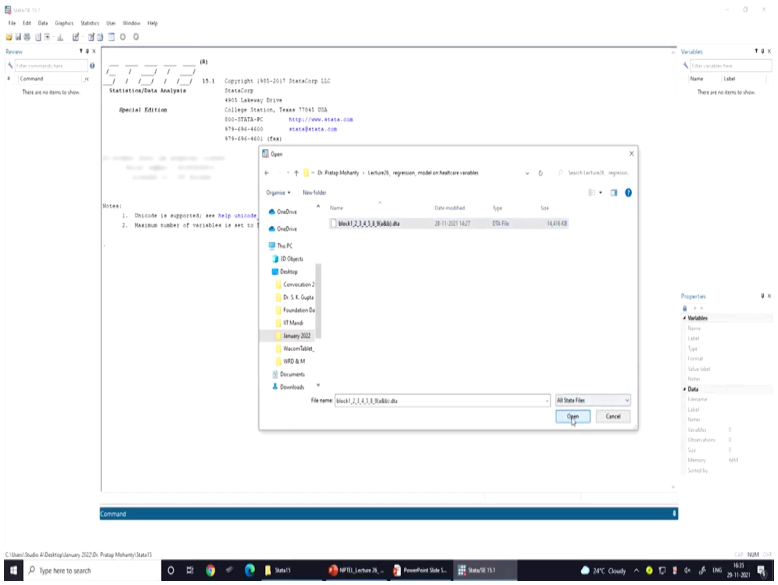


(Refer Slide Time: 30:30)



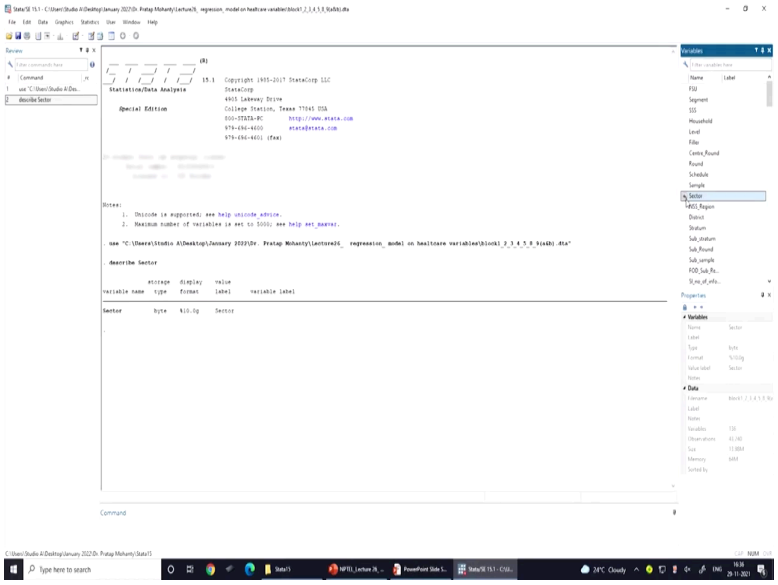
Which we have kept for this purpose.

(Refer Slide Time: 30:33)



And I am and we will experiment with this some of the variables derive results.

(Refer Slide Time: 30:35)

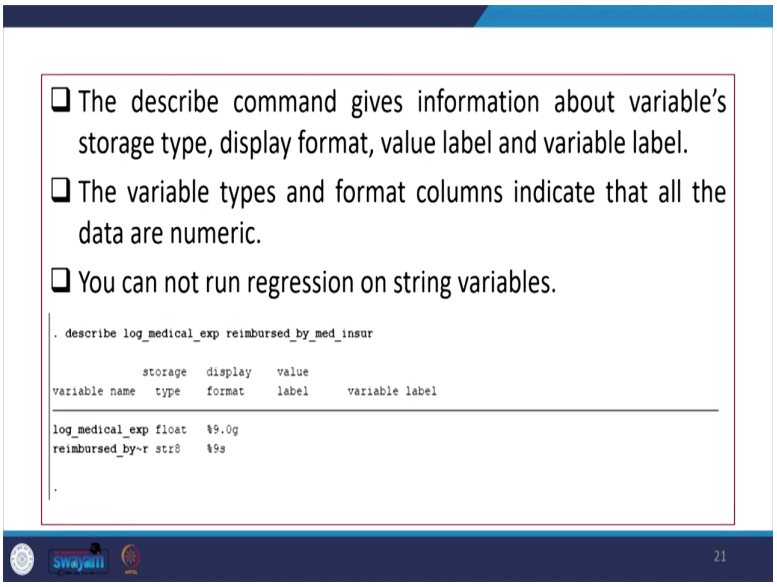


Now what I will do? I will first go through explaining describe; describe command that will give you the basic description of the variable. I will write down des or describe in full that will give us the correct result. Any variables I can just write it down.

Sector I have taken here alright. So, now, the describe I have entered with that variable. Now, this has given me information about what is the storage type of this particular variable and what is the label of this, since label are not defined correctly what categories it has. So, label it is not showing in detail that we will show it later. But at this moment I will explain it differently.

Similarly coming to another command like; so, like in the describe command once again let me explain that this list various features of the variables to be used in the linear regression. So, by typing describe all your set of variables may be starting with dependent independent it will give you the complete list of which is the description.

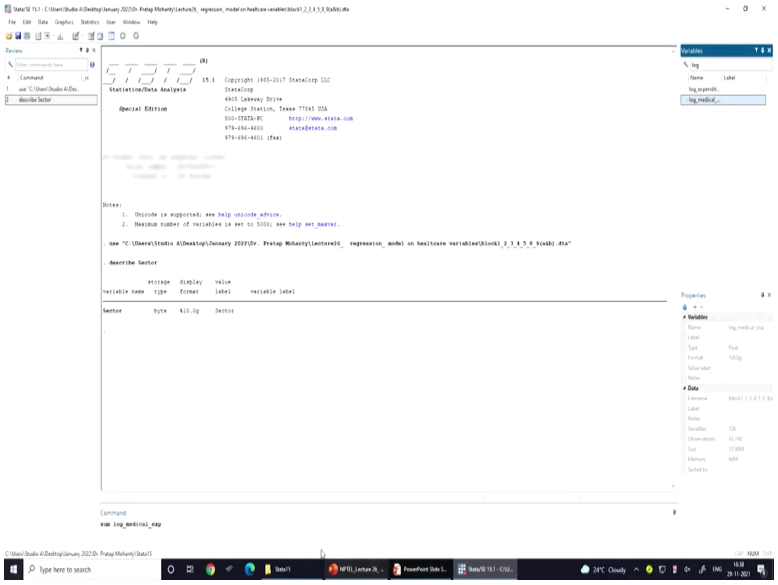
(Refer Slide Time: 32:03)



Like here if you have two variables one is dependent that log medical expenditure then the second one is reimbursement of medical expenditure. Then it has given which kind of storage it is and how it is presented. Variable time and in format columns indicate that all the data are in this case numeric or not it is since we have not in destring it, but at this moment it is not required for you to destring even if the data is in string this describe table will give you result.

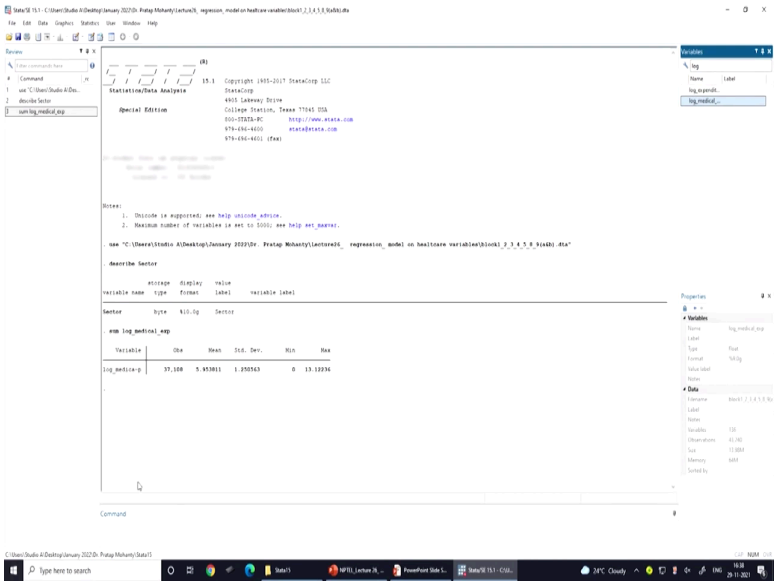
But in case of summary, it is not going to give you, we are just coming to it. The next important estimation is called the summary table, I will straight away go to this particular software. And now I will summarize the sum; the only a sum is fine or a complete summary of the variables.

(Refer Slide Time: 33:09)



If I just take one variable for your reference this is going to give you the result.

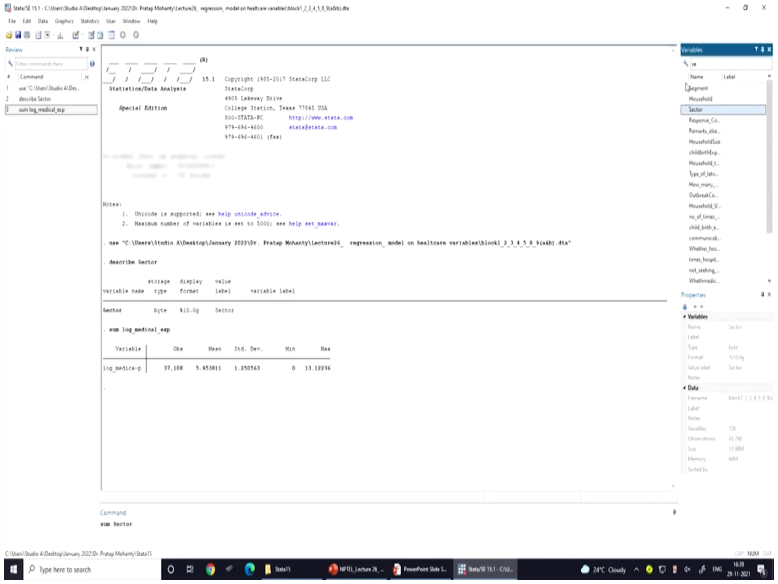
(Refer Slide Time: 33:15)



Now, this has given us information that log medical expenditure which we wanted to understand. The summary table is going to give us information about how many observations are there, what is the mean value of the expenditure, what is the standard deviation, and what is the mean and maximum value.

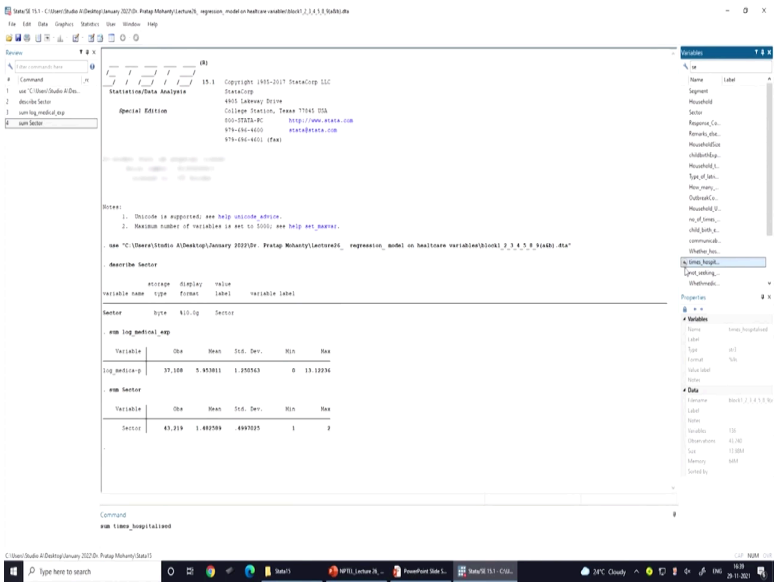
Now, just for your clarity, I am going to give you another summary which is string data, let it be the string data of let it be the sector the same sector variable which we have already done it.

(Refer Slide Time: 33:51)

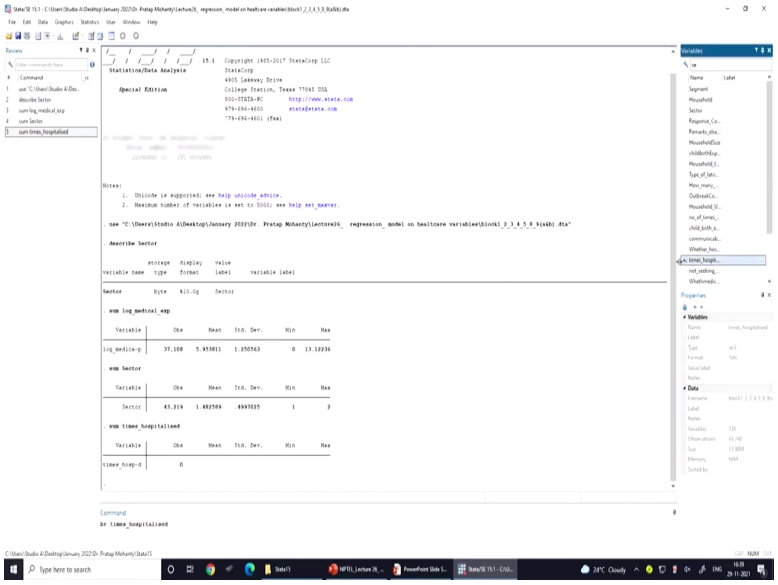


Now we will summarize the sector variable here now you will find the difference between string variable and nonstring variable.

(Refer Slide Time: 34:05)

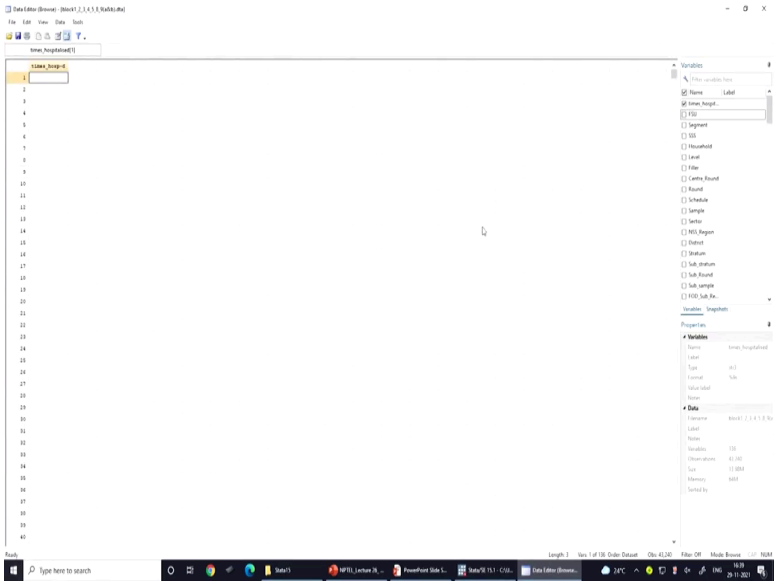


(Refer Slide Time: 34:18)



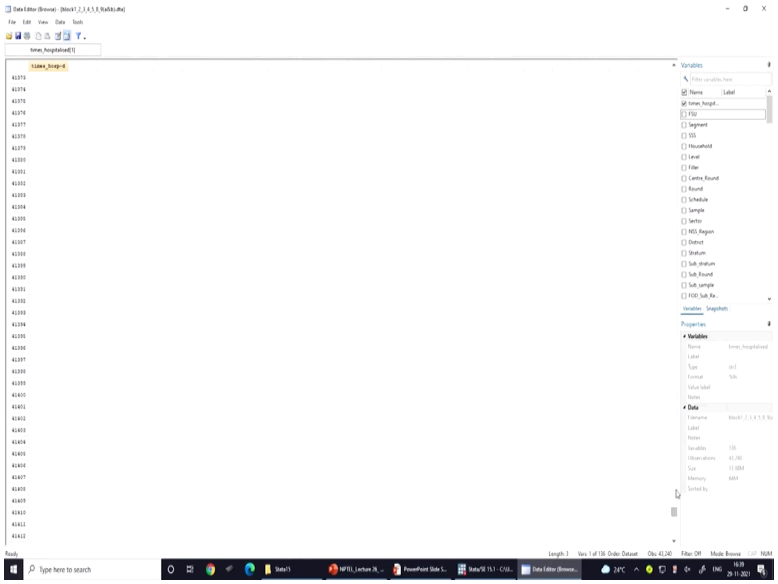
Now enter you will find that this also is going to give us values because this has already been destringed. Will give you another one let it be this is time spent in hospitals. Since this is a destring variable this is a string variable, the summarize table is going to give us 0. How do that this is a string variable? Just check browse and this variable br and times hospital.

(Refer Slide Time: 34:42)

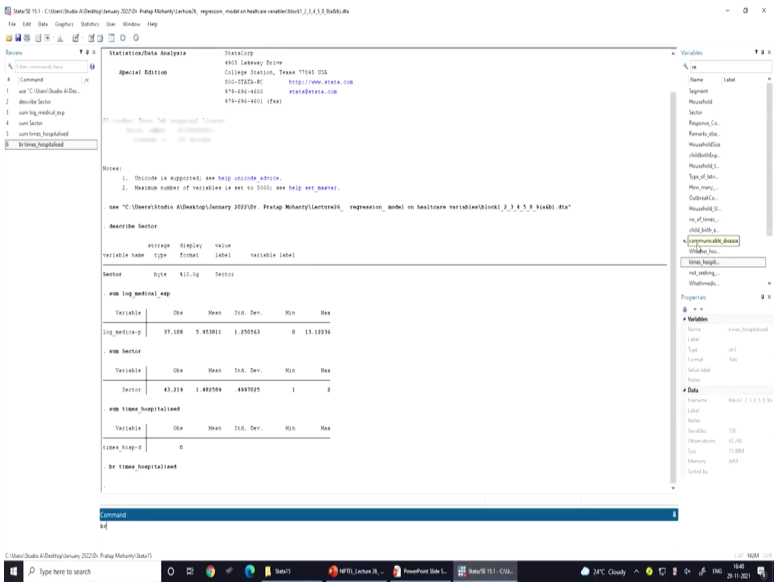


The font of this one must be in red or there are no entries at all or I will just do one thing there is no observation here.

(Refer Slide Time: 34:51)

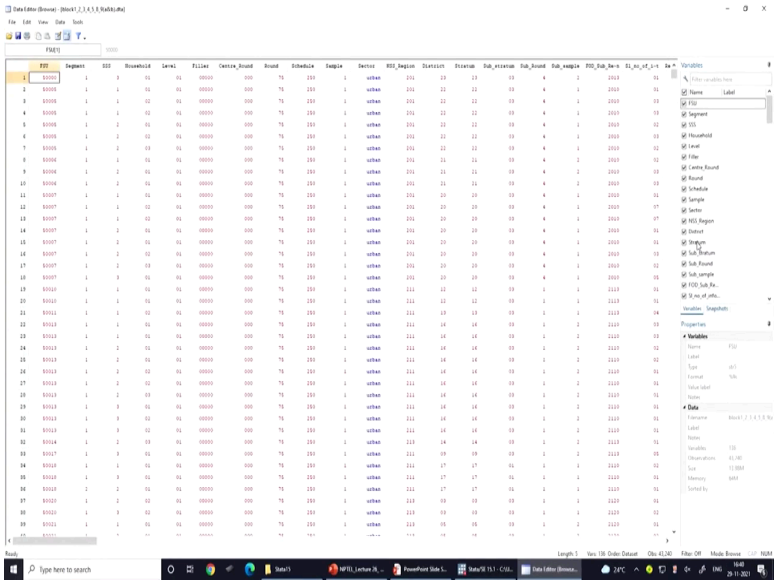


(Refer Slide Time: 34:53)



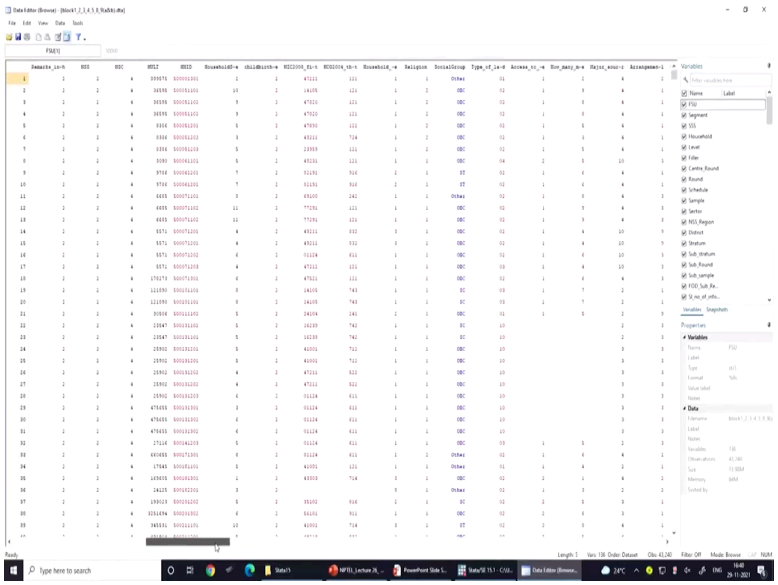
Just browse we will just see which are the string variable browse only browse br enter.

(Refer Slide Time: 35:00)



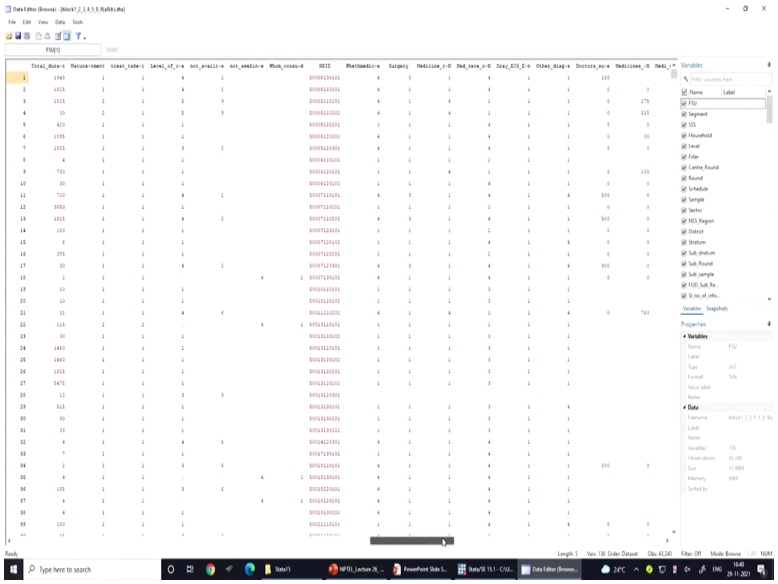
So now, you can see which variables are in fact your string end which are nonstring or numeric, or labeled numeric. I have already clarified to you that the values or the variables which are in red font red color are string variables.

(Refer Slide Time: 35:23)



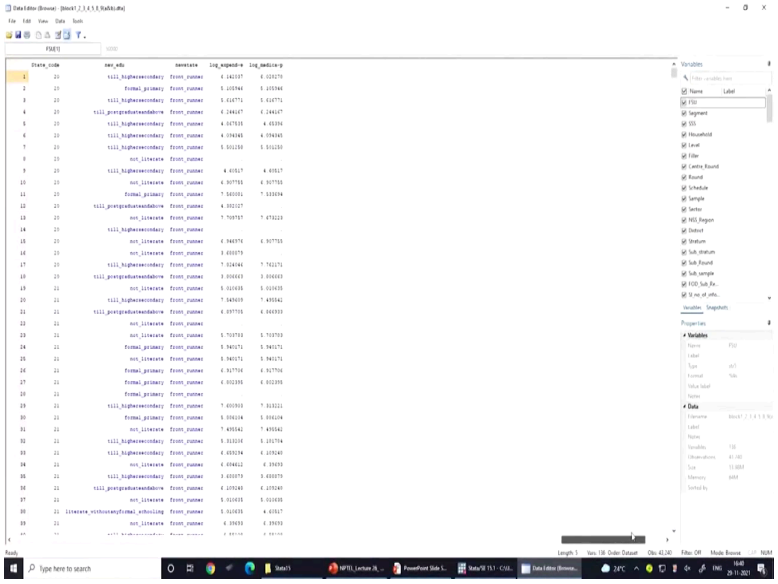
Those values are in black font.

(Refer Slide Time: 35:34)



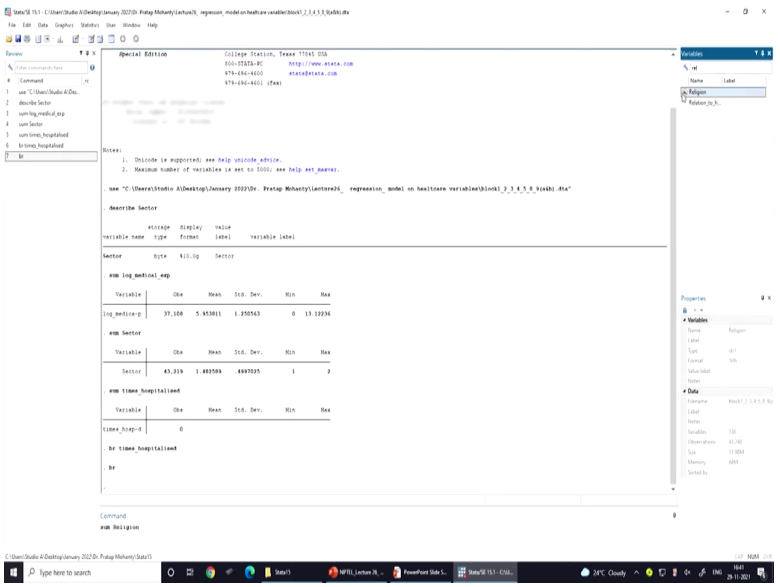
They are all black font they are in fact like the black font they are called numeric values.

(Refer Slide Time: 35:46)



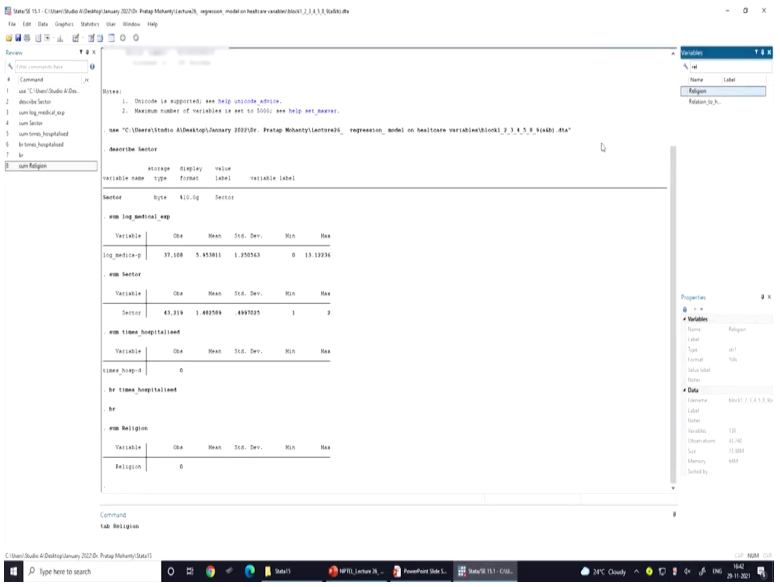
Those who are in blue color they are called labeled numeric they are not exact numeric values they are in fact labeled. So, if I just do one piece of information like at the last these are all our numeric values. From the beginning I will just do one thing, I will just try to summarize one piece of information on your screen; religion let it be. So, now, I will go to explain religion here in red font.

(Refer Slide Time: 36:15)



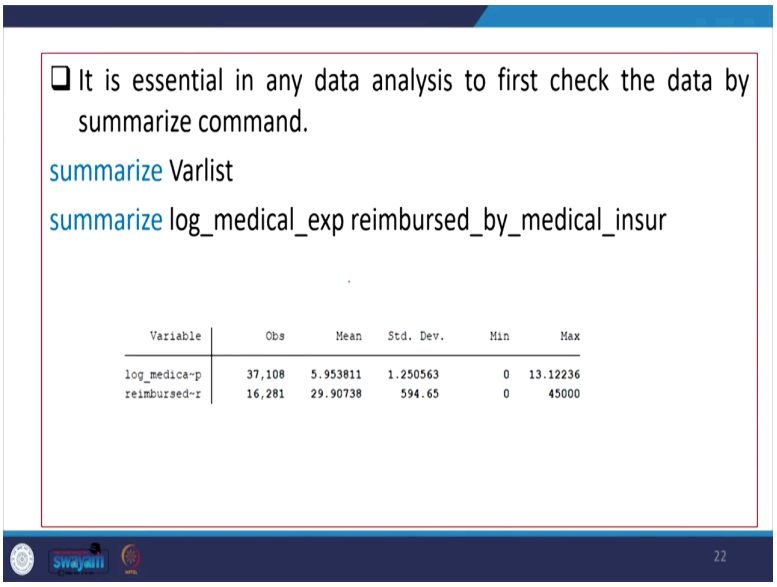
So, now, summarize religion.

(Refer Slide Time: 36:23)



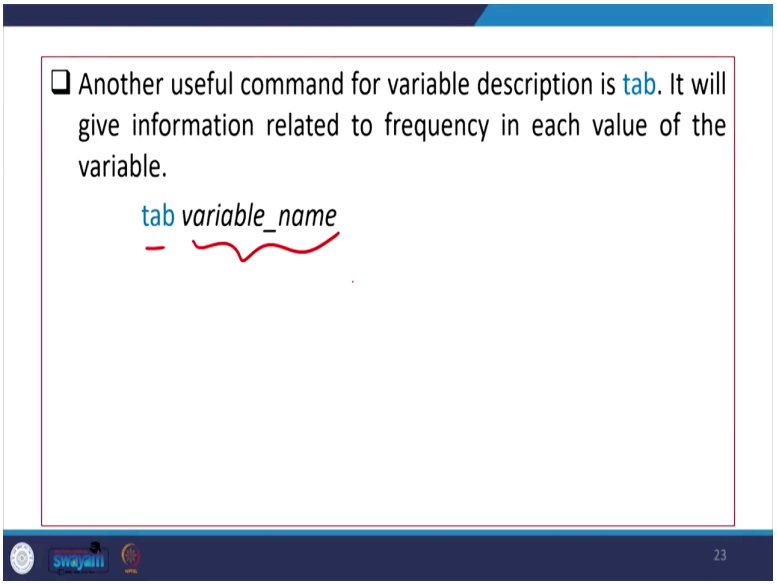
Now, this is going to give us 0 values. So, 0 values, because this is in this, is in a string value. Now, if I destring and run the same command it will give me the right result.

(Refer Slide Time: 36:39)



I can destring it quickly and then I can save it. I think I have already discussed that and we will also use it several times later. So, now, we will go to our PPT and stick to the requirement.

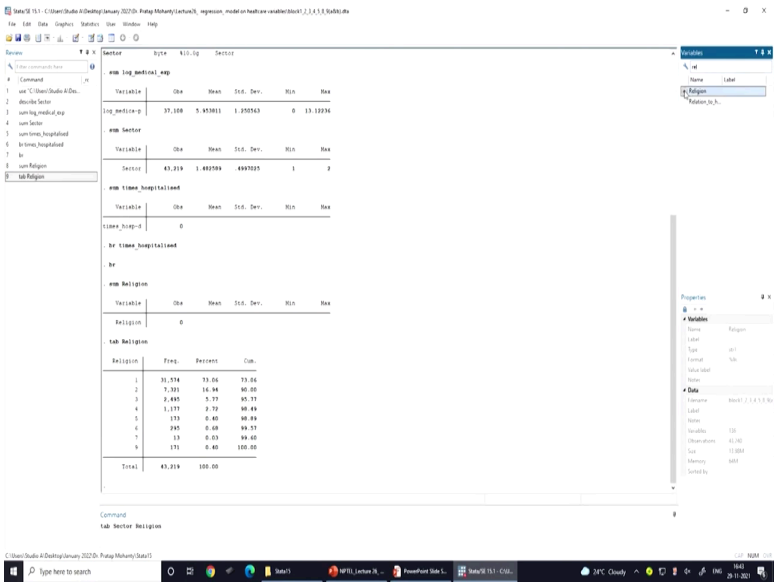
(Refer Slide Time: 36:52)



Similarly next is our understanding frequencies; tab, tabulate, there are tables, they are different commands, I know you will get it very correctly from my modules. So, tab is going to give frequencies of a particular variable tab command you can give it and then the variable

name. But if you like here if I say t a b tab the only t a then any variable here let it be same religion enter.

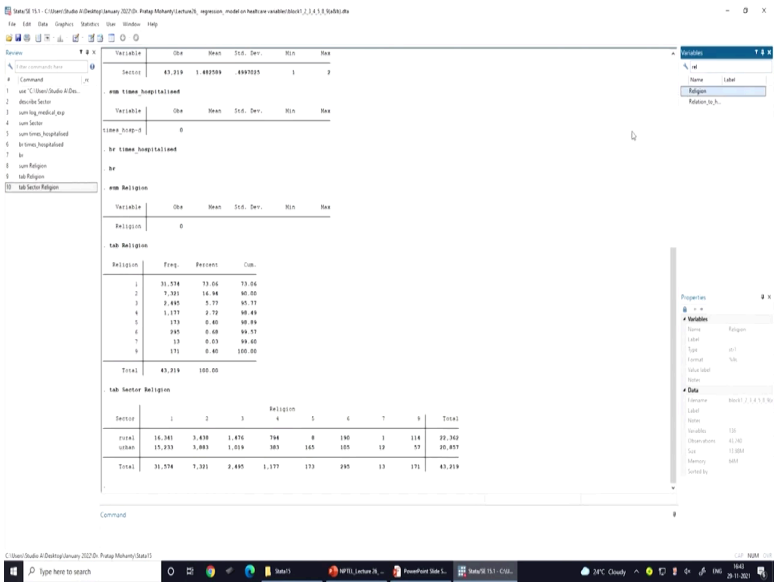
(Refer Slide Time: 37:29)



This is going to give me what? This is going to give me frequency and its cumulative frequencies. And tabulate is the command is also possible to operate in case of the string variable. But since it is not calculating anything once it is estimating anything the string variable is not going to give you any result.

So, now if in the place of double tab and two variables that will be giving you the values of two that is frequencies cross-tabulation of the variables. So, tab then two variables let us take tab and two variables sector and religion.

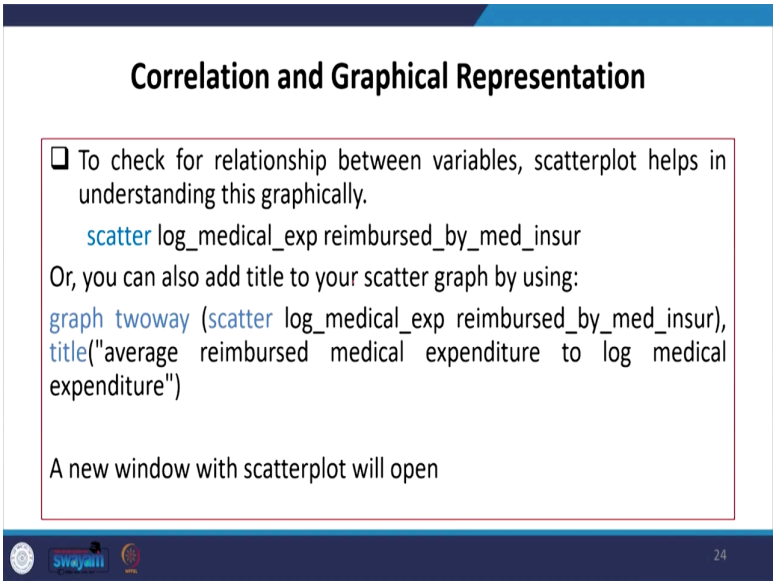
(Refer Slide Time: 38:25)



Then religion gives us information about the cross-tabulation of how rural and urban sectors, are actually distributed across religion, and its clearly understood. Now there are further sophisticated commands possible and we have already explained I have explained them all in detail in the review of command section in my previous module as well you can just follow and find it out.

We will do some new addition in this module on understanding some graphical representation correlation.

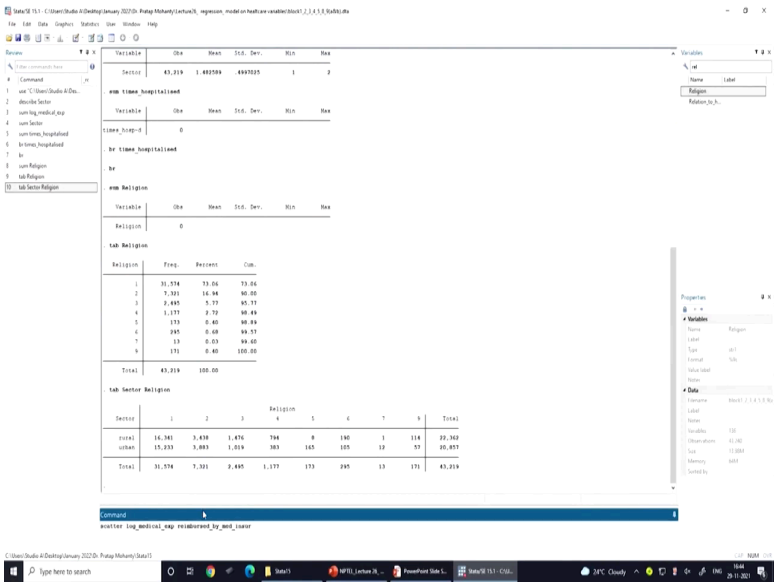
(Refer Slide Time: 39:02)



So, in order to understand the correlation or to check the relationship between variables scatter plot helps in understanding this graphically. Now what we will do? We will copy this command and simply run it. So, and just to show you quickly how it follows and that that is the reason why we have kept it on your screen and I am sure you can operate on your own as well.

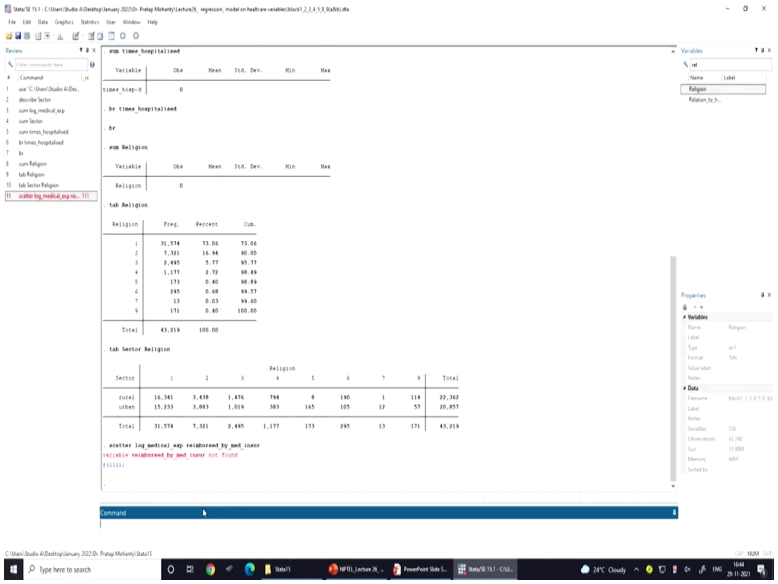
So, we will copy this scatter we could have entered manually, but keeping the time constant.

(Refer Slide Time: 39:43)



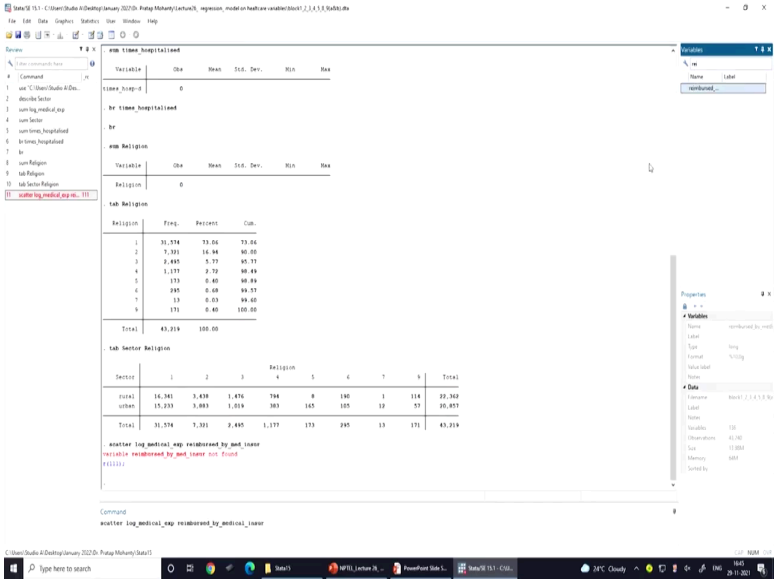
I am only directly operating it in the screen.

(Refer Slide Time: 39:45)



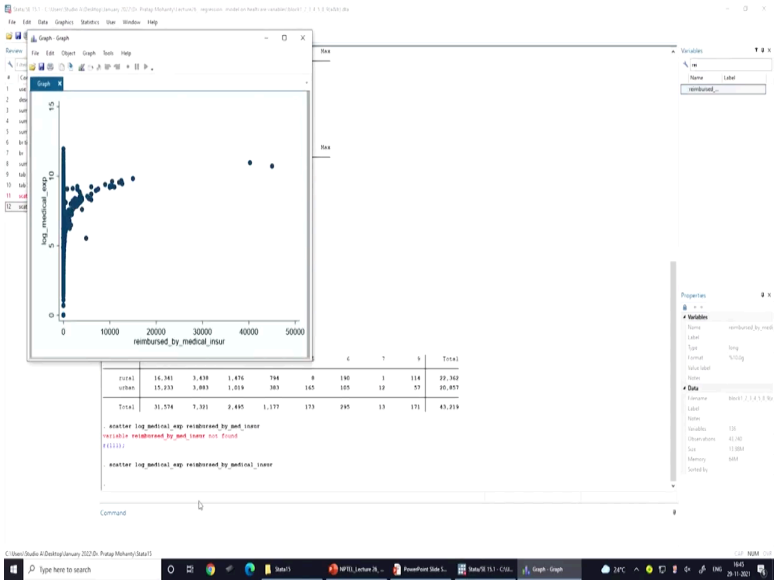
And this is going to give us the scatter plot.

(Refer Slide Time: 39:50)



So, variable symbols are not available we can just change them.

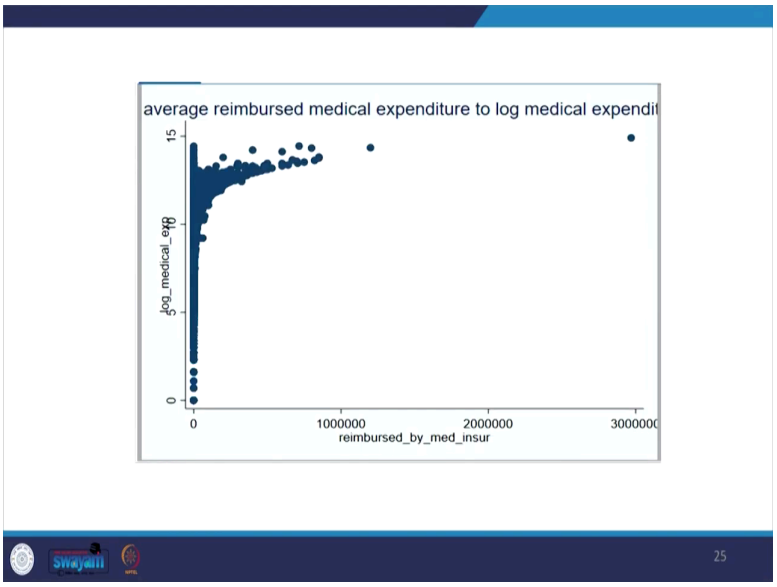
(Refer Slide Time: 40:05)



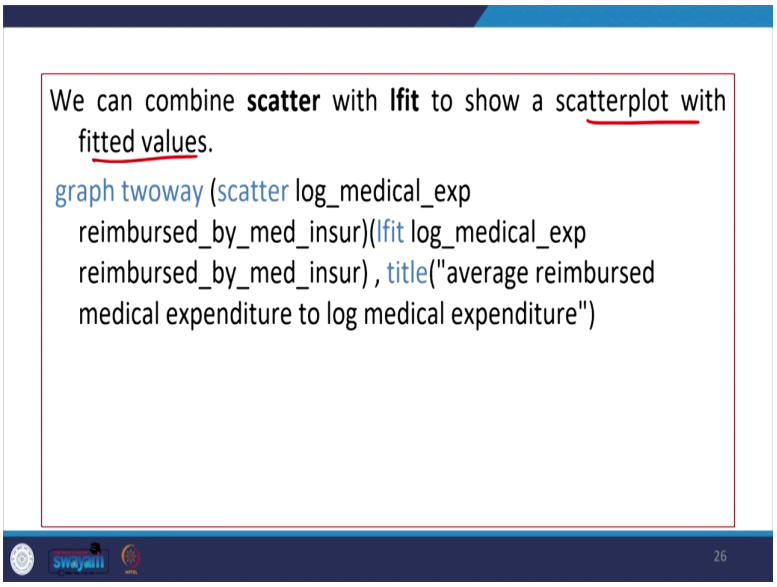
Now it has the name has been changed you can just check, how we also find that there is some errors in the variable. The variable name has been changed in my our data, which is why the model the command was not running. You have to give the exact name by searching through the window.

Now, you can find out how the two variables are related and where they are concentrating what is in fact the trend line all those things we are going to explain, in this model. Then next I am going to operate with we can also operate two-way table as well two ways scatter plate as well, then the average could be also run we are not going to do it at this moment. We will simply skip it and that is kept for your reference only.

(Refer Slide Time: 40:55)

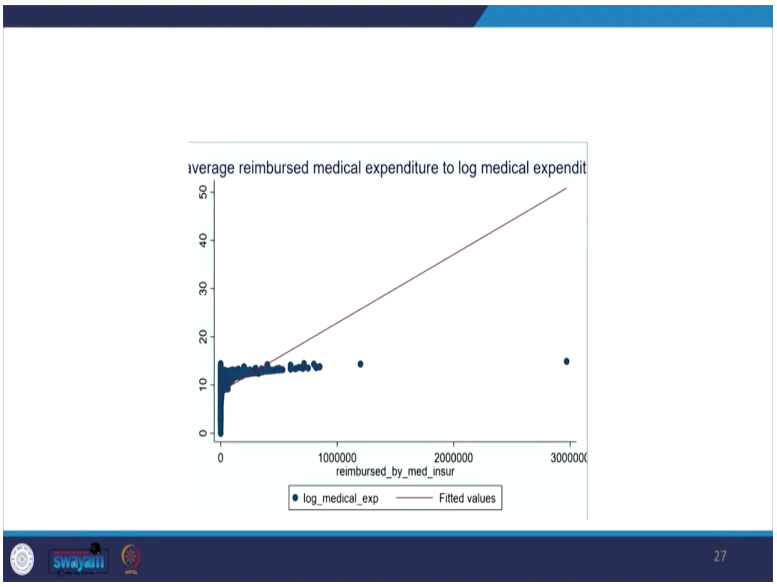


(Refer Slide Time: 40:56)



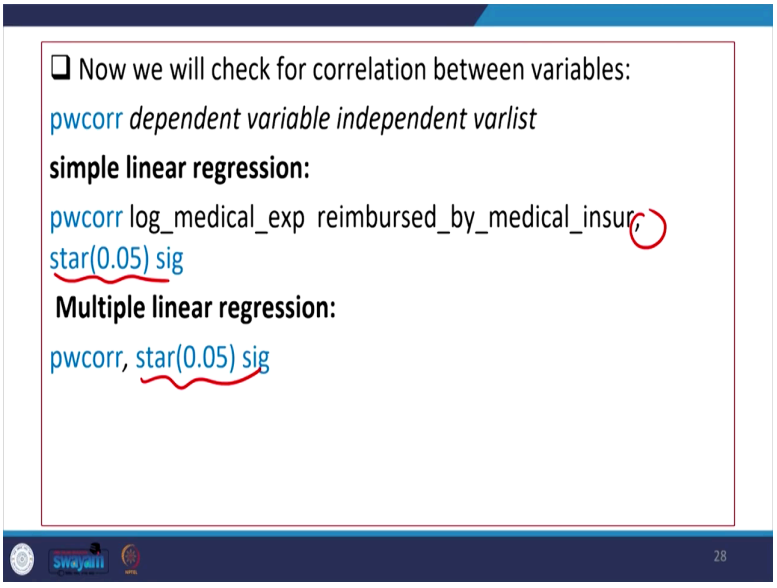
Like here you can do like you can combine scatter with ifit to show the scatter plate with fitted values. So, the fitted values can also be explained, so that command we have already given for your reference.

(Refer Slide Time: 41:16)



I am sure the trend line the result will look like this and the trend line basically the fitted line command if you attach it will give you this kind of result. So, fitted line in this case it seems that the fitted line is not that correct. So, you have to include multiple variables and the relevant variables to draw the fitted line.

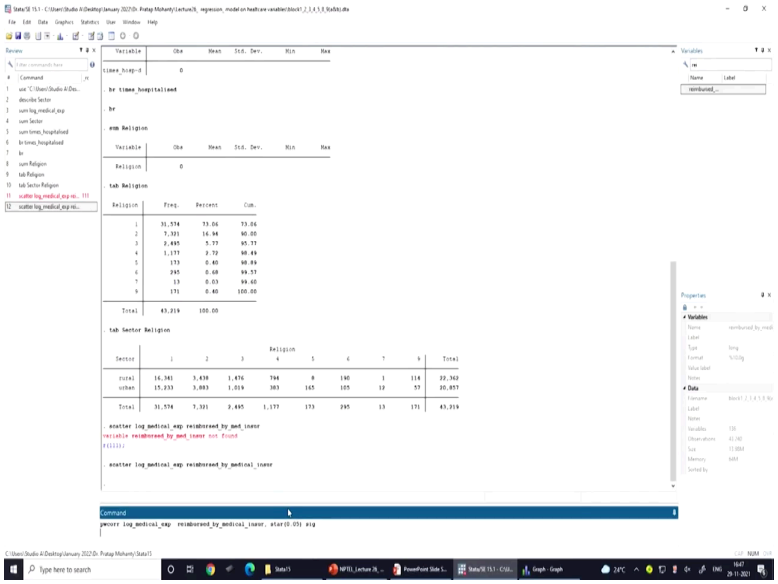
(Refer Slide Time: 41:39)



Similarly, correlation can also be checked through this command where I am just going to run it as per this one. So, it is here. I am just copying it once again and we will go to that

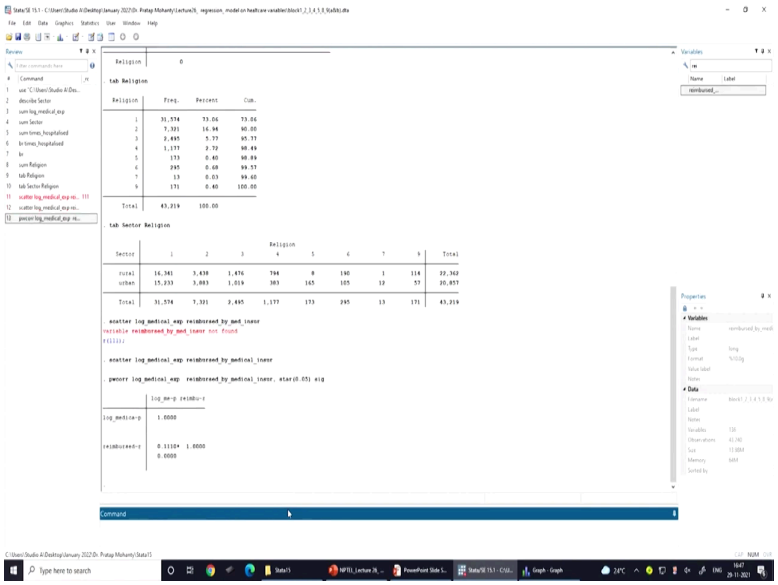
particular page. And we have kept all those things for your reference. Yes, this is the one. We are just copying it and running it on the window.

(Refer Slide Time: 42:12)



The window is here alright.

(Refer Slide Time: 42:15)



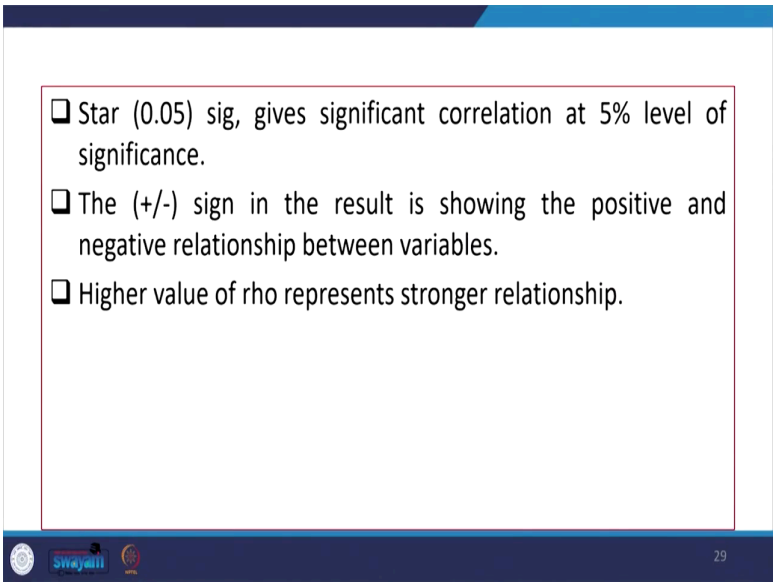
So, the correlation and the significance value basically the power of the correlation we have estimated. So, the star mark represents the level of significance at the 5 percent level 0.045

command is attached to this star. So, wherever the relationship is established at a 5 percent significance level it indicates the star mark.

Therefore, star mark you can just see, that I am just explaining once again. Similarly multiple linear regression linear in case of for the multiple linear regression we can also understand the power of the correlation basically the level of significance. You need to just attach this part.

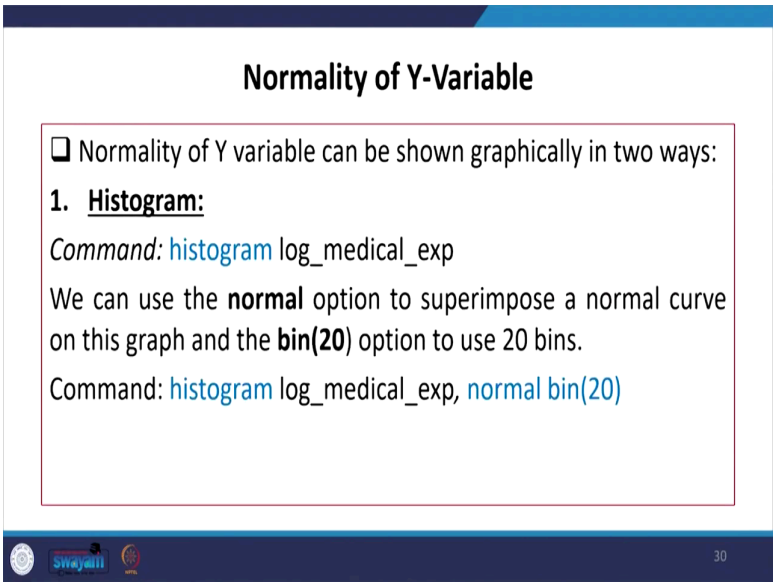
In addition to that comma followed by comma then star with the bracket at 0.5 percent etcetera if you give it then it will give you the right result. So, the plus and minus sign indicates the value and the relationship whether positive related or negative related.

(Refer Slide Time: 43:26)



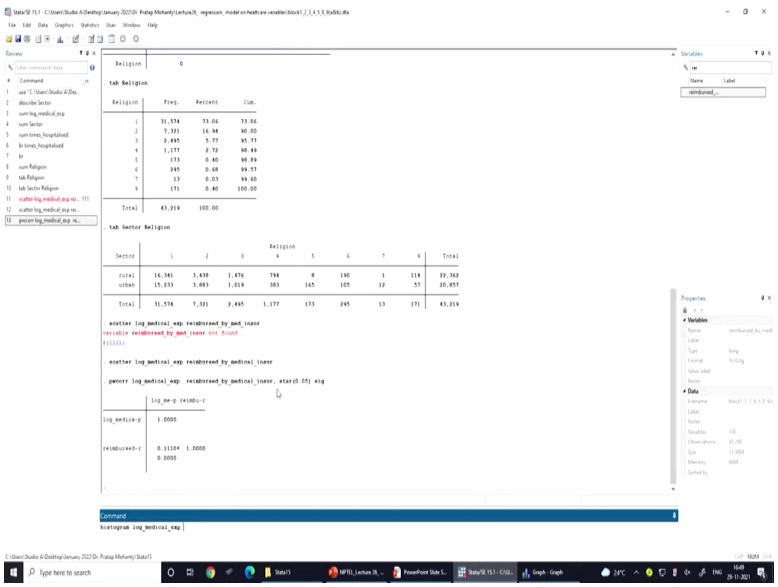
Higher value of rho represents stronger relationship or weaker relationship.

(Refer Slide Time: 43:27)



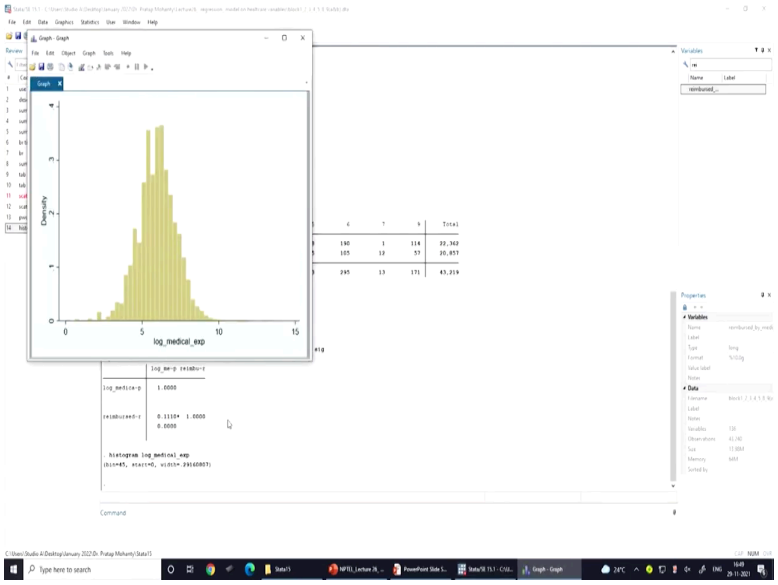
The next check is on the normality of the Y variable, the dependent variable how it is normal, and what extent we emphasize the normality of the distribution. The command we usually follow to go for a histogram how the histogram looks like, just to check histogram log a medical expenditure which the variable we have taken that we can check to do it once again it is here. So, we will run this command histogram to log medical expenditure that is that has been already taken.

(Refer Slide Time: 44:10)



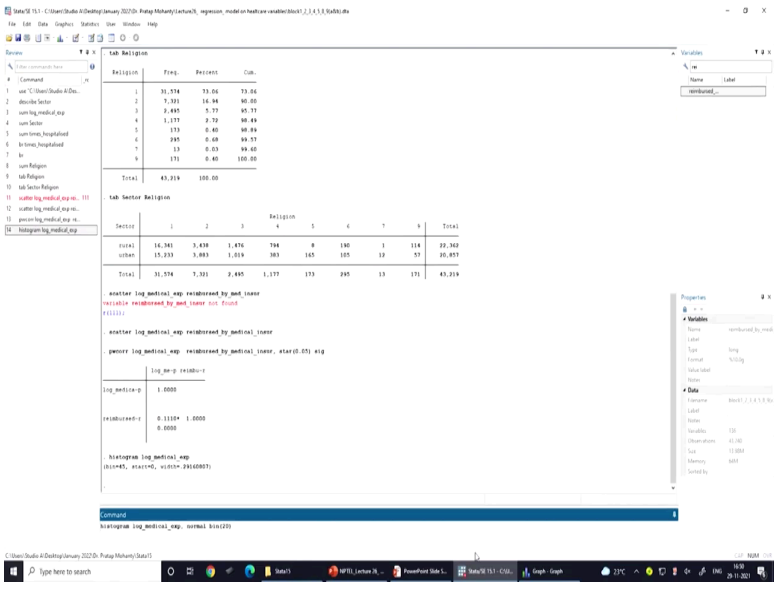
So, the histogram will be displayed on your screen.

(Refer Slide Time: 44:11)



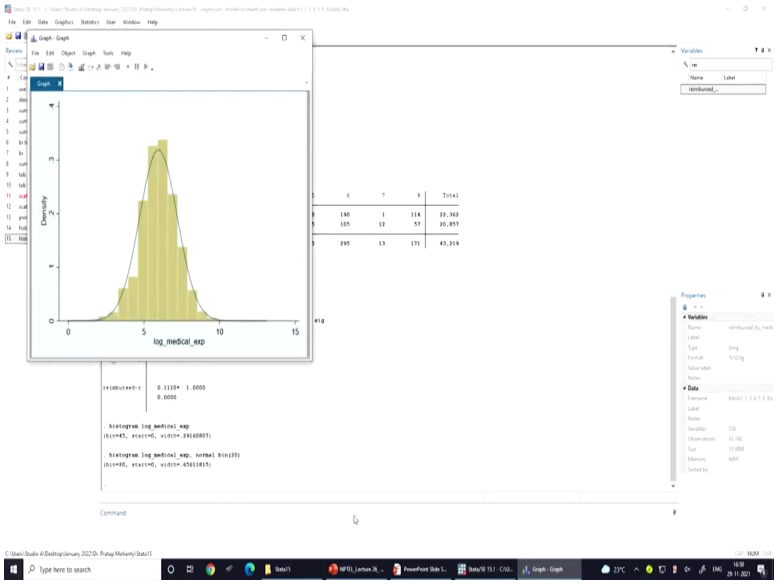
Now, in this one, you can easily see what the trend line looks like. The line we can also draw through the next command is the 1. So, histogram, then we will specify it.

(Refer Slide Time: 44:37)



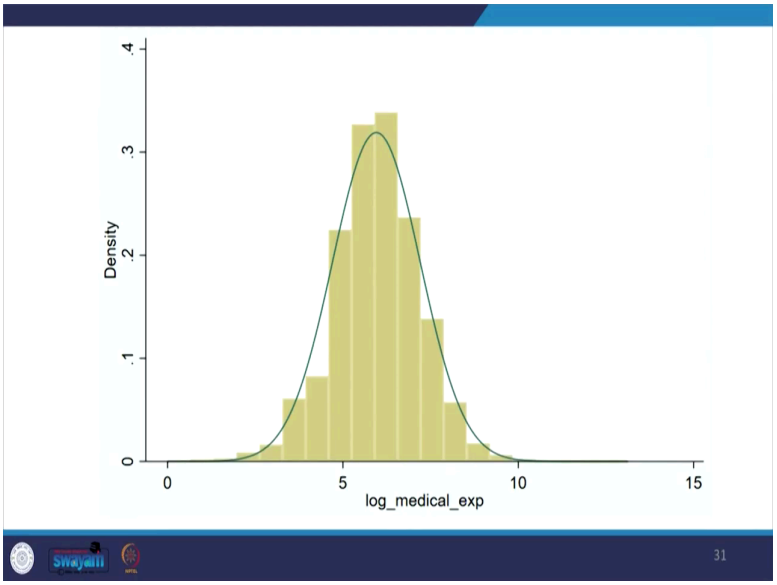
So, it is here.

(Refer Slide Time: 44:39)



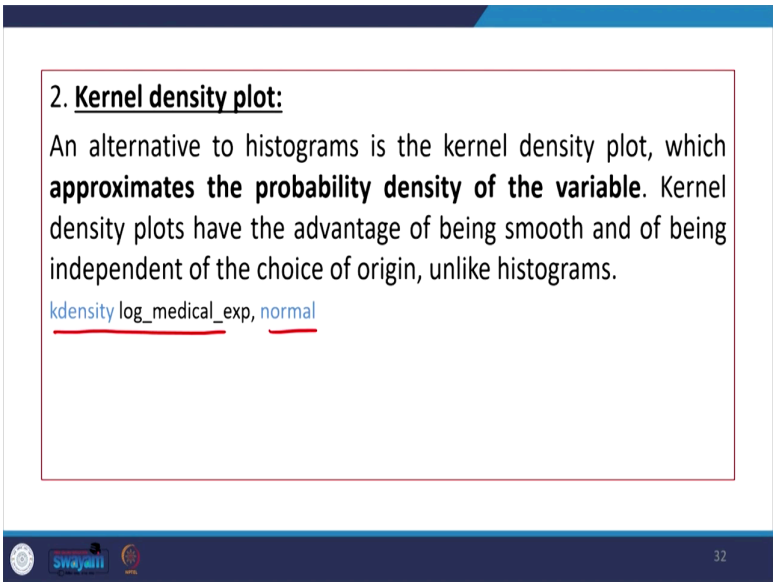
And it will specify the graph correctly. The graph now will you can identify whether the graph is approximating a normal distribution or not. This seems that this is a kind of leptokurtic diagram and then more it is symmetric though not mesokurtic. So, it is leptokurtic, but it has normality assumptions fulfilled. So, then coming to the next one this is what we have already explained.

(Refer Slide Time: 45:15)



So, this is what we have derived.

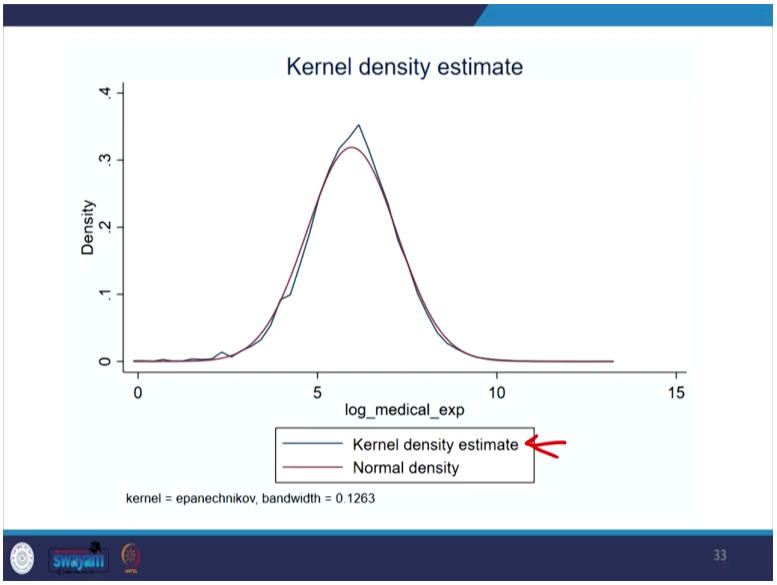
(Refer Slide Time: 45:17)



Then another one is called kernel density plot the through that you can also check the normality of the distribution. This is an alternative way to check the plots. [FL] this also gives approximate values of the probability density function of the particular variable.

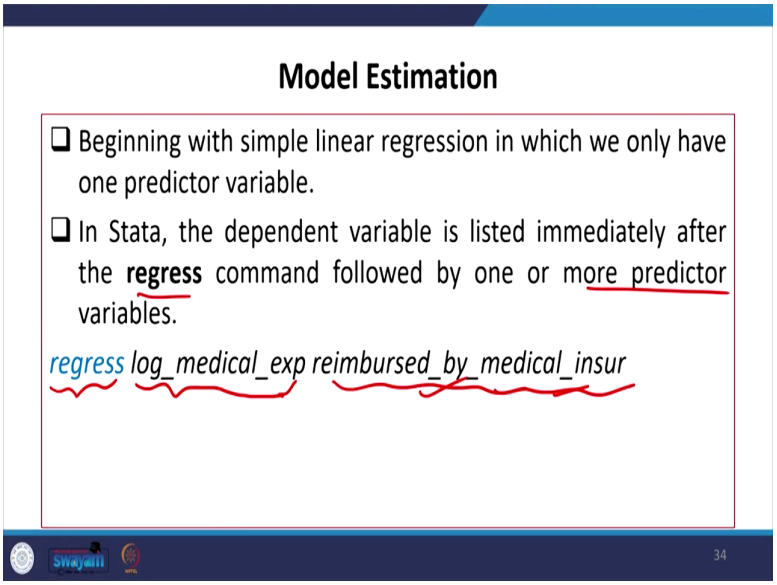
So, here we need to specify kdensity the variable name then normal the word we need to I mention. Kernel density plots have the advantage of being smooth and of being independent of the choice of origin like histograms.

(Refer Slide Time: 45:57)



So, this will look like this. I am not going to draw it here once again. So, you can see how the normal distribution plot is derived kernel distribution estimation is highlighted with the blue line. This is almost overlapping with the normal density function. Here you can easily say that yes your data is more normal.

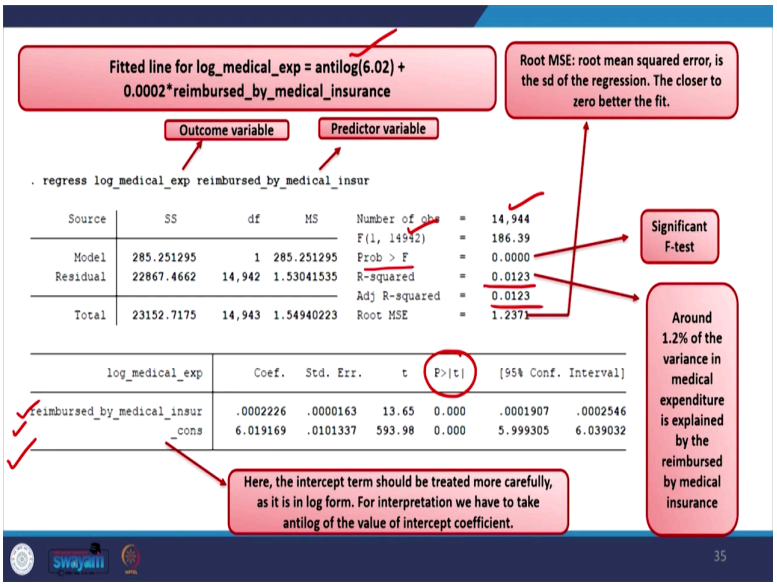
(Refer Slide Time: 46:20)



The next one is about model estimation beginning with sample simple linear regression in which we only have one predicted variable. In Stata the dependent variable is listed immediately after the regress command, regress is the command we give. So, the dependent variable will quickly follow the regress command.

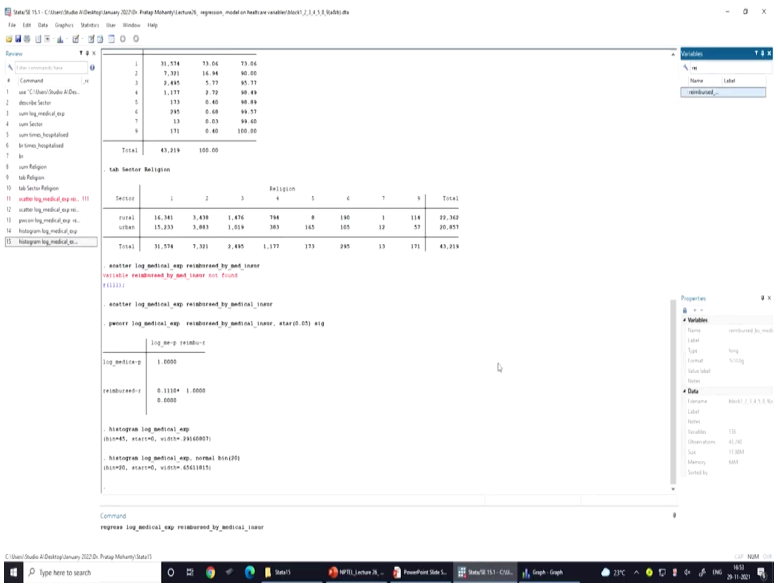
Then one or more predictor variables are specified, is like this. This is one, this is our dependent variable; then all sorts of independent variables could be mentioned alright. So, this is what is going to give us the estimation.

(Refer Slide Time: 47:01)



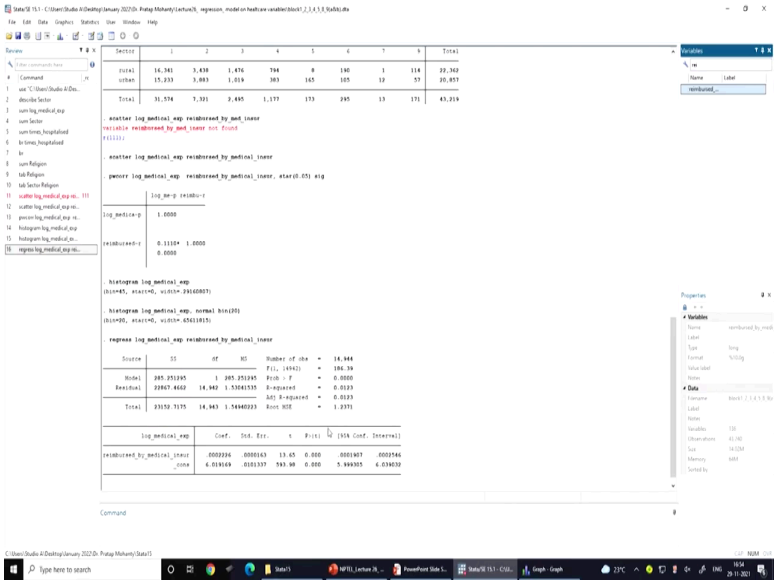
Therefore, first of all, we will copy this then if I remember this variable we can simply click that log medical expenditure then reimbursement then medical insurance etcetera. Therefore, will straight away go to Stata we will type regress.

(Refer Slide Time: 47:17)



Therefore, then all the variables are listed here log medical expenditure we are taking at the dependent variable, then we will take reimbursement.

(Refer Slide Time: 47:32)



Then related to reimbursement of insurance reimbursement for the medical expenditure. Now, on the screen you have got your regression result the ordinary least square regression model. This suggests that how the f coefficient beta value and beta naught and beta 1 the constant is in fact is your beta naught that is 6.01 9169 and your beta 1 is 0.0002226.

Most importantly we need to understand whether the data the coefficient is actually significantly estimating our model or not. Now, the P value are very effective P value if it is less than 0.05; that means, it is significant at the 5 percent level if it is less than 0.01; that means, it is significant at the 1 percent level.

Therefore, here it seems that your beta value is significant at even 1 percent level as well. Therefore, I am just explaining through the diagram through the PPT, here are the results of many interpretations you can do it. First, you need to check about the entire model and whether it is significant or not through the F test by default it considers the F test comparing the differences of residuals.

Therefore, the F test is defined, and the degrees of freedom is more important. So, F test values and its degrees of freedom are numerator and denominator are 1 and 14942, this is this given. And because a number of numbers in a number of observations is 14944 and accordingly others are understood. Now, the fitted line for log medical expenditure is in fact that we can derive, now at this moment let me just clarify its P values which I have just said is given here.

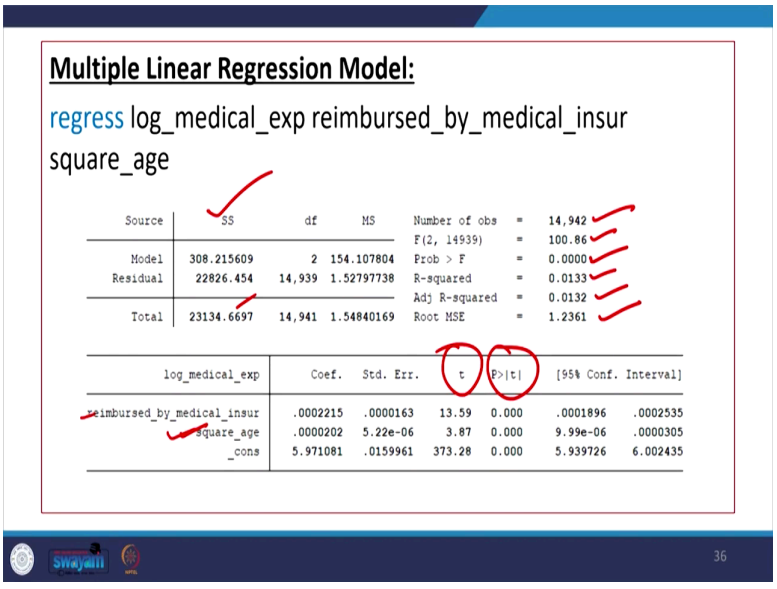
The model says that regarding whether it has a perfect fitted line or not R square value which we explained earlier R square and adjusted R square value is important. R square suggested that around 1.2 percent of the variance in medical expenditure is explained by the reimbursement by medical insurance. So, similarly adjusted R square where the degrees of freedom due to the parameter is changed. So, the number of our estimated parameters is changed.

So, accordingly, since you are only one parameter, we have taken one independent variable you have taken. So, a number of parameters is the same as for the entire model. The R square and adjusted R square are the same. It is not different When you increase the number of more than 1 independent variable your degrees of freedom will be different, and the adjusted R square value will be lesser than that of the R square value.

Adjuste R square value could be also negative which is all those things we already explained. Another piece of information in this chart around about linear or null in linear regression is on root means the sum of the square that is root means square error. root means square error is in fact the standard deviation of the regression, this is the closer to 0 better the fit.

And from all those things we can estimate a fitted line. So, fitted line for log medical expenditure is equal to antilog is 6.02 of the constant variable plus 0.00 0002 times the variable that is reimbursed by medical insurance. If I do this calculation I can get a fitted line, this is what your Excel or even your software usually follows to derive the fitted trend line.

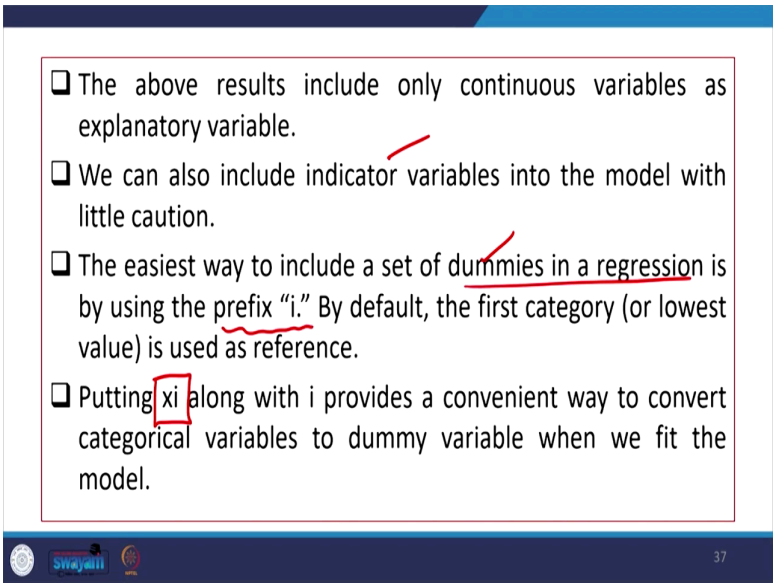
(Refer Slide Time: 52:01)



Similarly, we can do a multiple linear regression by adding another variable here we have added in addition to this insurance we have added now square age. So, with the same regress command, we have got the information, all those are a very important number of observations F values whether the model is significant or not R square adjusted R square root mean square we have already all explained.

Therefore, similarly residual sum of the square is total sum of the square. So, you can also estimate t values and then P values. So, and their coefficient everything is presented, so I have already explained I think I need not spend more time here.

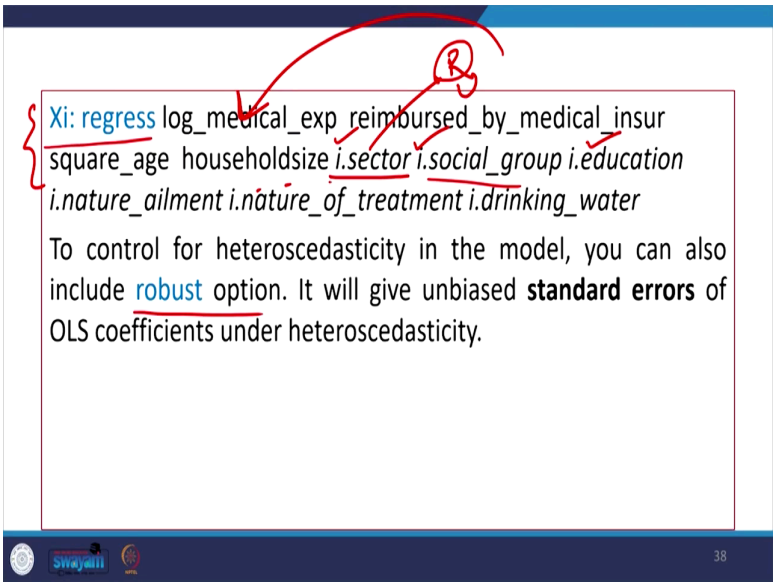
(Refer Slide Time: 52:52)



The above result including only continuous variables as an explanatory variables we can also include indicator variables into the model with a little caution. If there are certain indicator variables or dummy variables the easiest way to include a set of dummies in regression is by using the prefix I.

If there are set of dummies is included or an indicator variables is included, the i dot command is more suggested that will be comparing your categorical variable from a base category. By default, the first category is considered the reference category. In that case, some conditioning should be given. So, putting xi should be there, xi command along with the regress must be there and that will give you better information related to the interpretation of the dummy variables.

(Refer Slide Time: 53:45)

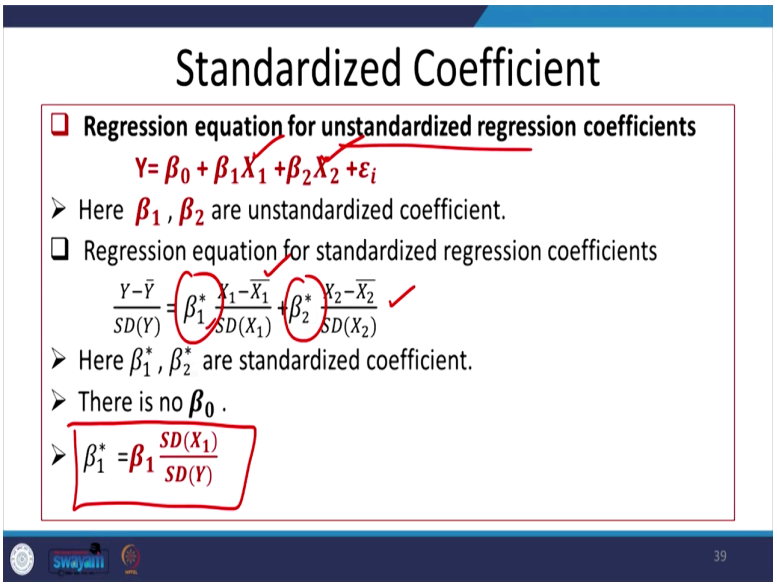


Therefore, this is what the command looks like. In our next module, we will explain all those things in detail, but at this moment we are not drawing. I will interpret the result to some extent. If I give Xi along with regress I am doing it because I have some categorical variables like sector, so i dot is taken. Then another categorical variable is here social groups, education etcetera similarly here as well.

Therefore, this will compare its base category like if sector we have rural and urban; rural is our best category and urban is another category. So, the coefficient whatever we have received rural by default is the base category you can also change the base category your coefficient is going to interpret as, how urban is comparatively better off or incurring expenditure as compared to rural on the medical expenses.

Therefore, in that case, another interesting aspect is to control for heteroscedasticity in the model, you can also include a robust option. So, the robust option is going to control your heteroscedasticity, it will give one by standard errors of OLS coefficients under heteroscedasticity.

(Refer Slide Time: 55:03)



So, there are some other details like standardized coefficient, unstandardized coefficient since we have already run for huge minutes in this lecture. We are not spending much time. I am just wrapping up all those things detail in our next episode maybe in the next year we will come up with detailed lecture of all those things.

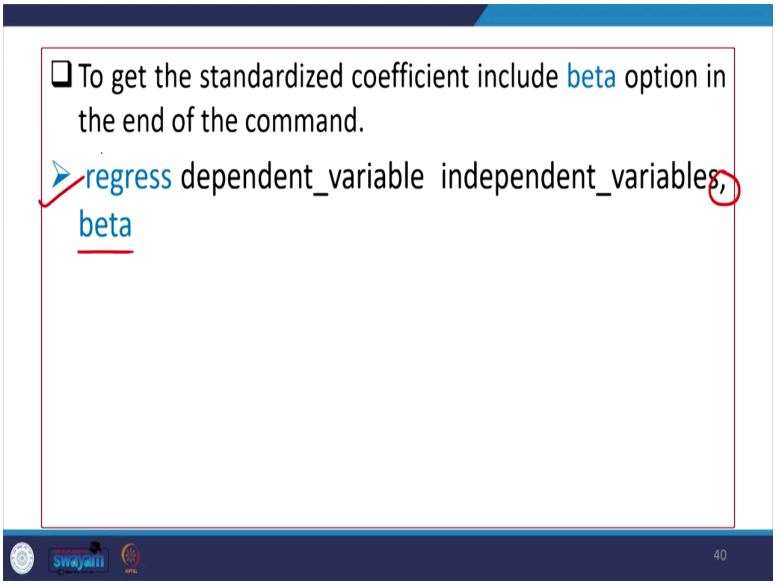
At this moment I am just clarifying this concept called what is what you mean by standardized coefficient, what do you mean by unstandardized coefficient. Regression equation for standardized regression coefficients it is simply unstandardized regression coefficient is simply taking its variable as it is X1, X2 etcetera. So, beta 1, beta 2 are called unstandardized coefficients.

Regression equation for standardized regression coefficients if anything any other variables are standardized. Like if beta 1 is not just with it is X i rather X i is taken from their mean value divided by standard deviation. That means the X variable X i variable has been standardized. X 1 variable has been standardized here also X 2 variable has been standardized.

The coefficient that we derived that is beta 2 star and beta 1 star are actually called standardized coefficients. And there is no beta naught because if you standardize that since it is already a constant. And if you take its mean value from its constant value, that we mean and the value itself is the same. So, the difference is going to be 0, so there will be no question of beta naught.

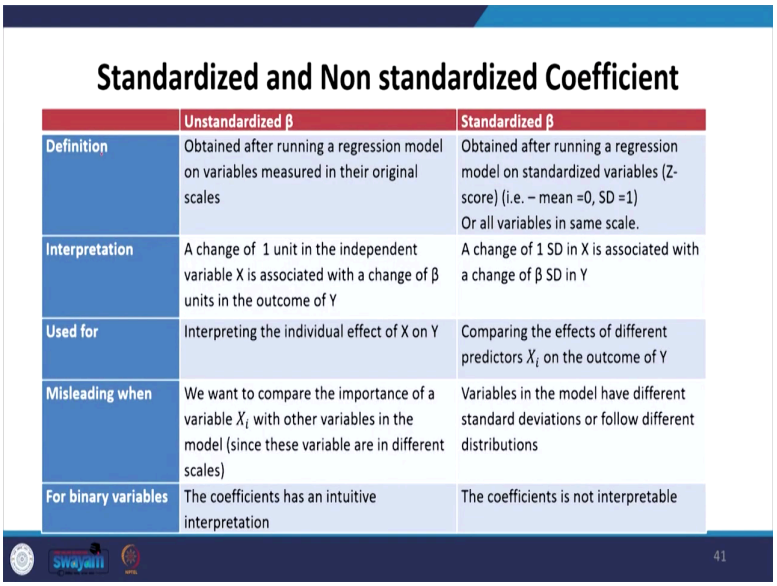
Therefore, when we have a standardized coefficient. So, beta 1 star is equal to beta 1 times the standard deviation of X1 divided by the standard deviation of the dependent variable. So, this is very important so far as any quiz question is concerned you can just note and you will understand further.

(Refer Slide Time: 56:59)



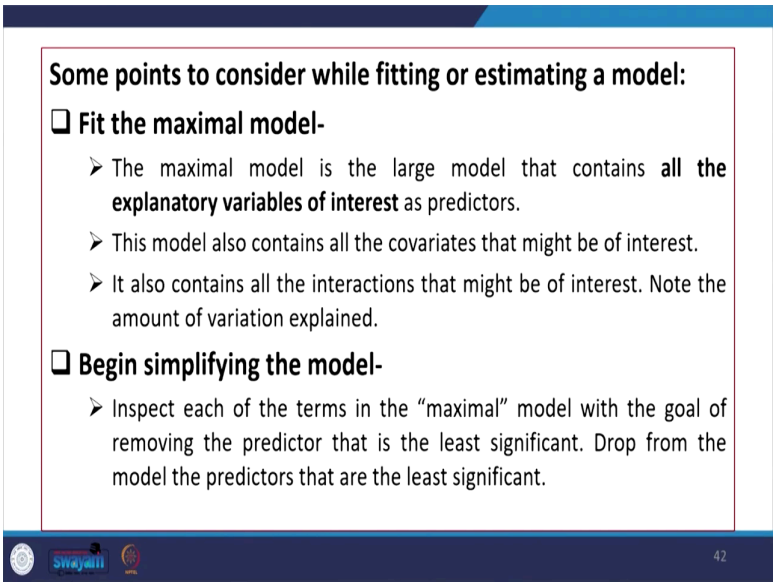
To get the standardized coefficient include the beta option at the end of the command. So, the beta if you are just by adding comma then followed by beta, the beta coefficient is going to give you standardized coefficient value. The rest of the commands are the same.

(Refer Slide Time: 57:16)



I am not calculating step by step; we will certainly come up with these in the next episode at this moment we are just clarifying it. Standardize and non-standardized coefficient by definition by interpretation, how it is misleaded, how these are regarded in case of binary variables. All those things I am sure if you read you can able to enjoy and understand the difference between non-standardized beta and standardized beta.

(Refer Slide Time: 57:47)

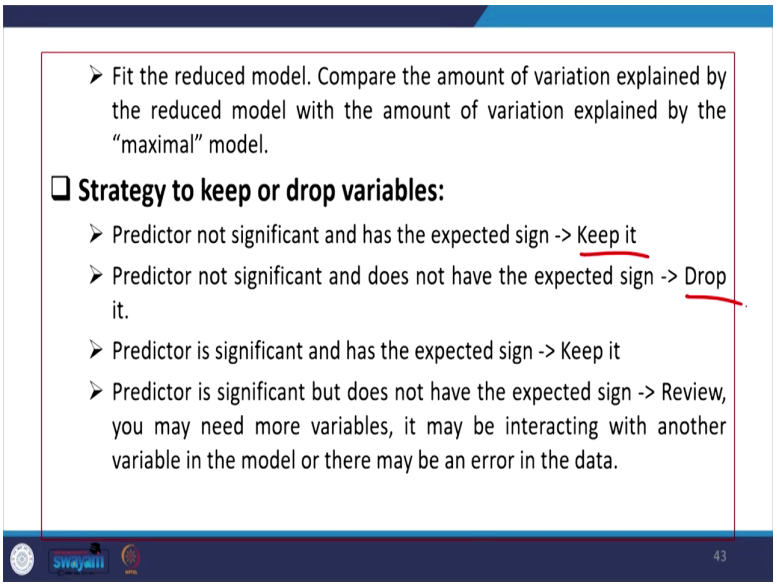


I am not going going through them in detail these are all explaining the basic meaning I have already explained I am sure you will understand this. Some points to consider while fitting or estimating a model like feeding the maximal model beginning simplifying the model etcetera can be done like.

The maximal model is the large model that contains all the explanatory variables of interest as predictors. That is called a maximal model. This model also contains all the covariates that might be of interest. It also explains or it also contains all the interactions that might be of interest note the amount of variation explained is also captured correctly.

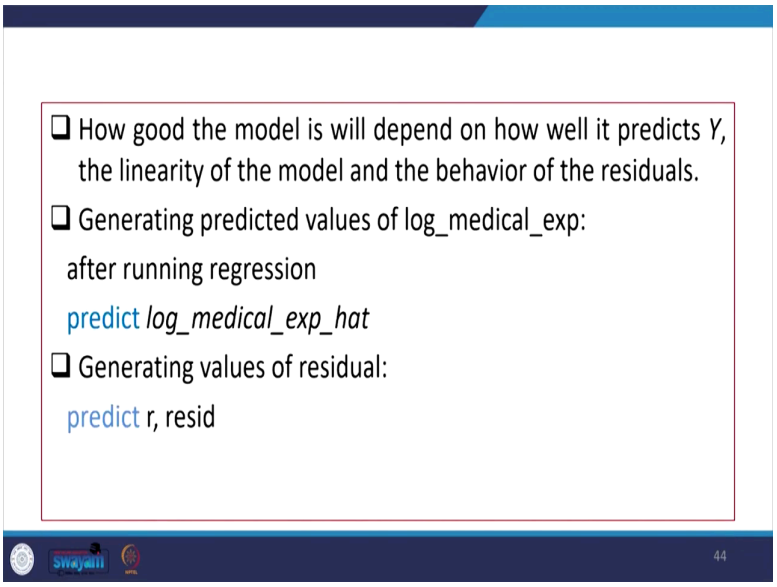
Then begin simplifying the model inspect each of the terms in maximal model with the goal of removing the predictor, that is this least significant one. Then drop those from the model and those are least significant and then further you run the regression.

(Refer Slide Time: 58:51)



Now, feed the reduce model compare the amount of variation explained by the reduced model with the amount of variation explained by the maximal model. So, the strategy to keep or draw variables is like predictor not significant and has the highest expected sign, in that case you can keep the command we have already set keep or drop command that you can show up or keep. And then further review and find out which one is in fact your better model.

(Refer Slide Time: 59:20)

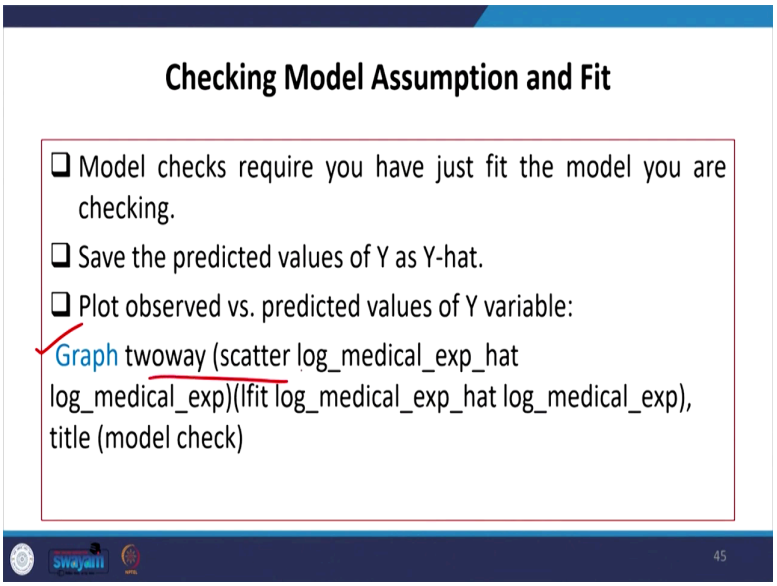


Then you can actually also predict by predicting you have to give the variable command is predict then that variable with a hat is there. Now, how good the model is dependent on how

well it predicts the linearity of the model and the behavior of the residuals etcetera generating predicted values of log medical expenditure.

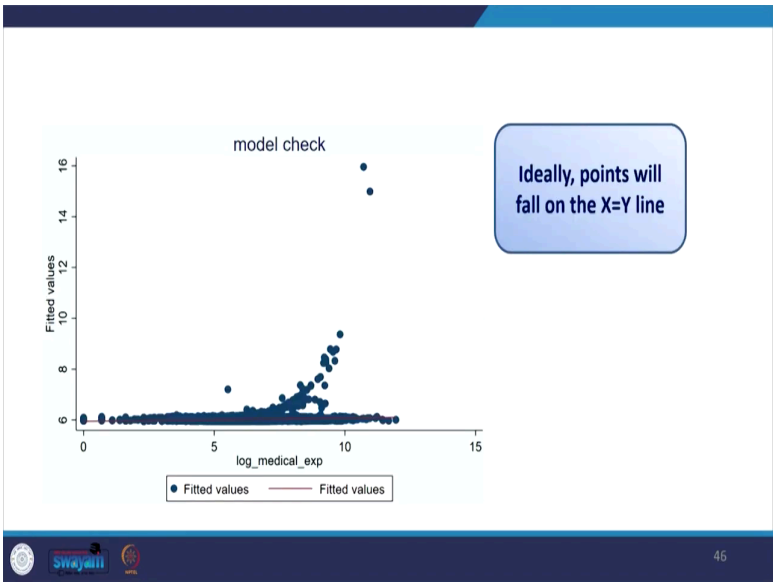
After running regression simply, you give predict at the command. So, basically whatever the result we have got that is in terms of the hat. By hat it should save it as by adding with a hat in addition to that name. Then predict r and resid that will give you the distribution of his residual terms.

(Refer Slide Time: 60:13)

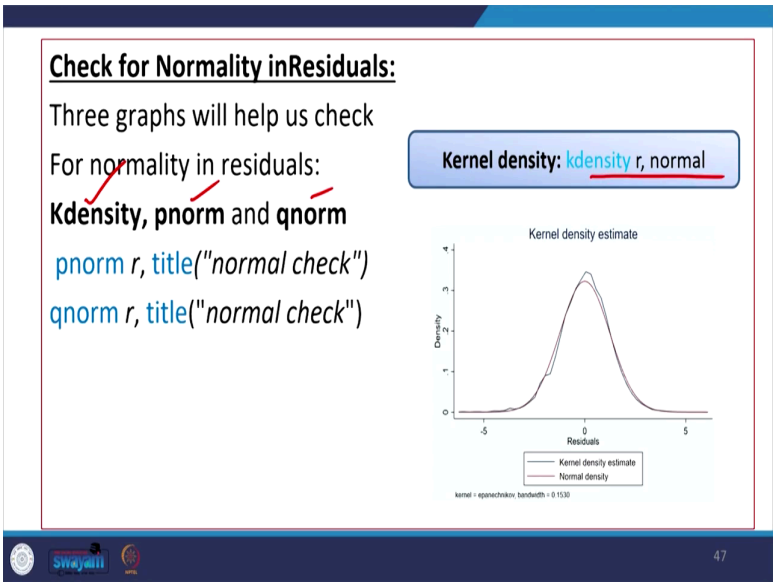


Checking model assumption and in fitting how clearly these are going to give it.

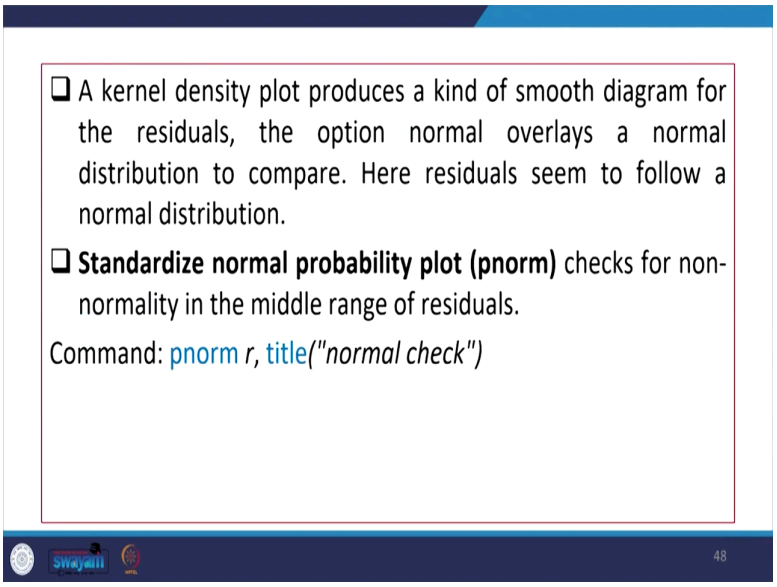
(Refer Slide Time: 60:18)



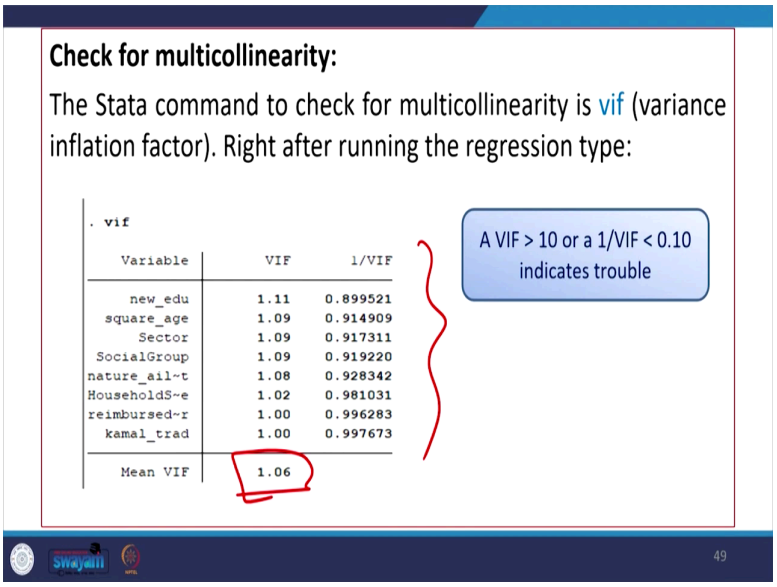
(Refer Slide Time: 60:19)



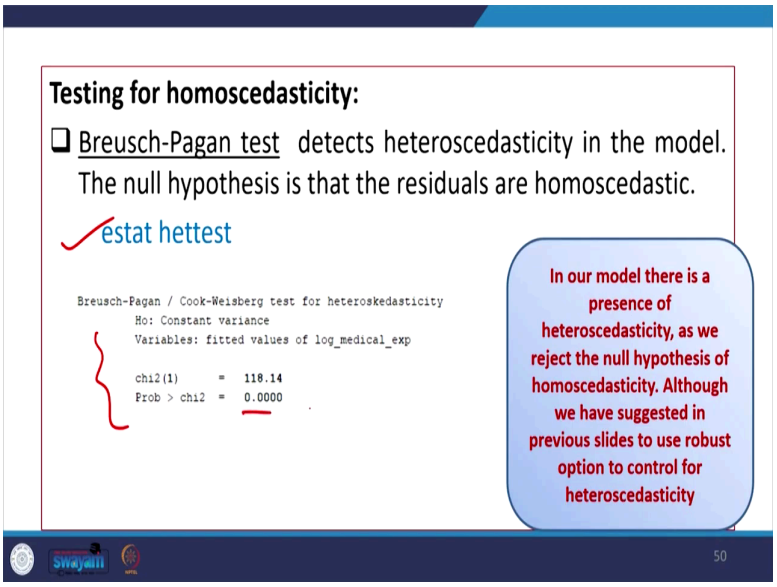
(Refer Slide Time: 60:20)



(Refer Slide Time: 60:20)



(Refer Slide Time: 60:21)



I am just explaining the overview of it. So, checking whether the model assumptions are fit or not. We have already started with some model called graph, if you go by graph two-way scatter plot of all those things will give you a graph. So, graph then the fitted line can also be derived.

That fitted line if it is perfectly going now with the distribution then your model is fine, here it seems your model is not perfectly fit. Similarly normality in residuals I have already told you one command called kdensity, but in this case pnorm and or and qnorm to be to be given to check whether they are following normal distribution or not.

So, then kdensity and normal if you do it will be plotting like this and accordingly, we can understand. Similarly, standardized normal probability plot that is pnorm checks for non-normality in the middle range of residuals. In that case pnorm r then you can give title of that particular a chart with an in with the bracket about the normality normal check, it will give you information.

Some other assumptions checked are through multicollinearity I know we will explain in our other models as well, but at this moment if you go by the statistics books called VIF Variance Inflation Factor, the mean VIF value should be no less than 10.

If it is greater than 10 then there will be multicollinearity. So, the mean VIF value is here if it is very less; that means, your model is perfectly fine. So, the VIF command is an important

VIF that will give you the result like this. I am not drawing it, we will draw it sometimes in our other modules. Similarly testing for homoscedasticity usually we go for the Breusch Pagan test.

So, that basically detects heteroscedasticity in the model the null hypothesis is that the results are homoscedastic whether; which means the having constant variance. So, that command is called estat hettes. estat hettest my previous module on large scale data handling we have also included this, if you want further reading you can also refer to that as well.

But otherwise, if you simply run with this command you will get this kind of information. Our assumption is that our residuals are homoscedastic there is no variation in the residuals the standard deviation is constant. But here it seems that your result is significant the assumption is constant variance; that means, it is deviating it is actually rejecting your assumption. So that means, there is heteroscedasticity.

So, this is what is interpreted, and the rest you can just read between the line and I am sure you will enjoy reading.

And this is a very big lecture and we have skipped some of this section deliberately and I am I know that, if you go through you will have lots of queries and we will be happy to deal with them in our live session class. So, with this, I think it is time to close and I look forward to your queries.

Thank you.