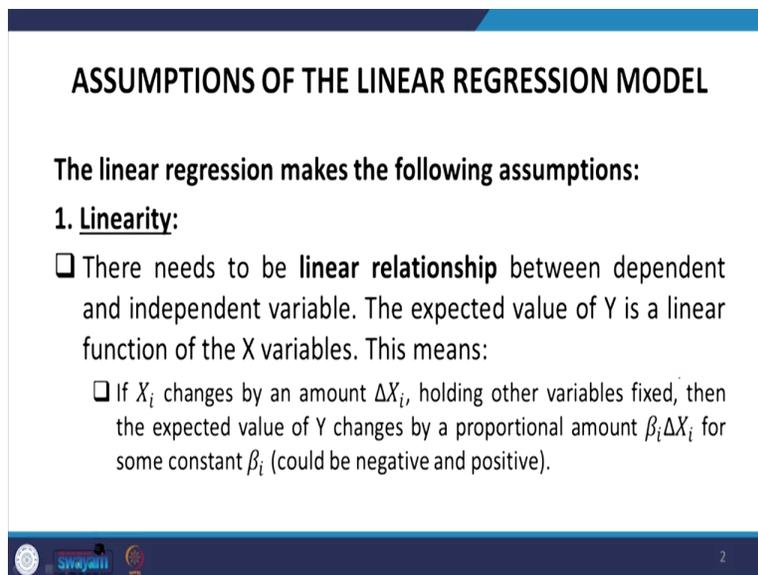


Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology Roorkee
Lecture 30
Linear Regression Analysis in STATA-III

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)



ASSUMPTIONS OF THE LINEAR REGRESSION MODEL

The linear regression makes the following assumptions:

1. Linearity:

- There needs to be **linear relationship** between dependent and independent variable. The expected value of Y is a linear function of the X variables. This means:
 - If X_i changes by an amount ΔX_i , holding other variables fixed, then the expected value of Y changes by a proportional amount $\beta_i \Delta X_i$ for some constant β_i (could be negative and positive).

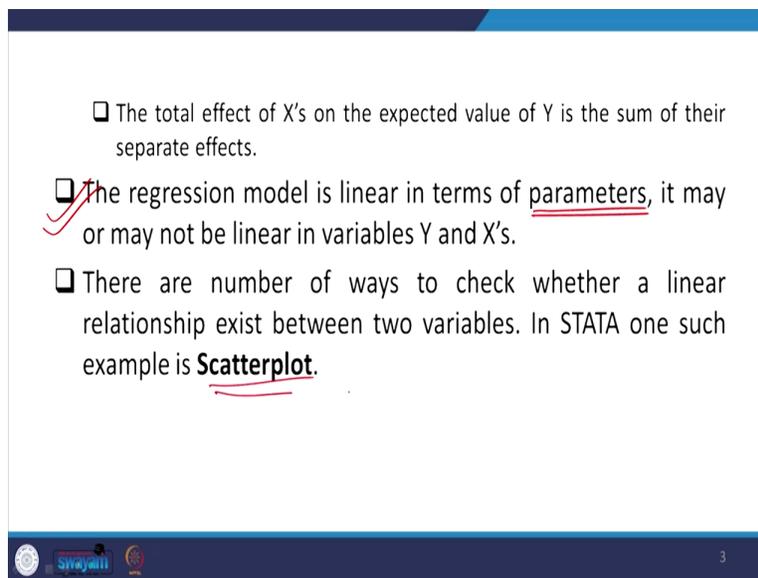
2

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 ΔX , 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 β_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)

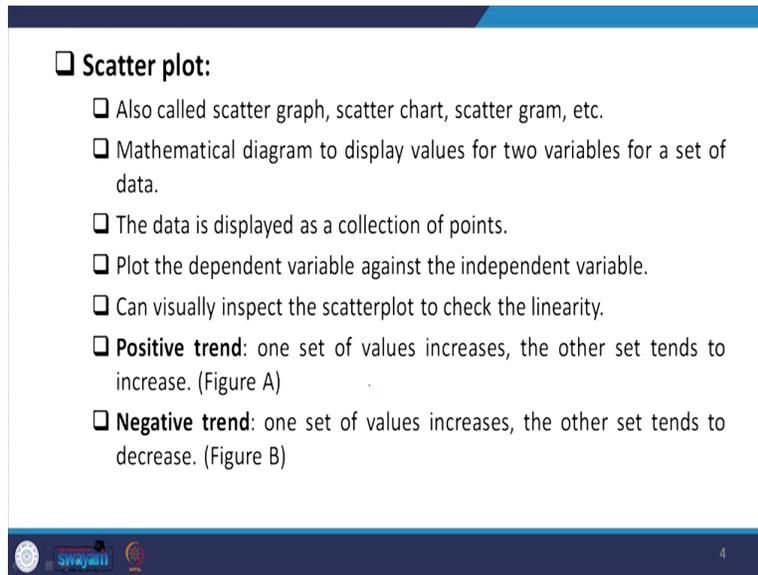


- The total effect of X 's on the expected value of Y is the sum of their separate effects.
- The regression model is linear in terms of parameters, it may or may not be linear in variables Y and X 's.
- There are number of ways to check whether a linear relationship exist between two variables. In STATA one such example is Scatterplot.

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)



Scatter plot:

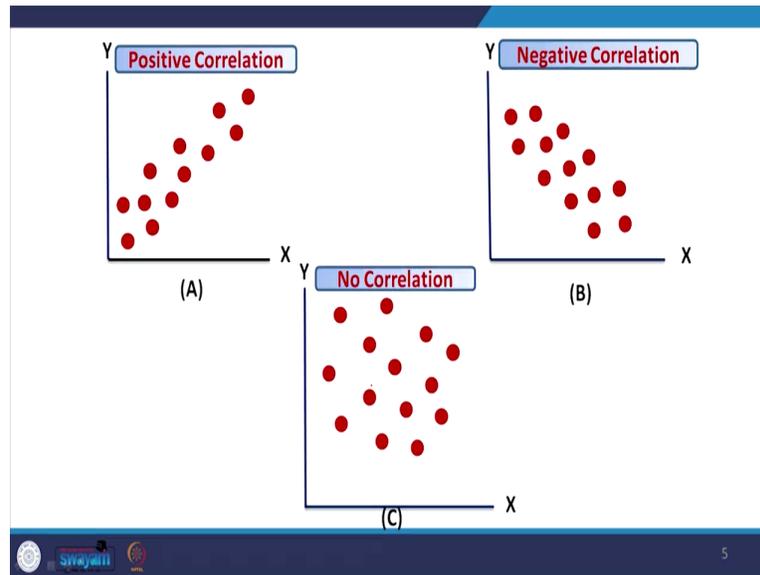
- ❑ Also called scatter graph, scatter chart, scatter gram, etc.
- ❑ Mathematical diagram to display values for two variables for a set of data.
- ❑ The data is displayed as a collection of points.
- ❑ Plot the dependent variable against the independent variable.
- ❑ Can visually inspect the scatterplot to check the linearity.
- ❑ **Positive trend:** one set of values increases, the other set tends to increase. (Figure A)
- ❑ **Negative trend:** one set of values increases, the other set tends to decrease. (Figure B)

4

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)

2. Independence of Error (Autocorrelation):

- ❑ The residuals are assumed to be uncorrelated with each other, which implies that the Y's are also uncorrelated.
- ❑ The correlation happens because of two reasons:
 - ❑ **Model misspecification**- If an important independent variable is omitted or if an incorrect functional form is used, the residuals may not be independent. So, it is suggested to use proper functional form or use of multiple regression.
 - ❑ **Time-sequenced data**- if regression analysis is performed on data taken over time, the residuals may be correlated. These kind of correlation are termed as serial correlation. So, it is common in time-series data analysis.

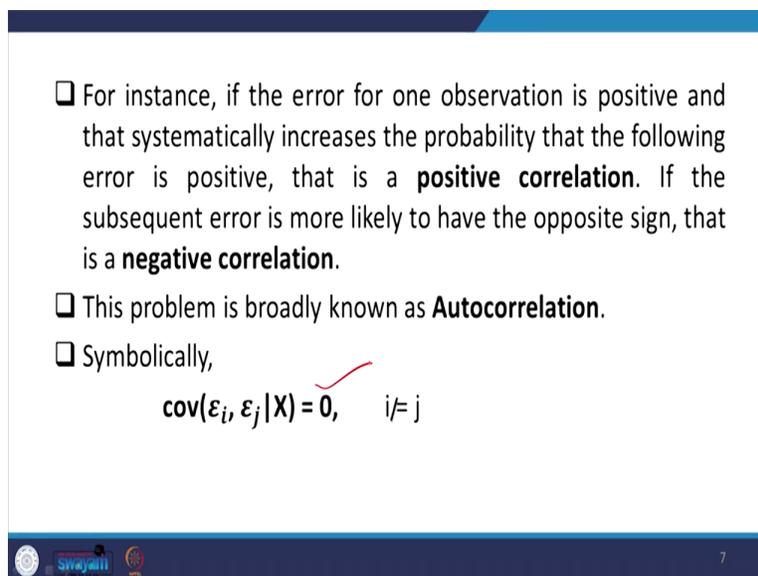
Coming to other backdrop of understanding the linear regression and its applications. The residuals are assumed to be; we are now going by another assumption called independence of error. So, independence of error also referred to as autocorrelation. Let us clarify autocorrelation

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, 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 is a **negative correlation**.

□ This problem is broadly known as **Autocorrelation**.

□ Symbolically,

$$\text{cov}(\varepsilon_i, \varepsilon_j | \mathbf{X}) = 0, \quad i \neq j$$

7

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 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)

3. The Error Term has Population Mean Zero:

- Given the values of the X variables, the expected or mean value of the error term is zero.

$$E(\epsilon_i | X) = 0$$

As a result of this assumption:

$$\begin{aligned} E(Y_i | X) &= \beta X + E(\epsilon_i | X) \\ &= \beta X \end{aligned}$$

In regression analysis, our main objective is to estimate this function.

- 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 is a **negative correlation**.

- This problem is broadly known as **Autocorrelation**.

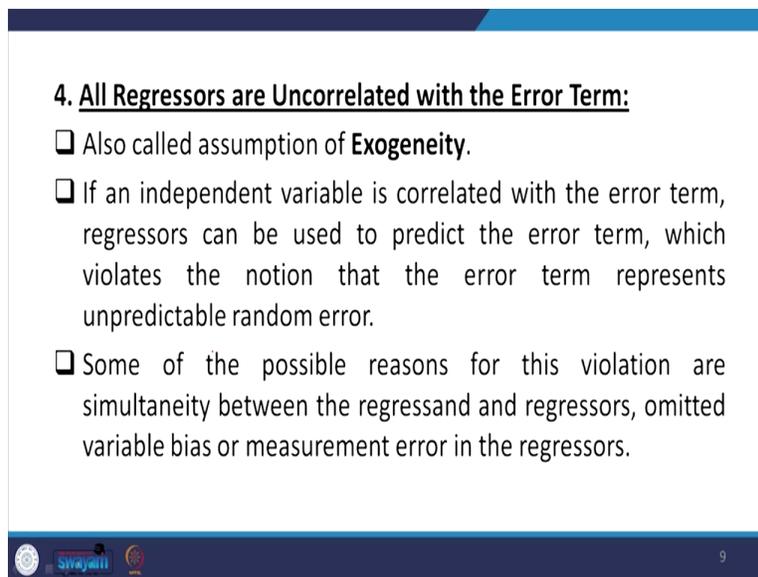
- Symbolically,

$$\text{cov}(\epsilon_i, \epsilon_j | X) = 0, \quad i \neq j$$

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 βX_i plus α or βX_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)



4. All Regressors are Uncorrelated with the Error Term:

- Also called assumption of **Exogeneity**.
- 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 error.
- Some of the possible reasons for this violation are simultaneity between the regressand and regressors, omitted variable bias or measurement error in the regressors.

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 exogeneity. 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)

5. Homoscedasticity:

- ❑ Homoscedasticity is a Greek word where “homo” means equal and “scedastic” means variance or scatter.
- ❑ The variance of each ϵ_i , given the values of X , is constant. Symbolically,
$$\text{var}(\epsilon_i | X) = \sigma^2$$
- ❑ If the variance changes, it is referred as **heteroscedasticity** (different scatter).
- ❑ To check this assumption we can create residual versus fitted value plot. If the spread of the residual increases in one direction and appears in cone shape, can be concluded as presence of heteroscedasticity.

The slide includes a diagram where the symbol σ^2 in the equation is circled in red, and another circled σ^2 is shown with a red arrow pointing to the word "heteroscedasticity" in the text below.

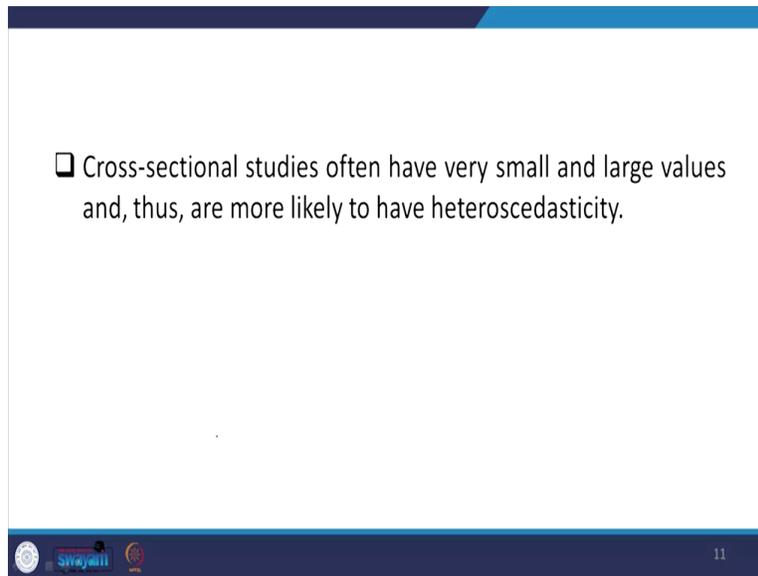
Coming to the issues of variance of the distribution error term, we are now going to emphasize the another assumption called homoscedasticity. The meaning of homoscedasticity derived from the Greek word with homo refers to equal and scedastic refers to variance or scatter.

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)

6. No Multicollinearity:

- ❑ There are *no perfect linear relationship among the X variables.*
- ❑ Perfect correlation suggests that two variables are different forms of the same variable. For example, games won and games lost have a perfect negative correlation (-1). The temperature in Fahrenheit and Celsius have a perfect positive correlation (+1).
- ❑ If these correlations are high enough, they can reduce the precision of the estimates in OLS linear regression.



12

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)

7. Normality of the Error Term:

- ❑ Error term follows the normal distribution with zero mean and constant variance σ^2 . Symbolically,

$$\varepsilon_i \sim N(0, \sigma^2)$$



13

6. No Multicollinearity:

- There are *no perfect linear relationship among the X variables.*
- Perfect correlation suggests that two variables are different forms of the same variable. For example, games won and games lost have a perfect negative correlation (-1). The temperature in Fahrenheit and Celsius have a perfect positive correlation (+1).
- If these correlations are high enough, they can reduce the precision of the estimates in OLS linear regression.



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 ϵ distribution follows normality with 0 mean and variance σ^2 .

(Refer Slide Time: 15:22)

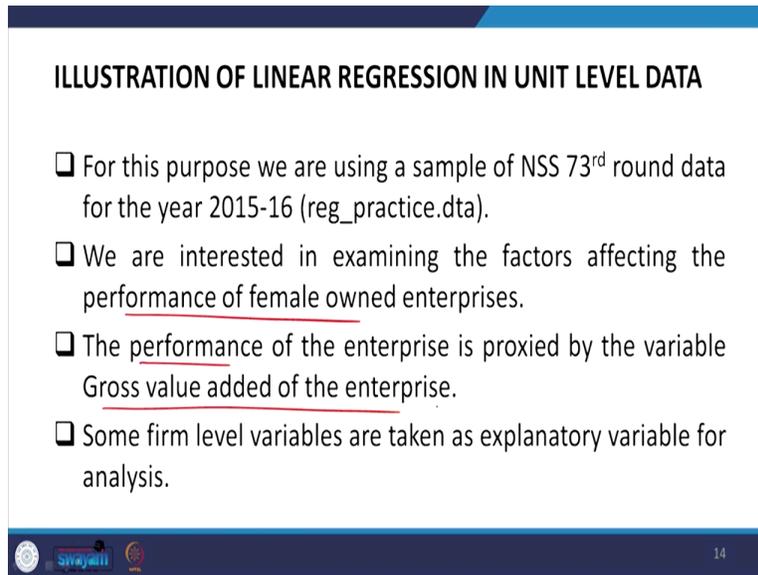


ILLUSTRATION OF LINEAR REGRESSION IN UNIT LEVEL DATA

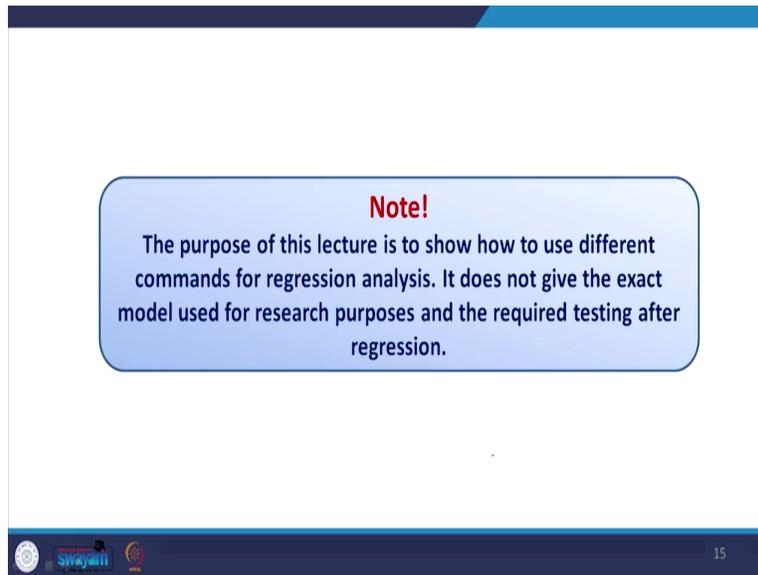
- ❑ For this purpose we are using a sample of NSS 73rd round data for the year 2015-16 (reg_practice.dta).
- ❑ We are interested in examining the factors affecting the performance of female owned enterprises.
- ❑ The performance of the enterprise is proxied by the variable Gross value added of the enterprise.
- ❑ Some firm level variables are taken as explanatory variable for analysis.

14

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)



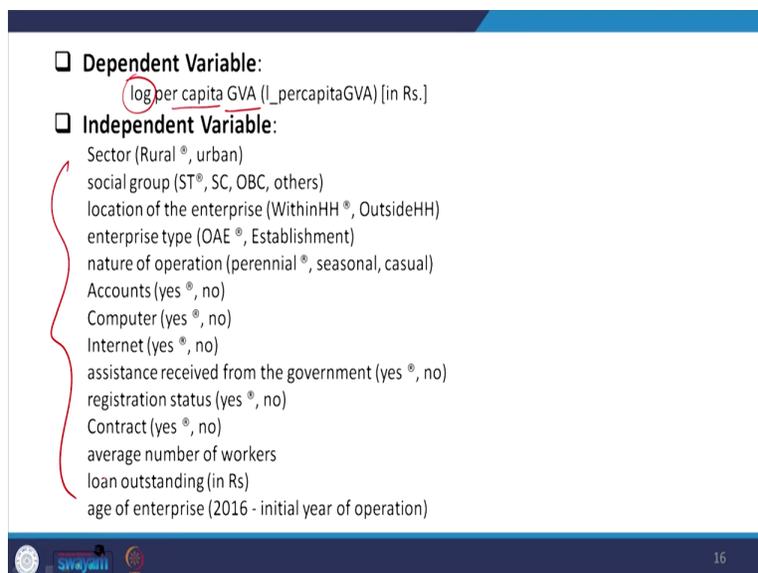
Note!

The purpose of this lecture is to show how to use different commands for regression analysis. It does not give the exact model used for research purposes and the required testing after regression.

15

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)



Dependent Variable:
log per capita GVA (l_percapitaGVA) [in Rs.]

Independent Variable:

- Sector (Rural^o, urban)
- social group (ST^o, SC, OBC, others)
- location of the enterprise (WithinHH^o, OutsideHH)
- enterprise type (OAE^o, Establishment)
- nature of operation (perennial^o, seasonal, casual)
- Accounts (yes^o, no)
- Computer (yes^o, no)
- Internet (yes^o, no)
- assistance received from the government (yes^o, no)
- registration status (yes^o, no)
- Contract (yes^o, no)
- average number of workers
- loan outstanding (in Rs)
- age of enterprise (2016 - initial year of operation)

16

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)

STEPS FOR RUNNING REGRESSION

1. Examine descriptive statistics
2. Look at relationship graphically and test correlation(s).
3. Run and interpret regression
4. Test regression assumptions

Swayam 17

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)

EXAMINATION OF DESCRIPTIVE STATISTICS

It is always advisable first to examine the variables in the model to check for possible errors.

begin by using the `describe` command to list various features of the variables to be used in the linear regression.

Type in command window:

`describe dependent variable independent variable`

`describe l_per capitaGVA total_worker`

For simple linear regression model, we are considering only two variables. Will consider remaining for the multiple regression.

18

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)

The screenshot shows the STATA 15.1 software interface. The command window contains the following text:

```
STATA 15.1 Copyright 1985-2017 StataCorp LLC
Statistica/Data Analysis
Special Edition
100-STAT2A-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)

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

. use "G:\Regression\reg_prac..."
. des l_perceptiva age_of_enterprise loan_outstanding nature_operat location_enterprise

+-----+-----+-----+-----+-----+-----+
| variable name | type | format | label | variable label |
+-----+-----+-----+-----+-----+
| l_perceptiva | float | %9.0g | | |
| age_of_enterp | float | %9.0g | | |
| loan_outstand | double | %10.0g | Amount outstanding as on last date of reference year (t.) |
| nature_operat | byte | %10.0g | nature_operat | Nature of operation |
| location_enter | byte | %10.0g | location | Location of the enterprise |
+-----+-----+-----+-----+-----+

Command
```

The Variables window on the right lists the following variables:

Name	Label
location_enter	Location of the enterprise
enterprise_by	Enterprise type
nature_operat	Nature of operation
accounts	Whether accounts
computer	Did the enterprise
internet	Did the enterprise
assistance_rev	Did the enterprise
registration_st	Whether registered
contract	Does the enterprise
GVA	Value(R)
total_worker	average number of
NIC_MAXOR	major nc activity
activity_group	

The screenshot shows the STATA 15.1 software interface with a keyboard overlay. The command window contains the following text:

```
STATA 15.1 Copyright 1985-2017 StataCorp LLC
Statistica/Data Analysis
Special Edition
100-STAT2A-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)

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

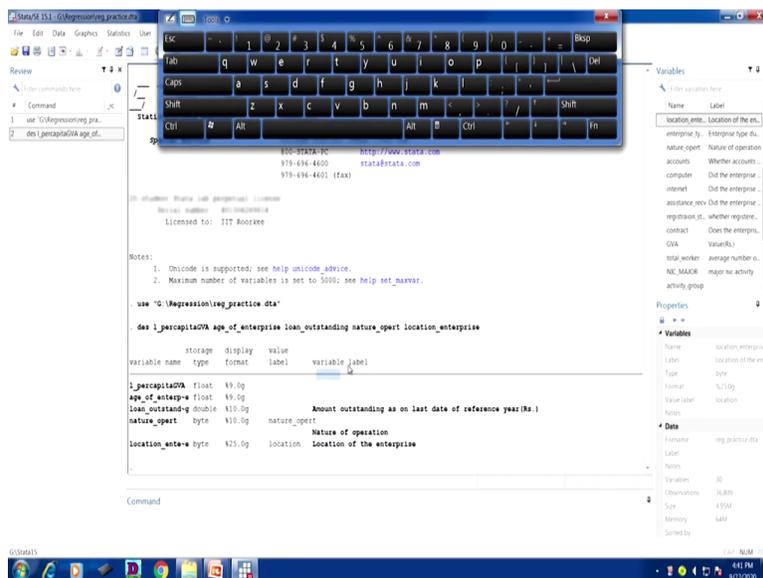
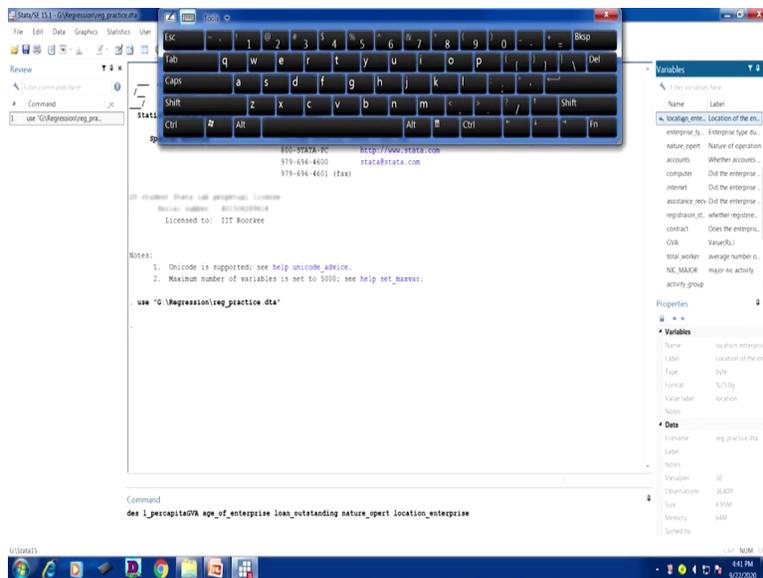
. use "G:\Regression\reg_prac..."
. des l_perceptiva

+-----+-----+-----+-----+-----+
| variable name | type | format | label | variable label |
+-----+-----+-----+-----+-----+
| l_perceptiva | float | %9.0g | | |
+-----+-----+-----+-----+-----+

Command
des l_perceptiva
```

The Variables window on the right lists the following variables:

Name	Label
registration_st	whether registered
contract	Does the enterprise
GVA	Value(R)
total_worker	average number of
NIC_MAXOR	major nc activity
activity_group	
loan_outstand	Amount outstand
min_GVA	
max_GVA	
normal_GVA	
age_of_enter	
perceptiva_gva	
l_perceptiva	



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)

- The describe command gives information about variable's storage type, display format, value label and variable label.
- The variable types and format columns indicate that all the data are numeric.
- You can not run regression on string variables.

```
. describe l_per capitaGVA total_worker
```

variable name	storage type	display format	value label	variable label
l_per capitaGVA	float	%9.0g		
total_worker	double	%10.0g		average number of workers-total

19

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)

□ It is essential in any data analysis to first check the data by summarize command.

`summarize varlist`

`summarize l_percapitaGVA total_worker`

```
. sum l_percapitaGVA total_worker
```

Variable	Obs	Mean	Std. Dev.	Min	Max
l_percapitaGVA	36,809	8.130111	1.028166	1.609438	12.44784
total_worker	36,809	1.910131	3.129021	1	208

The observation column tells us that there is no missing values as the number of observation in both the variables are same.

The screenshot shows the Stata software interface. The Command window contains the following text:

```
. use "G:\Regression\reg.jps."  
1 use "G:\Regression\reg.jps."  
2 des l_percapitaGVA age_of_...  
3 sum l_percapitaGVA age_of_...  
Notes:  
1. Unicode l...  
2. Maximum...  
... use "G:\Regression...  
... des l_percapitaGVA age_of_enterprise loan_outstanding nature_oper location_enterprise  
variable name type format label variable label  
l_percapitaGVA float 19.0g  
age_of_estre float 19.0g  
loan_outst double Amount outstanding as on last date of reference year (Rs.)  
nature_oper byte 19.0g nature_oper  
location_entre byte 125.0g location Location of the enterprise  
... sum l_percapitaGVA age_of_enterprise loan_outstanding nature_oper location_enterprise  
Variable Obs Mean Std. Dev. Min Max  
l_percapitaGVA 36 809 8.130111 1.028166 1.609438 12.44784  
age_of_estre 36 809 8.32636 7.226667 0 84  
loan_outst 36 809 28522.38 540847.2 0 5.63e+07  
nature_oper 36 809 1.922332 3.129021 1 3  
location_entre 36 809 1.388988 4.970209 1 2
```

The Variables list on the right shows the following variables:

- location_entre: Location of the enterprise
- enterprise_ty: Enterprise type
- nature_oper: Nature of operation
- accounts: Whether accounts
- copudate: Did the enterprise
- internet: Did the enterprise
- assistance_rev: Did the enterprise
- registration_st: whether registered
- contract: Does the enterprise
- GVA: Value(Rs.)
- total_worker: average number of
- NIC_MAJOR: major nic activity
- activity_group

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)

```

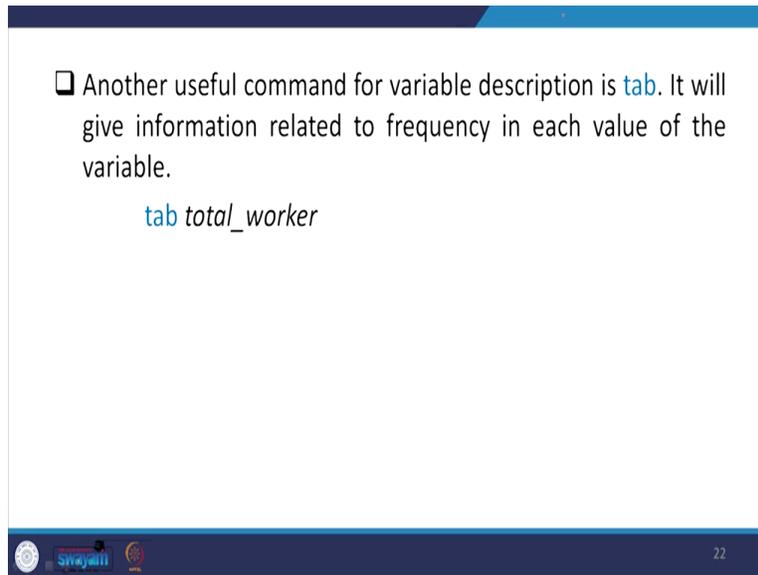
. sum l_percapitaGVA total_worker age_of_enterprise loan_outstanding sector social_g
> roup location_enterprise enterprise_type nature_opert accounts computer internet a
> ssistance_rcv registraion_status contract

```

Variable	Obs	Mean	Std. Dev.	Min	Max
l_percapit-A	36,809	8.130111	1.028166	1.609438	12.44784
total_worker	36,809	1.910131	3.129021	1	208
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
sector	36,809	1.518243	.4996739	1	2
social_group	36,809	2.993616	.9135279	1	4
location_e-e	36,809	1.358988	.4797103	1	2
enterprise-e	36,809	1.211878	.4086441	1	2
nature_opert	36,809	1.022331	.1870366	1	3
accounts	36,809	1.895732	.305612	1	2
computer	36,809	1.946834	.2243681	1	2
internet	36,809	1.955228	.206805	1	2
assistance-v	36,809	1.988046	.1086786	1	2
registraio-s	36,809	1.772311	.4193464	1	2
contract	36,809	1.906843	.2906557	1	2

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)



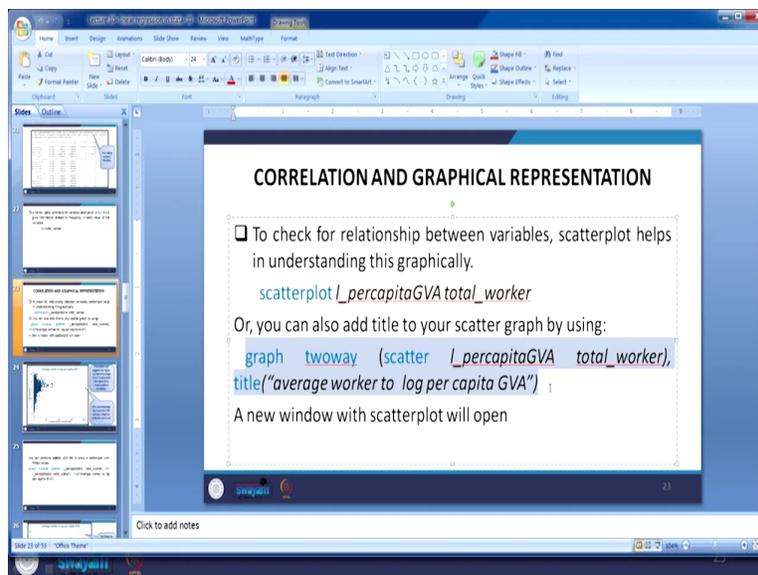
□ Another useful command for variable description is `tab`. It will give information related to frequency in each value of the variable.

```
tab total_worker
```

22

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)



CORRELATION AND GRAPHICAL REPRESENTATION

□ To check for relationship between variables, scatterplot helps in understanding this graphically.

```
scatterplot _percapitaGVA total_worker
```

Or, you can also add title to your scatter graph by using:

```
graph twoway (scatter _percapitaGVA total_worker),  
title("average worker to log per capita GVA")
```

A new window with scatterplot will open

23

StataSE11 - C:\Programing\practic01a

```

1. Unload all
2. Maximum memory
. use "G:\Regression\reg.prj"
. use "G:\Regression\reg.prj"
. des l_perceptiva age_of_entrprse loan_outstanding nature_operat location_enterprise
. sum l_perceptiva age_of_entrprse loan_outstanding nature_operat location_enterprise

```

Variable	Obs	Mean	Std. Dev.	Min	Max
l_perceptiva	36,809	0.130211	1.028166	1.609438	12.44784
age_of_entrprse	36,809	0.32836	7.224447	0	84
loan_outstanding	36,809	28522.38	540847.2	0	5.63e+07
nature_operat	36,809	1.022331	1875346	1	3
location_enterprise	36,809	1.358988	4797103	1	2

```

. scatterplot l_perceptiva total_worke
command scatterplot is unrecognized
r(199);

```

Command

Variables: Name: total_worke, Label: average number of workers, Type: double, Format: %10.0g, Value labels: none

Data: Filename: reg.practic01a, Labels: none, Variables: 10, Observations: 36,809, Size: 49304, Memory: 64M, Sorted by: none

StataSE11 - C:\Programing\practic01a

```

1. Unload all
2. Maximum memory
. use "G:\Regression\reg.prj"
. use "G:\Regression\reg.prj"
. des l_perceptiva age_of_entrprse loan_outstanding nature_operat location_enterprise
. sum l_perceptiva age_of_entrprse loan_outstanding nature_operat location_enterprise
. scatterplot l_perceptiva total_worke
command scatterplot is unrecognized
r(199);
. graph twoway (scatter l_perceptiva total_worke), title("average worke to log per capita GVA")

```

Variable	Obs	Mean	Std. Dev.	Min	Max
l_perceptiva	36,809	0.130211	1.028166	1.609438	12.44784
age_of_entrprse	36,809	0.32836	7.224447	0	84
loan_outstanding	36,809	28522.38	540847.2	0	5.63e+07
nature_operat	36,809	1.022331	1875346	1	3
location_enterprise	36,809	1.358988	4797103	1	2

```

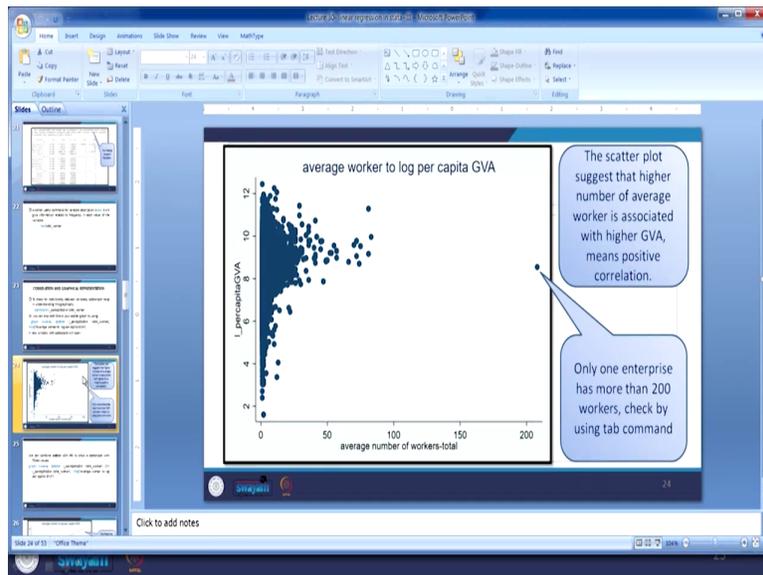
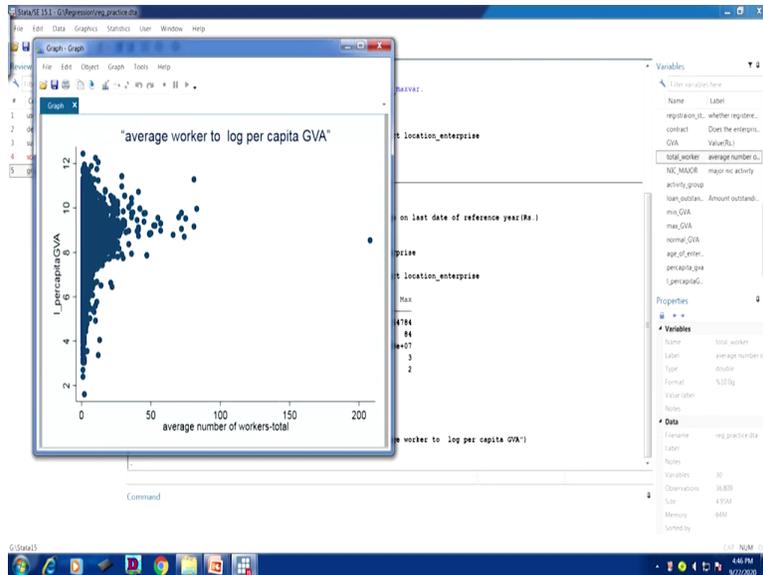
. graph twoway (scatter l_perceptiva total_worke), title("average worke to log per capita GVA")

```

Command

Variables: Name: total_worke, Label: average number of workers, Type: double, Format: %10.0g, Value labels: none

Data: Filename: reg.practic01a, Labels: none, Variables: 10, Observations: 36,809, Size: 49304, Memory: 64M, Sorted by: none



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 bivariate 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)

Click to add title

We can combine scatter with lfit to show a scatterplot with fitted values.

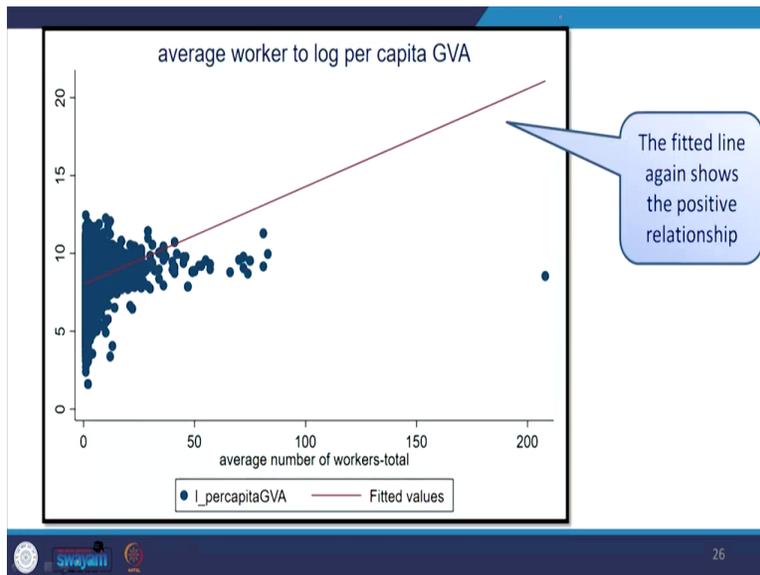
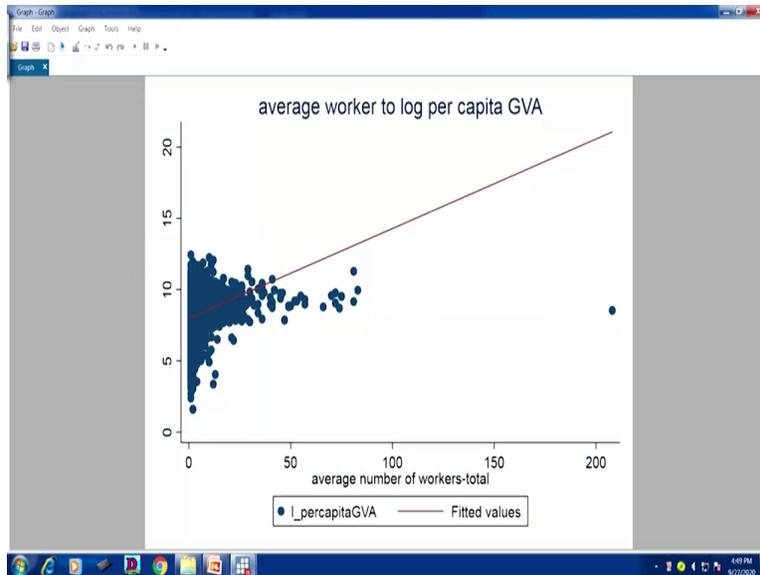
```
graph twoway (scatter l_per capitaGVA total_worker) (lfit l_per capitaGVA total_worker), title("average worker to log per capita GVA")
```

Click to add notes

```
1. use "G:\Regression\reg.prj"
2. des l_per capitaGVA age_of_...
3. sum l_per capitaGVA age_of_...
4. scatterplot l_per capitaGVA...
5. graph twoway (scatter l_per...
6. graph twoway (scatter l_per...

Variable      Obs      Mean      Std. Dev.      Min      Max
-----
l_per capitaG  36,809   8.130111   1.028166   1.609438   12.44784
age_of_ent-e  36,809   8.32636   7.224447       0         84
loan_outst-a  36,809  28522.38  540847.2     0   5.63e+07
nature_oper  36,809   1.022332   2870366       1         3
location_e-e  36,809   1.358988   4797103       1         2

. scatterplot l_per capitaGVA total_worker
. command: scatterplot l_per capitaGVA total_worker
. lfit1991:
. graph twoway (scatter l_per capitaGVA total_worker), title("average worker to log per capita GVA")
. graph twoway (scatter l_per capitaGVA total_worker) (lfit l_per capitaGVA total_worker), title("average worker to log per capita GVA")
```



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.

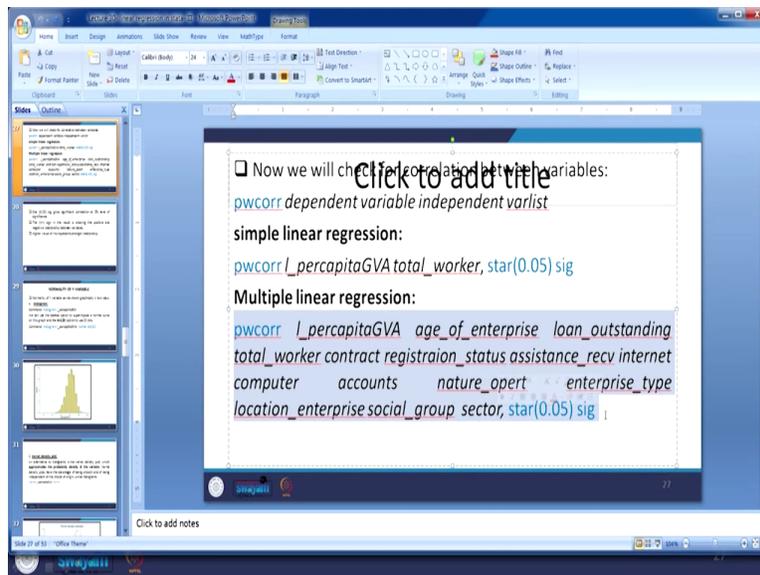
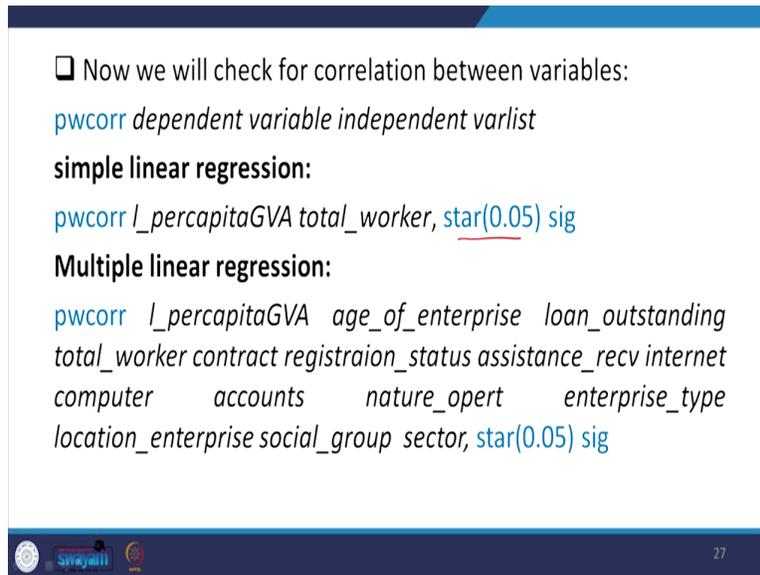
So this red line highlights fitted values. This shows that there exists a positive relationship. So, we are not spending much time on it and what it indicates is more important for our explanation and this is what we have explained. The fitted line shows the positive relationship as I just mentioned.

(Refer Slide Time: 29:41)

□ Now we will check for correlation between variables:
`pwcorr dependent variable independent varlist`

simple linear regression:
`pwcorr l_per capitaGVA total_worker, star(0.05) sig`

Multiple linear regression:
`pwcorr l_per capitaGVA age_of_enterprise loan_outstanding
total_worker contract registraion_status assistance_recv internet
computer accounts nature_opert enterprise_type
location_enterprise social_group sector, star(0.05) sig`



StataSE11 - C:\Programing\practic... [File Edit Data Graphics Statistics User Window Help]

Review

```

1 use "C:\Programing\pr...
2 des LpercapaGVA age_of...
3 sum LpercapaGVA age_of...
4 scatterplot LpercapaGVA...
5 graph hwekey scatter Lper...
6
7 pcor LpercapaGVA age...

```

	Internet	computer	accounts	nature_sport	enterprise_e	location_e	social_group	sector
Internet	1.0000							
computer	0.8123* 1.0000							
accounts	0.3904* 0.4477* 1.0000							
nature_sport	0.0188* 0.0231* 0.0203* 1.0000							
enterprise_e	-0.3314* -0.3812* -0.4281* -0.0452* 1.0000							
location_e	-0.2121* -0.2498* -0.2964* -0.0504* 0.4949* 1.0000							
social_group	-0.1150* -0.1281* -0.0932* -0.0373* 0.1030* 0.0021 1.0000							
sector	-0.0947* -0.0978* -0.1007* -0.0281* 0.1108* 0.0359* 0.1144* 1.0000							

Internet computer accounts nature-t enterprise locati-e social-g

command

Variables

- regpractic... Label: whether registe...
- contract Does the empres... Value(8)
- GVA Value(8)
- total_worke... average number of... Value(8)
- NIC_MAJOR... major inc activity
- activity_group
- loan_outstan... Amount outstand...
- min_GVA
- max_GVA
- normal_GVA
- age_of_entre...
- percapita_gva
- LpercapaG... Value(8)

Properties

Variables

- Name: total_worke...
- Label: average number of...
- Type: double
- Format: %10.0g
- Value label:
- Notes:

Data

- Filename: reg_practic... dia
- Label:
- Notes:
- Variables: 30
- Observations: 38,839
- Size: 49,934
- Memory: 64M
- Sorted by:

4:51 PM 9/22/2008

StataSE11 - C:\Programing\practic... [File Edit Data Graphics Statistics User Window Help]

Review

```

1 use "C:\Programing\pr...
2 des LpercapaGVA age_of...
3 sum LpercapaGVA age_of...
4 scatterplot LpercapaGVA...
5 graph hwekey scatter Lper...
6
7 pcor LpercapaGVA age...

```

	Internet	computer	accounts	nature_sport	enterprise_e	location_e	social_group	sector
Internet	1.0000							
computer	0.8123* 1.0000							
accounts	0.3904* 0.4477* 1.0000							
nature_sport	0.0188* 0.0231* 0.0203* 1.0000							
enterprise_e	-0.3314* -0.3812* -0.4281* -0.0452* 1.0000							
location_e	-0.2121* -0.2498* -0.2964* -0.0504* 0.4949* 1.0000							
social_group	-0.1150* -0.1281* -0.0932* -0.0373* 0.1030* 0.0021 1.0000							
sector	-0.0947* -0.0978* -0.1007* -0.0281* 0.1108* 0.0359* 0.1144* 1.0000							

Internet computer accounts nature-t enterprise locati-e social-g

command

Variables

- regpractic... Label: whether registe...
- contract Does the empres... Value(8)
- GVA Value(8)
- total_worke... average number of... Value(8)
- NIC_MAJOR... major inc activity
- activity_group
- loan_outstan... Amount outstand...
- min_GVA
- max_GVA
- normal_GVA
- age_of_entre...
- percapita_gva
- LpercapaG... Value(8)

Properties

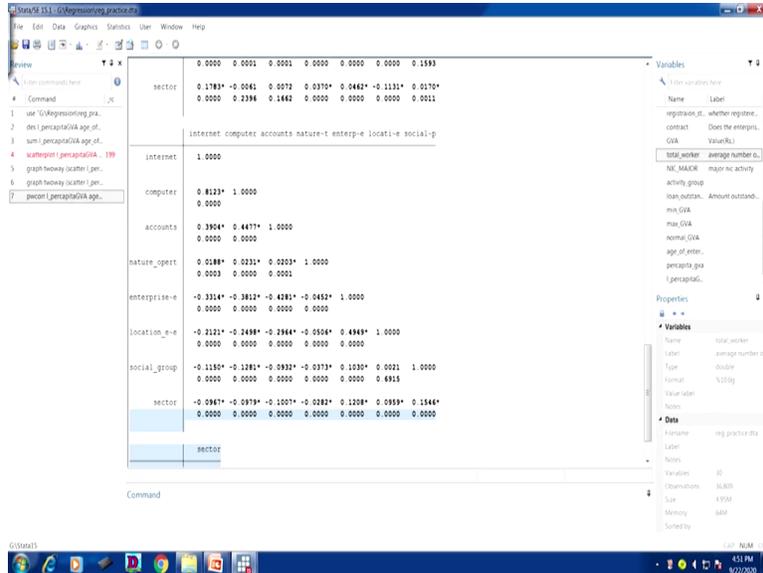
Variables

- Name: total_worke...
- Label: average number of...
- Type: double
- Format: %10.0g
- Value label:
- Notes:

Data

- Filename: reg_practic... dia
- Label:
- Notes:
- Variables: 30
- Observations: 38,839
- Size: 49,934
- Memory: 64M
- Sorted by:

4:51 PM 9/22/2008



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)

- ❑ Star (0.05) sig, gives significant correlation at 5% level of significance.
- ❑ The (+/-) sign in the result is showing the positive and negative relationship between variables.
- ❑ Higher value of rho represents stronger relationship.

The plus and minus sign regarding their relationship has already been mentioned. Somewhere we get positive. Somewhere we get negative. So, higher values of rho 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)

NORMALITY OF Y-VARIABLE

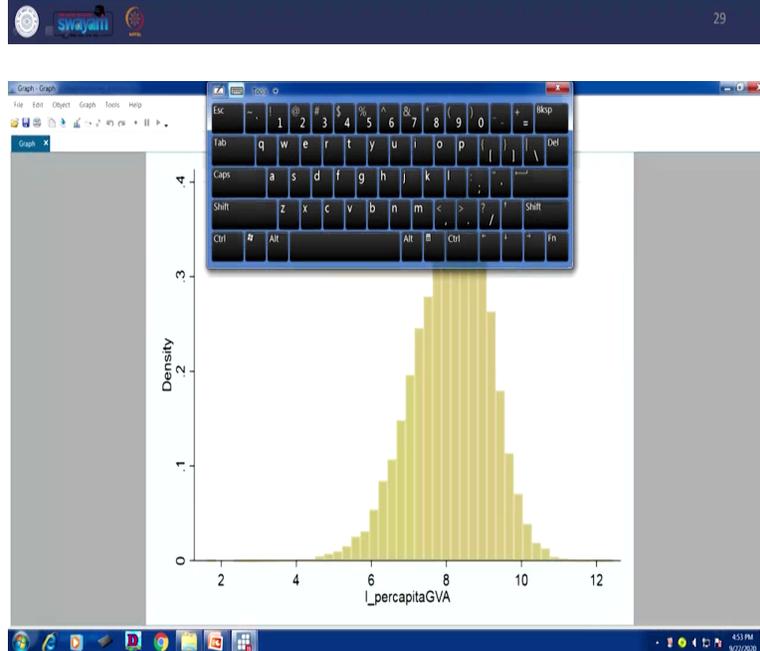
□ Normality of Y variable can be shown graphically in two ways:

1. Histogram:

Command: `histogram I_per capitaGVA`

We can use the **normal** option to superimpose a normal curve on this graph and the **bin(20)** option to use 20 bins.

Command: `histogram I_per capitaGVA, normal bin(20)`



Stata/11.1 - C:\Programing\practic... [log.c]

File Edit Data Graphics Statistics User Window Help

Review

1. use 'log' to log the results of the regression
 2. des _percapitaGVA age of...
 3. sum _percapitaGVA age of...
 4. scatterplot _percapitaGVA...
 5. graph htwway (scatter) _per...
 6. graph htwway (scatter) _per...
 7. percent _percapitaGVA age...
 8. histogram _percapitaGVA...
 9. histogram _percapitaGVA...

computer 0.0
 accounts 0.0
 nature_opert 0.0
 enterprise-e 0.0
 location_e-e -0.2121* -0.2498* -0.2964* -0.0504* 0.4949* 1.0000
 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
 social_group -0.1150* -0.1281* -0.0932* -0.0373* 0.1030* 0.0021 1.0000
 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.4915
 sector -0.0947* -0.0979* -0.1007* -0.0282* 0.1208* 0.0859* 0.1544*
 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

sector
 sector 1.0000

histogram _percapitaGVA
 (bin=45, start=-1.4594379, width=24.093342)

histogram _percapitaGVA, normal
 (bin=25, start=-1.4594379, width=34.192219)

command

Variables

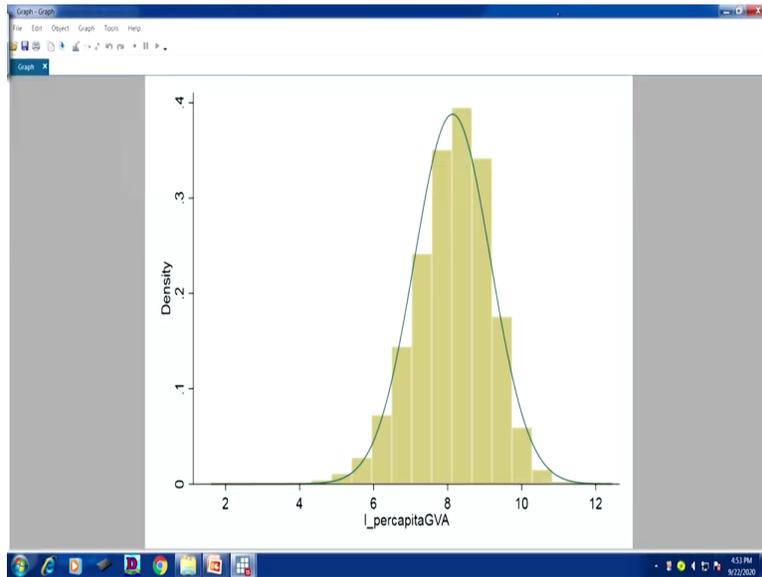
Name	Label
registrarship_it	whether registered...
operatad	Does the enterprise...
GVA	Value(RL)
total_workers	average number of...
NIC_MAXKOR	major sic activity
activity_group	
loan_outstak	Amount outstand...
min_GVA	
max_GVA	
normal_GVA	
age_of_enter	
percapita_gva	
_percapitaG...	

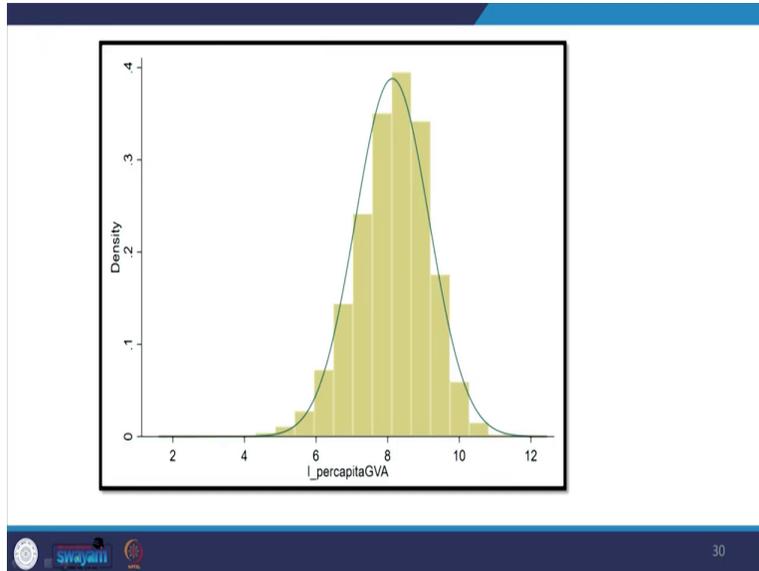
Properties

Variables

Name: _percapitaGVA
 Label:
 Type: float
 Format: %9.0g
 Value label:
 Notes:
 Data
 Examine: log practice.dta
 Label:
 Notes:
 Variables: 30
 Observations: 86,809
 Size: 4,9304
 Memory: 6484
 Sorted by:

@(stata11) 4:53 PM 9/22/2008





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)

2. kernel density plot:

An alternative to histograms is the kernel density plot, which **approximates the probability density of the variable**. Kernel density plots have the advantage of being smooth and of being independent of the choice of origin, unlike histograms.

`kdensity l_per capitaGVA, normal`

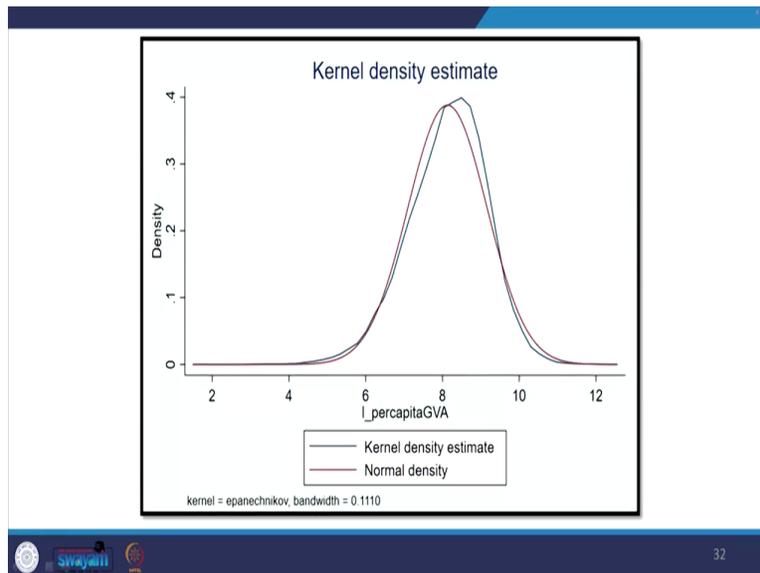
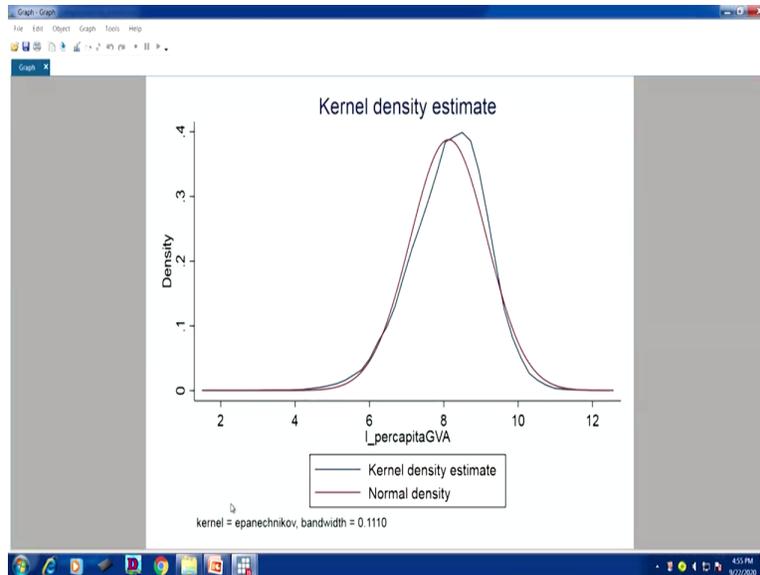
The screenshot displays the Stata software interface. The main window shows the following regression results for the variable 'sector':

Variable	coefficient	std. err.	z	prob > z	[95% conf. interval]			
computer	0.8123*	1.0000	0.0000					
accounts	0.3904*	0.4477*	1.0000					
nature_oper	0.0188*	0.0231*	0.0203*	1.0000				
enterprise_e	-0.3314*	0.3812*	-0.4281*	0.0452*	1.0000			
location_e	-0.3121*	0.2488*	-0.2964*	0.0504*	0.4945*	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*	0.1544*	1.0000

Below the regression results, the following commands are shown in the Command window:

```
. histogram l_per capitaGVA  
(bins=45, start=1.4094279, width=24083342)  
. histogram l_per capitaGVA, normal, bins(50)  
(bins=50, start=1.4094279, width=84130219)  
. kdensity l_per capitaGVA, normal
```

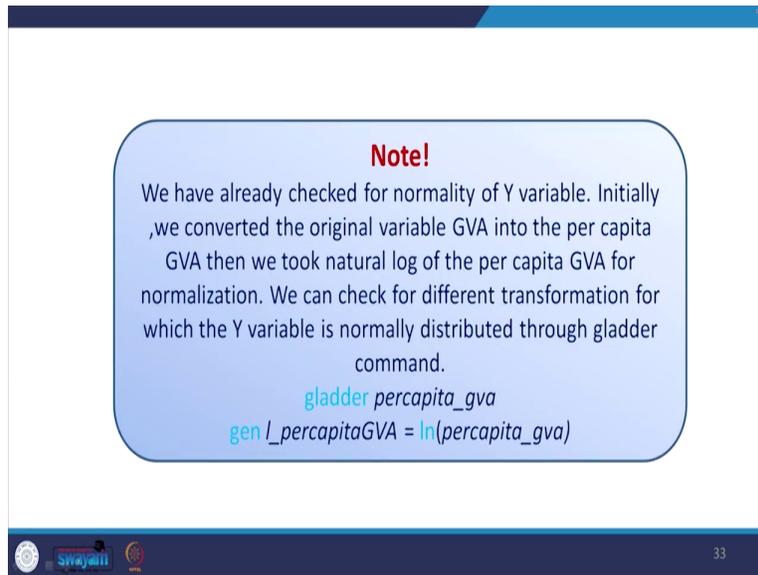
The right-hand side of the interface shows the Variables list and Properties panel for the variable 'l_per capitaGVA'.



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)



Note!

We have already checked for normality of Y variable. Initially ,we converted the original variable GVA into the per capita GVA then we took natural log of the per capita GVA for normalization. We can check for different transformation for which the Y variable is normally distributed through gladder command.

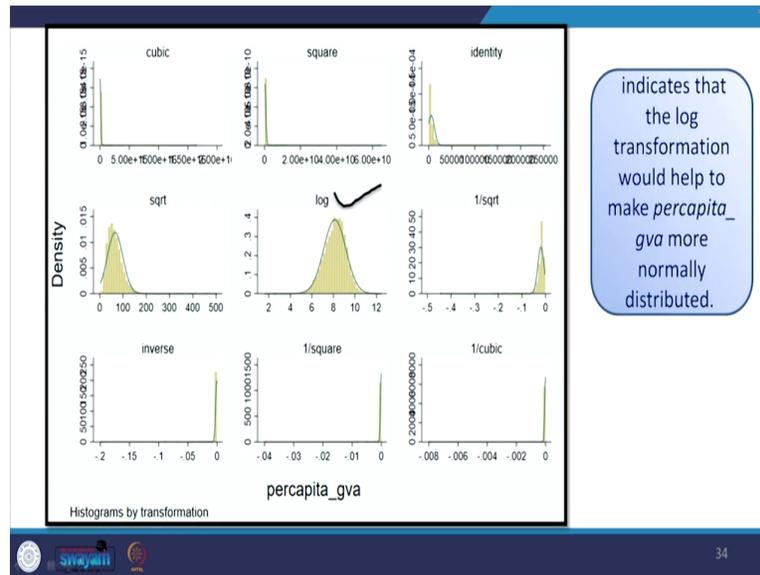
```
gladder percapita_gva  
gen l_percapitaGVA = ln(percapita_gva)
```

33

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)

MODEL ESTIMATION

- Beginning with simple linear regression in which we only have one predictor variable.
- In Stata, the dependent variable is listed immediately after the **regress** command followed by one or more predictor variables.

```
regress l_percapitaGVA total_worker  
eststo model1
```

Stata 11.1 - O:\regression_practice.dta

File Edit Data Graphics Statistics User Window Help

Review

```

1 use 'O:\regression_prac...
2 des l_perceptiva age of...
3 sum l_perceptiva age of...
4 scatterplot l_perceptiva...
5 graph twoway (scatter) l...
6 graph twoway (scatter) l...
7 percent l_perceptiva age...
8 histogram l_perceptiva...
9 histogram l_perceptiva...
10 identify l_perceptiva age...
11 search eststo

```

Variable	1	2	3	4	5	6	7	8	9	10	11
accounts	0.3924*	0.4477*	1.0000								
nature_operit	0.0188*	0.0231*	0.0203*	1.0000							
enterprise_e	-0.3514*	-0.3812*	-0.4281*	-0.0453*	1.0000						
location_e-e	-0.2121*	-0.2498*	-0.2364*	-0.0504*	0.4949*	1.0000					
social_group	-0.1150*	-0.1281*	-0.0932*	-0.0373*	0.1030*	0.0021	1.0000				
sector	-0.0947*	-0.0979*	-0.1007*	-0.0282*	0.1208*	0.0959*	0.1544*	1.0000			

. histogram l_perceptiva @V
 (Bin=45, start=1.4594379, width=24093342)

. histogram l_perceptiva @V, normal bin(20)
 (Bin=20, start=1.4594379, width=14192219)

. identify l_perceptiva @V, normal

. search eststo

Command

Variables

Name	Label
registration_it	whether registered...
contract	Does the enterprise...
GVA	Value(RL)
total_worker	average number of...
NIC_MAXOR	major nic activity
activity_group	
loan_outstak	Amount outstanding
mix_GVA	
max_GVA	
normal_GVA	
age_of_enter	
perceptiva_gva	
l_perceptiva @V	

Properties

Name: l_perceptiva @V
 Type: float
 Format: %9.0g
 Value label:
 Notes:
 Data: Estimation reg_practice.dta
 Variables: 30
 Observations: 36,809
 Size: 493M
 Memory: 64M
 Sorted by:

4:58 PM 9/22/2008

Stata 11.1 - O:\regression_practice.dta

File Edit Data Graphics Statistics User Window Help

Viewer: search eststo

search eststo

search for eststo (manual: [F] search)

search of official help files: FAQs, Examples, Stz, and STB

ST-14-2 st0085_2 Software update for `estadd`, `estout`, `eststo`, `esttab`, `esttab`
 (help `estadd`, `estout`, `eststo`, `esttab` if installed) - B. Jann
 Q2/07 82 7123:1951
 new features added and various problems fixed

ST-1-2 st0085_1 Making regression tables simplified
 (help `estadd`, `estout`, `eststo`, `esttab` if installed) - B. Jann
 Q2/07 82 7123:227-244
 introduces the `eststo` and `esttab` commands (stemming from
estadd), 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_3 from <http://www.stata-journal.com/software/sj14-2>
 STJ4-2 st0085_2, Update: Making regression... / Update: Making regression
 tables from stored / estimates / by Ben Jann, University of Bern /
 Support: jann@fsoz.unibe.ch / After installation, type `help estout`,
`esttab`, `eststo`, `estadd`, and `estpost`

st0085_1 from <http://www.stata-journal.com/software/sj3-2>
 STJ3-2 st0085_1, Update: Making regression tables simplified / Update:
 Making regression tables simplified / by Ben Jann, ETH Zurich / Support:
 jann@fsoz.unibe.ch / After installation, type `help estout`, `esttab`,

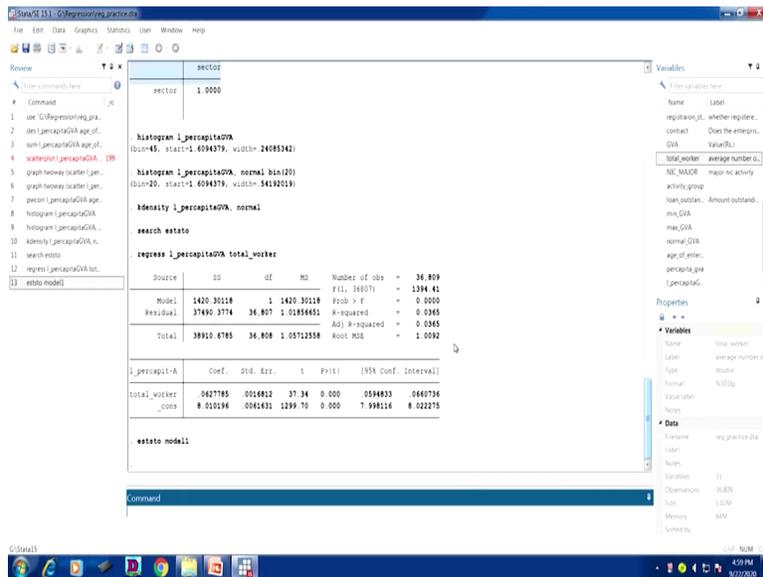
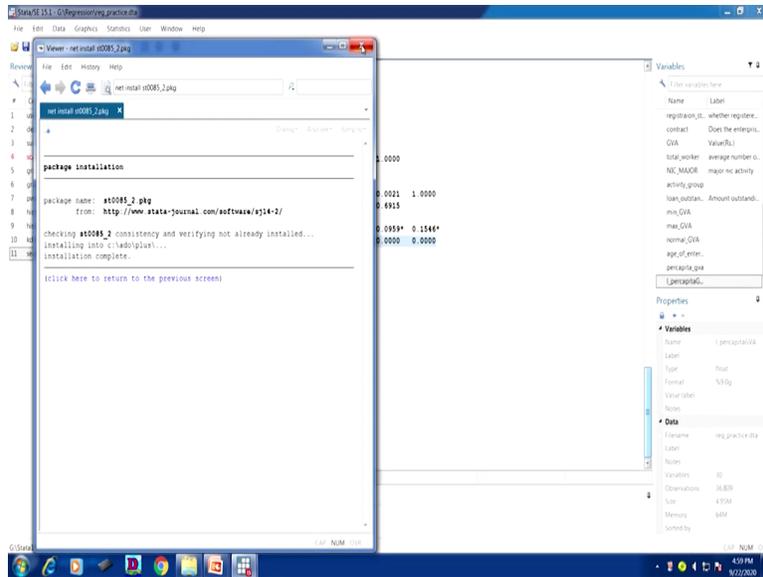
Variables

Name	Label
registration_it	whether registered...
contract	Does the enterprise...
GVA	Value(RL)
total_worker	average number of...
NIC_MAXOR	major nic activity
activity_group	
loan_outstak	Amount outstanding
mix_GVA	
max_GVA	
normal_GVA	
age_of_enter	
perceptiva_gva	
l_perceptiva @V	

Properties

Name: l_perceptiva @V
 Type: float
 Format: %9.0g
 Value label:
 Notes:
 Data: Estimation reg_practice.dta
 Variables: 30
 Observations: 36,809
 Size: 493M
 Memory: 64M
 Sorted by:

4:58 PM 9/22/2008



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 predictor 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.

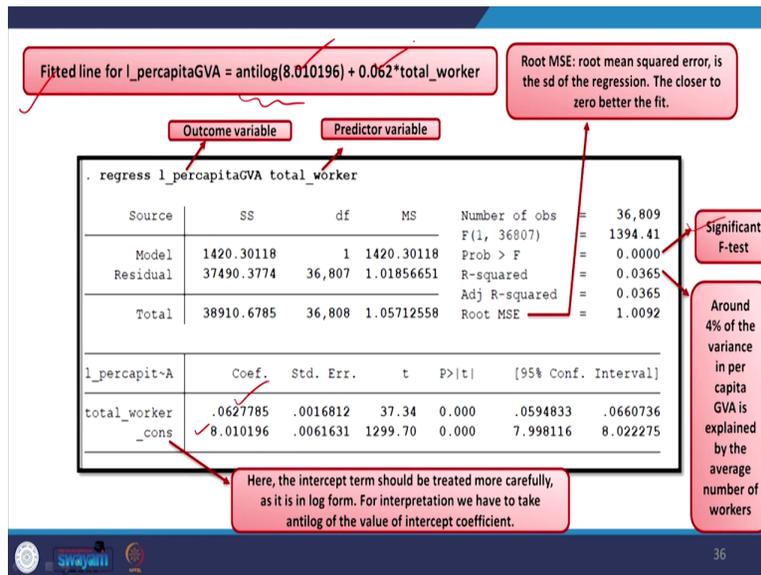
One of the very important suggestions to you that please try to run another regression command to it after regress so that you can compare all the models, eststo command of that model 1, similarly, in the model 2, I am going to show you in model 2. So, let us see, if it is not there, we

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)



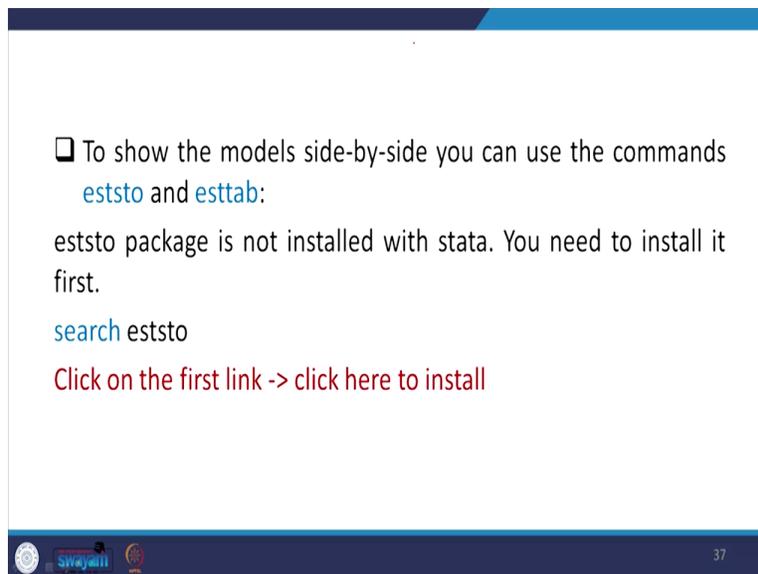
So, here is the result and what is important to be mentioned, everything is given here, that we start with the command here as outcome variable is here, then the predictor variable is here highlighted very clearly. This result is also gives the significance level, the F-test value and it suggests that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



❑ To show the models side-by-side you can use the commands `eststo` and `esttab`:

`eststo` package is not installed with stata. You need to install it first.

`search eststo`

Click on the first link -> [click here to install](#)

37

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)

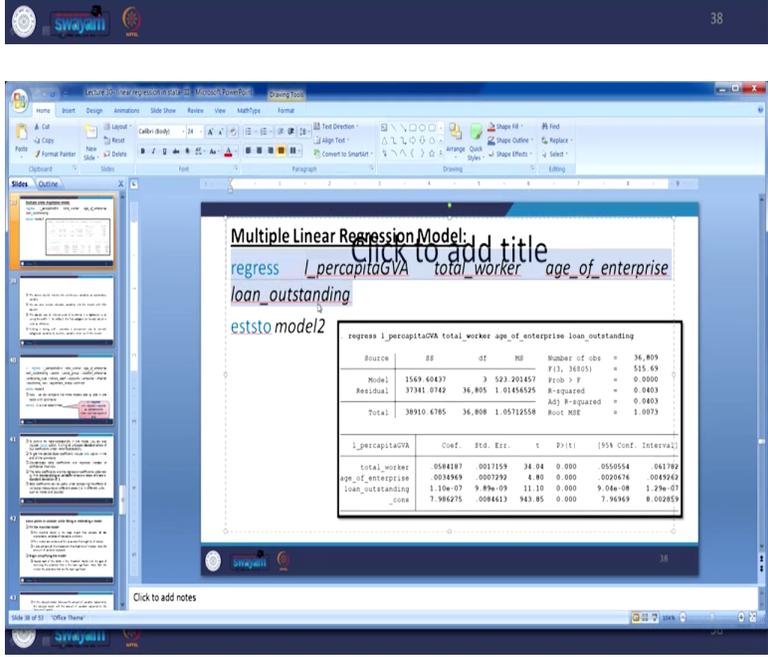
Multiple Linear Regression Model:

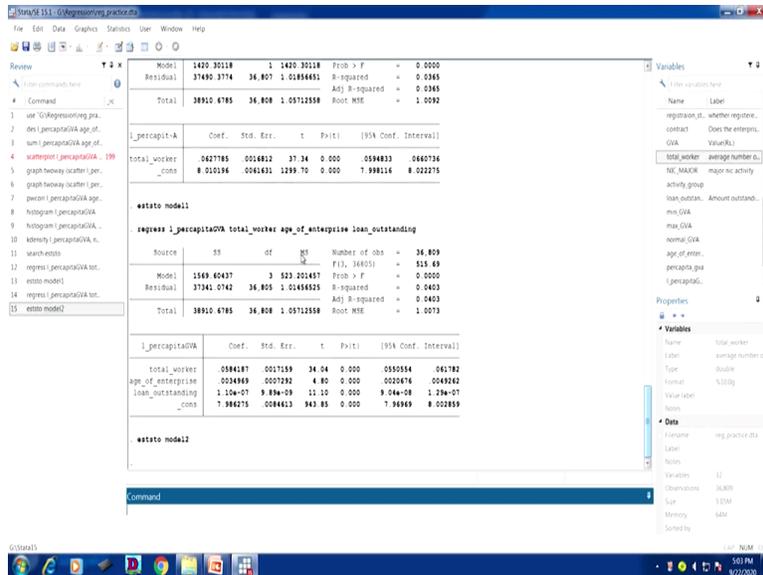
```
regress l_percapitaGVA total_worker age_of_enterprise  
loan_outstanding  
eststo model2
```

```
. regress l_percapitaGVA total_worker age_of_enterprise loan_outstanding
```

Source	SS	df	MS	Number of obs =	36,809
Model	1569.60437	3	523.201457	F(3, 36805) =	515.69
Residual	37341.0742	36,805	1.01456525	Prob > F =	0.0000
				R-squared =	0.0403
				Adj R-squared =	0.0403
				Root MSE =	1.0073
Total	38910.6785	36,808	1.05712558		

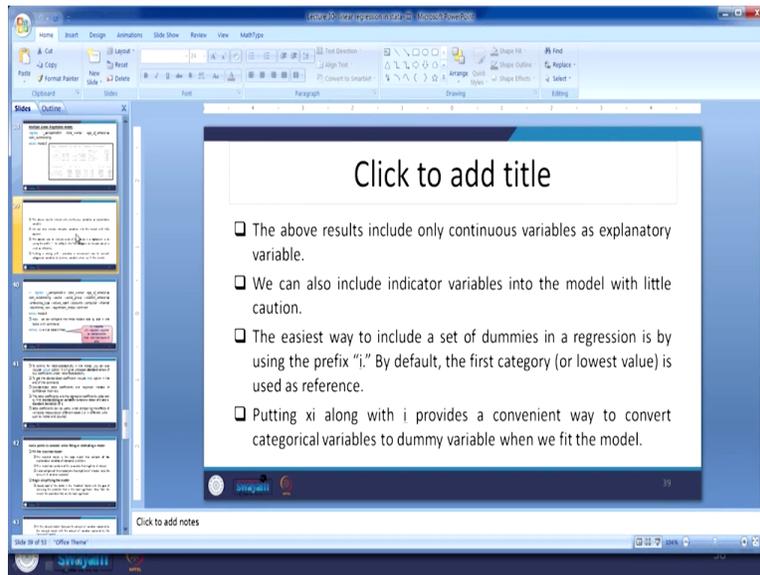
l_percapitaGVA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
total_worker	.0584187	.0017159	34.04	0.000	.0550554 .061782
age_of_enterprise	.0034969	.0007292	4.80	0.000	.0020676 .0049262
loan_outstanding	1.10e-07	9.89e-09	11.10	0.000	9.04e-08 1.29e-07
_cons	7.986275	.0084613	943.85	0.000	7.96969 8.002859





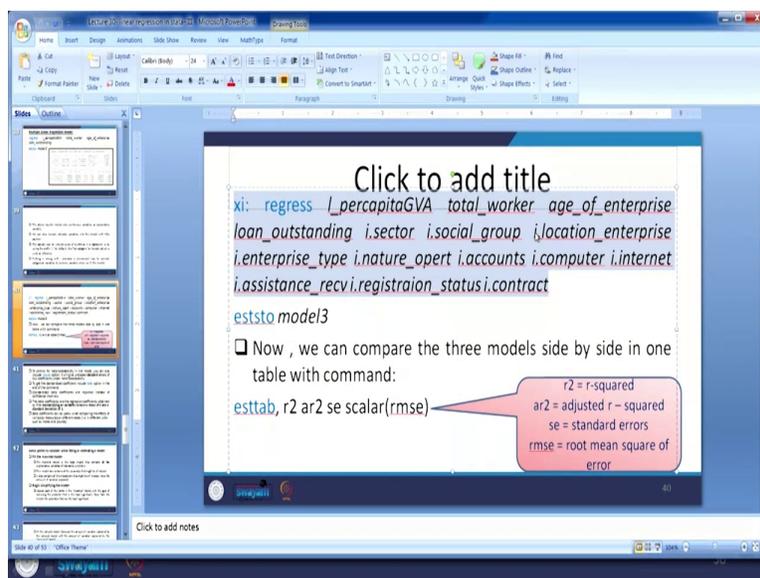
So, let us understand another approach for another model also. So, regress then we will copy this or we will operate these variables and I wanted to show that how eststo command works. So, it is here. So let us go to that window. Simply we will copy it and operate this through our STATA and STATA gives 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)



Stata 11.1 - O:\regression\practice.dta

File Edit Data Graphics Statistics User Window Help

Review

```

1 use 'O:\regression\practice.dta'
2 des _lpercapitaGVA age_of_
3 sum _lpercapitaGVA age_of_
4 scatterplot _lpercapitaGVA,
5 graph htwyway (scatter) _lper_
6 graph htwyway (scatter) _lper_
7 percent _lpercapitaGVA age_of_
8 histogram _lpercapitaGVA
9 histogram _lpercapitaGVA,
10 idensity(_lpercapitaGVA,n,
11 search=estab
12 regress(_lpercapitaGVA tot_
13 eststo model1
14 regress(_lpercapitaGVA tot_
15 eststo model2
16 regress(_lpercapitaGVA tot_
17 eststo model3

```

Source

Source	SS	df	MS	Number of obs	F(1, 3619)	Prob > F	R-squared	Adj R-squared	Root MSE
Model	12243.1411	17	720.194760	36,809	993.58	0.0000	0.3146	0.3143	853.37
Residual	26667.5375	36,191	738.08214						
Total	38910.6786	36,800	1.05712558						

Dependent Variable: _lpercapitaGVA

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_total_worker	-0.155178	0.014832	-8.98	0.000	-0.184169 - 0.131888
_age_of_enterprise	0.048973	0.006195	8.07	0.000	0.037803 0.060216
_loan_outstanding	8.42e-08	8.41e-09	10.02	0.000	6.77e-08 1.01e-07
_i.sector_2	-221743	2099594	-24.38	0.000	-2039288 2390702
_i.social_group_2	-4602884	6193003	-23.84	0.000	-4979275 - 4222492
_i.social_group_3	-2812272	2155541	-18.08	0.000	-3117136 - 2507407
_i.social_group_4	-222004	2163777	-13.54	0.000	-2941048 - 1899032
_i.location_2	4402816	6121303	40.47	0.000	4380581 4435051
_i.enterprise_2	3523321	624564	24.20	0.000	323794 3808702
_i.nature_op_2	-2929279	6466302	-4.23	0.000	-3824421 - 1984137
_i.nature_op_3	8542399	6549851	11.90	0.000	7619023 9483576
_i.accounts_2	-2221837	6271548	-12.84	0.000	-2973875 - 1477789
_i.computer_2	-9326709	6355613	-2.83	0.008	-1633712 - 6239699
_i.internet_2	-2515048	6389630	-4.80	0.000	-3239537 - 1790516
_i.assistance_recv	6849273	6411523	2.11	0.238	-1475869 9562416
_i.registraion_status	3555737	6238249	18.79	0.000	3712789 3338485
_i.contract_2	4744939	6156411	30.34	0.000	443842 505156
_cons	8.476285	6536737	157.92	0.000	8.371083 8.581487

eststo model3

command

Variables

Name	Label
registraion_status	whether registra...
contract	Does the enterpr...
GVA	Value(\$)
total_worker	average number of...
NIC_MAXOR	major nic activity
activity_group	
loan_outstanc	Amount outstand...
min_GVA	
normal_GVA	
age_of_enter	
percapitaGVA	

Properties

Variables

Name	Label
total_worker	average number of...
GVA	Value(\$)
%50log	

Data

Filename	log
reg_practice.dta	

Variables: 47
Observations: 36,809
Size: 1.58K
Memory: 64K
Sorted by:

Click to add title

```

xi: regress _lpercapitaGVA 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

Stata 11.1 - C:\Programing\practice.dta

```

Review
+-----+-----+-----+-----+
+ | _lnature_g-3 | -2303279 | 0444802 | -4.23 | 0.000 | -3824421 | -1394137 |
+ | _lnature_g-3 | -6541239 | 0548851 | -11.90 | 0.000 | -7619223 | -5463576 |
+ | _laccounts_2 | -2221837 | 0171568 | -12.54 | 0.000 | -2573875 | -1877759 |
+ | _lcomputer_2 | -936709 | 0395613 | -2.43 | 0.008 | -143372 | -623969 |
+ | _linternet_2 | -2515049 | 0368622 | -6.80 | 0.000 | -3239537 | -1790516 |
+ | _lassistant_2 | -086973 | 0412522 | -2.11 | 0.035 | -1475849 | -6262476 |
+ | _lregistra_2 | -3525737 | 0136249 | -25.73 | 0.000 | -3772789 | -3238485 |
+ | _lcontact_2 | 474499 | 0194412 | 2.49 | 0.000 | 443642 | 505156 |
+ | _lcons | 8.476285 | 0336737 | 127.92 | 0.000 | 8.371263 | 8.581487 |
+-----+-----+-----+-----+

. eststo model3

. esttab, x2 a12 no scalar(m3)

(1) (2) (3)
1 _lpercapit-A 1 _lpercapit-A 1 _lpercapit-A
total_worker 0.0428*** 0.0584*** -0.0151***
(0.002168) (0.00212) (0.002168)
age_of_ent-e 0.00350*** 0.00350***
(0.000729) (0.000480)
loan_outst-g 0.000000101*** 8.42e-08***
(9.89e-09) (8.42e-09)
_sector_2 0.222***
(0.00910)
_social_g-2 -0.460***
(0.0193)
_social_g-3 -0.281***

```

Command

Stata 11.1 - C:\Programing\practice.dta

```

Review
+-----+-----+-----+-----+
+ | _lnature_g-3 | (0.0447) |
+ | _lnature_g-3 | -0.654*** |
+ | _lnature_g-3 | (0.0550) |
+ | _laccounts_2 | -0.223*** |
+ | _laccounts_2 | (0.0178) |
+ | _lcomputer_2 | -0.0937** |
+ | _lcomputer_2 | (0.0256) |
+ | _linternet_2 | -0.282*** |
+ | _linternet_2 | (0.0370) |
+ | _lassistant_2 | -0.0869* |
+ | _lassistant_2 | (0.0412) |
+ | _lregistra_2 | -0.351*** |
+ | _lregistra_2 | (0.0136) |
+ | _lcontact_2 | 0.474*** |
+ | _lcontact_2 | (0.0156) |
+ | _lcons | 8.476*** |
+ | _lcons | (0.00216) | (0.00464) | (0.00371) |
+-----+-----+-----+-----+
n 36809 36809 36809
F-rsq 0.037 0.040 0.315
adj-r-sq 0.036 0.040 0.316
mse 1.009 1.007 0.851

Standard errors in parentheses
* p<.05, ** p<.01, *** p<.001

```

Command

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 `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, you can also include **robust** option. It will give unbiased **standard errors** of OLS coefficients under heteroscedasticity.
- ❑ To get the standardized coefficient include **beta** option in the end of the command.
- ❑ Standardized beta coefficients are reported instead of confidence intervals.
- ❑ 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.**
- ❑ Beta coefficients can be useful when comparing the effects of variables measured on different scales (i.e. in different units such as inches and pounds)

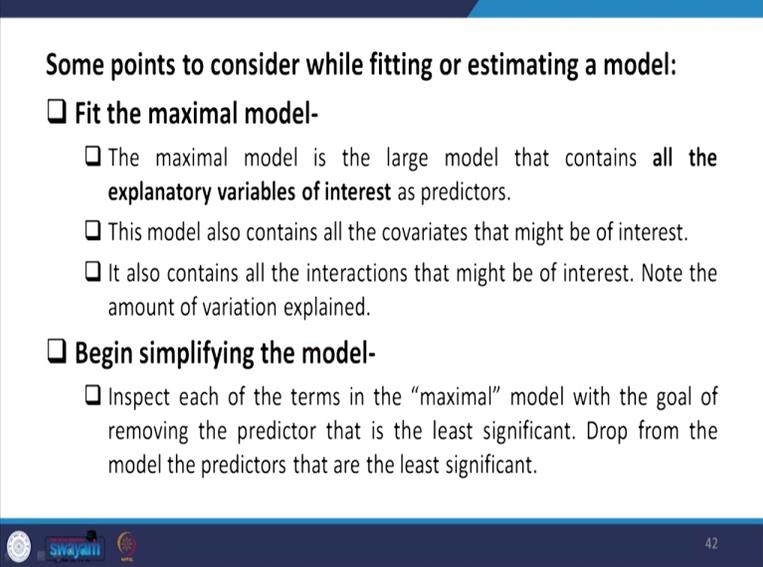


41

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)



Some points to consider while fitting or estimating a model:

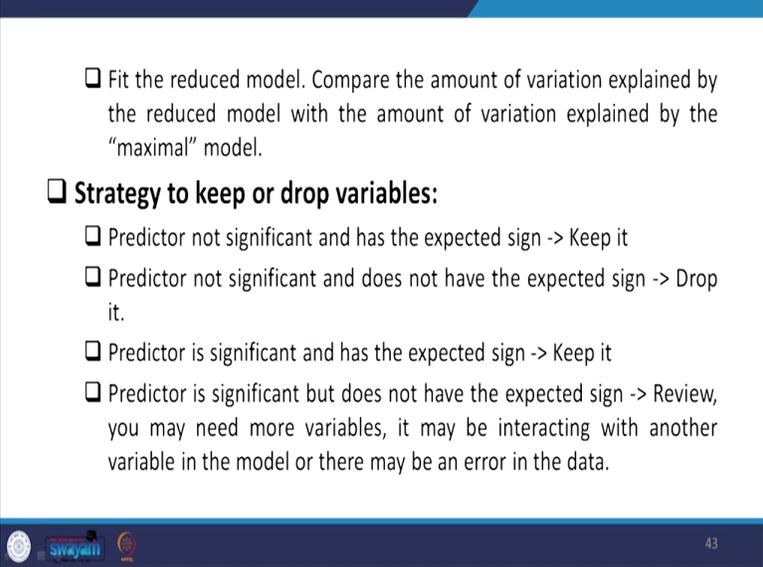
- ❑ **Fit the maximal model-**
 - ❑ The maximal model is the large model that contains **all the explanatory variables of interest** as predictors.
 - ❑ This model also contains all the covariates that might be of interest.
 - ❑ It also contains all the interactions that might be of interest. Note the amount of variation explained.
- ❑ **Begin simplifying the model-**
 - ❑ Inspect each of the terms in the “maximal” model with the goal of removing the predictor that is the least significant. Drop from the model the predictors that are the least significant.

42

So, there are some points for consideration in order to understand the fitting or estimating 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)



Fit the reduced model. Compare the amount of variation explained by the reduced model with the amount of variation explained by the “maximal” model.

Strategy to keep or drop variables:

- Predictor not significant and has the expected sign -> Keep it
- Predictor not significant and does not have the expected sign -> Drop it.
- Predictor is significant and has the expected sign -> Keep it
- Predictor is significant but does not have the expected sign -> Review, you may need more variables, it may be interacting with another variable in the model or there may be an error in the data.

43

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 the model is will depend on how well it predicts Y , the linearity of the model and the behavior of the residuals.

□ Generating predicted values of \hat{Y} after running regression

`predict _l_per capitaGVA_hat`

□ Generating values of residual:

`predict r, resid`

The screenshot displays the Stata software interface. The main window shows the command window with the following text:

```
. nl regress _l_per capitaGVA total_worker age_of_enterprise loan_outstanding i. sector i. social_group i. location_enterprise i. ente  
> rprise_type i. nature_of_ent i. accounts i. computer i. internet i. assistance_new i. registration_status i. contract  
i. sector      _i.factor_1-3 (naturally coded: _i.factor_1 omitted)  
i. location_en i. location_1-4 (naturally coded: _i.location_1 omitted)  
i. enterprise_e i. enterprise_1-2 (naturally coded: _i.enterprise_1 omitted)  
i. nature_of_ent i. nature_of_1-3 (naturally coded: _i.nature_of_1 omitted)  
i. accounts     i. accounts_1-2 (naturally coded: _i.accounts_1 omitted)  
i. computer     i. computer_1-2 (naturally coded: _i.computer_1 omitted)
```

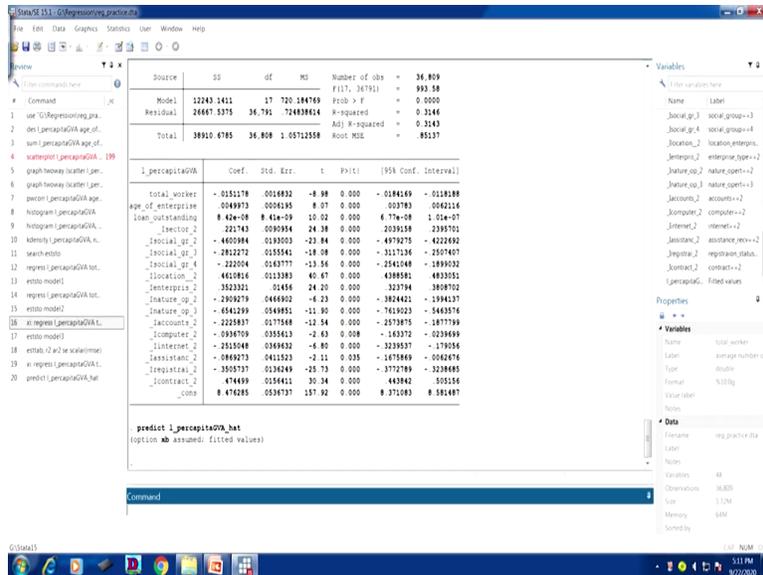
The Results window shows the following regression output:

	0.016***	7.986***	0.474***
	(0.00416)	(0.00846)	(0.0537)
N	36809	36809	36809
R-sq	0.037	0.040	0.315
adj. R-sq	0.036	0.040	0.314
F(3)	1.009	1.007	0.882

The Command window shows the command used to generate predicted values and residuals:

```
. nl predict _l_per capitaGVA_hat  
. nl predict r, resid
```

The Variables window shows the list of variables in the dataset, including `total_worker`, `age_of_enterprise`, `loan_outstanding`, `sector`, `social_group`, `location_enterprise`, `enterprise_type`, `nature_of_enterprise`, `accounts`, `computer`, `internet`, `assistance_new`, `registration_status`, and `contract`.



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 draw 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 underscored 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)

CHECKING MODEL ASSUMPTION AND FIT

- Model checks require you have just fit the model you are checking.
- Save the predicted values of Y as Y-hat.
- Plot observed vs. predicted values of Y variable:
`Graph twoway (scatter $\hat{l}_{percapitaGVA}$ $l_{percapitaGVA}$) (fit $l_{percapitaGVA}$ $\hat{l}_{percapitaGVA}$), title("Model Check")`

45

The screenshot shows a Beamer presentation slide with the same content as the first image. The slide is titled "CHECKING MODEL ASSUMPTION AND FIT" and contains three bullet points. The third bullet point includes a code snippet for a Stata graph. The slide is displayed within a Beamer window, with a slide navigation pane on the left and a status bar at the bottom. The status bar shows "Slide 45 of 53" and "Office Theme".

CHECKING MODEL ASSUMPTION AND FIT

- Model checks require you have just fit the model you are checking.
- Save the predicted values of Y as Y-hat.
- Plot observed vs. predicted values of Y variable:
`Graph twoway (scatter $\hat{l}_{percapitaGVA}$ $l_{percapitaGVA}$) (fit $l_{percapitaGVA}$ $\hat{l}_{percapitaGVA}$), title("Model Check")`

45

Click to add notes

Stata/MP 15.1 - O:\regression\practic01.d

File Edit Data Graphics Statistics User Window Help

Review

Source SS df MS Number of obs = 36,809
F(17, 36791) = 993.58
Model 12243.1451 17 720.18469 Prob > F = 0.0000
Residual 24667.5375 36,792 724.83614 R-squared = 0.3246
Total 36910.6826 36,800 1.0512558 Adj R-squared = 0.3143
Root MSE = .85137

	_l_perceptaGVA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
total_worke		-0.015178	0.0048932	-4.89	0.000	-0.0241659 - 0.0118108
age_of_ent		0.049973	0.0061955	8.07	0.000	0.037803 0.0621216
loan_outstand		8.424e-08	6.414e-09	10.02	0.000	6.774e-08 1.024e-07
_insector_2		221.743	0.090364	24.38	0.000	202.9518 239.5701
_social_of_3		-440.094	0.230029	-23.84	0.000	-447.9275 - 432.2602
_social_of_2		-282.272	0.155542	-18.08	0.000	-311.7136 - 250.7407
_social_of_4		-222.004	0.163777	-13.56	0.000	-254.4048 - 189.6032
_location_2		461.0816	0.113389	40.67	0.000	438.9581 483.2051
_enterprise_2		353.3251	0.4166	84.20	0.000	337.914 368.7372
_nature_op_2		-295.979	0.446902	-4.23	0.000	-382.4421 - 199.4137
_nature_op_3		-654.299	0.548851	-11.90	0.000	-763.9233 - 544.6576
_accounts_2		-222.897	0.177548	-12.54	0.000	-257.3875 - 187.7999
_computer_2		-0.84709	0.036613	-2.31	0.022	-1.433172 - 0.260999
_internet_2		-251.5048	0.036632	-4.80	0.000	-323.9537 - 179.056
_assistant_2		-0.89379	0.411523	-2.11	0.035	-1.675869 - 0.062476
_registrar_2		-365.579	0.126249	-28.79	0.000	-377.2789 - 353.8885
_contract_2		474.489	0.154412	30.34	0.000	449.742 505.236
_cons		8.474285	0.536737	157.92	0.000	8.371083 8.581487

- predict _l_perceptaGVA_hat
 (option xb assumed: fitted values)
 - Graph twoway (scatter _l_perceptaGVA_hat _l_perceptaGVA) if (_l_perceptaGVA_hat > _l_perceptaGVA), ltitle("Model Check")

command

Variables

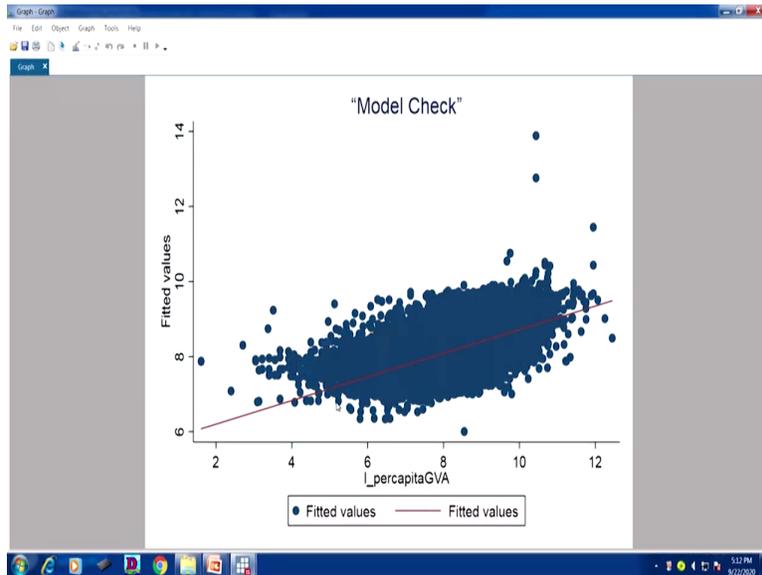
Name Label
 _social_of_3 social_group=3
 _social_of_4 social_group=4
 _location_2 location_enterp...
 _enterprise_2 enterprise_age=2
 _nature_op_2 nature_opent=2
 _nature_op_3 nature_opent=3
 _accounts_2 accounts=2
 _computer_2 computer=2
 _internet_2 internet=2
 _assistant_2 assistance_prov=2
 _registrar_2 registrar_status...
 _contract_2 contract=2
 _l_perceptaGVA Fitted values

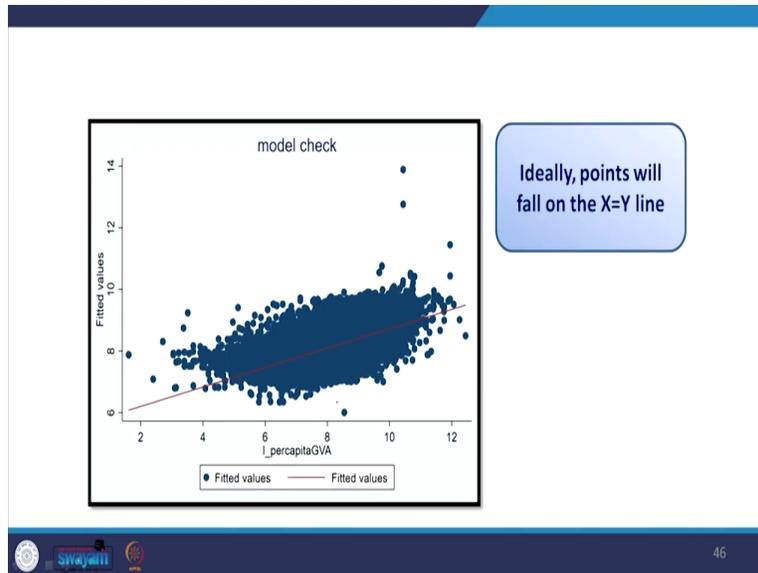
Properties

Name total_worke
 Label average number of
 Type double
 Format %10.0g
 Value label
 Notes

Data

Filename reg_practic01.dta
 Label
 Notes
 Variables 48
 Observations 36,809
 Size 3.72M
 Memory 64M
 Sorted by





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)

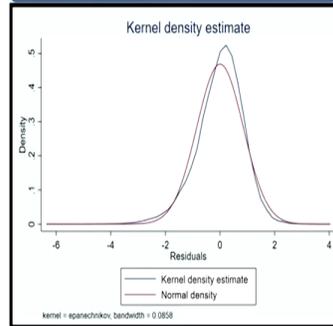
Check for Normality in Residuals:

Three graphs will help us check

For normality in residuals:

Kdensity, pnorm and qnorm

Kernel density: **kdensity** r, normal



Review

1 use 'O' to Regression practice.prj
2 des i_percepta07a age_of_...
3 sum i_percepta07a age_of_...
4 nlreg i_percepta07a ...
5 graph twoway (scatter) i_per...
6 graph twoway (scatter) i_per...
7 predict i_percepta07a_hat
8 histogram i_percepta07a_...
9 kernel i_percepta07a_...
10 identify i_percepta07a_...
11 search eststo
12 regress i_percepta07a_hat...
13 eststo model1
14 regress i_percepta07a_hat...
15 eststo model2
16 nlreg i_percepta07a_hat...
17 eststo model3
18 esttab i2 w2 se scalar(mse)
19 nlreg i_percepta07a_hat...
20 predict i_percepta07a_hat...
21 graph twoway (scatter) i_per...
22 predict r_resid
23 identify normal

	Residual	Total	26667.5375	36.791	724838614	R-squared	Adj. R-squared	Root MSE
						0.3146	0.3143	4812.7

Variables

Name Label

total_worker average number of...

age_of_entrepria amount of outstand...

loan_outstanding amount of outstand...

director_2

director_3

director_4

director_5

director_6

director_7

director_8

director_9

director_10

director_11

director_12

director_13

director_14

director_15

director_16

director_17

director_18

director_19

director_20

director_21

director_22

director_23

director_24

director_25

director_26

director_27

director_28

director_29

director_30

director_31

director_32

director_33

director_34

director_35

director_36

director_37

director_38

director_39

director_40

director_41

director_42

director_43

director_44

director_45

director_46

director_47

director_48

director_49

director_50

Properties

Variables

Name total_worker

Label average number of...

Type double

Format %30.0g

Value Label

None

Data

Filename reg_practice.do

Label

Notes

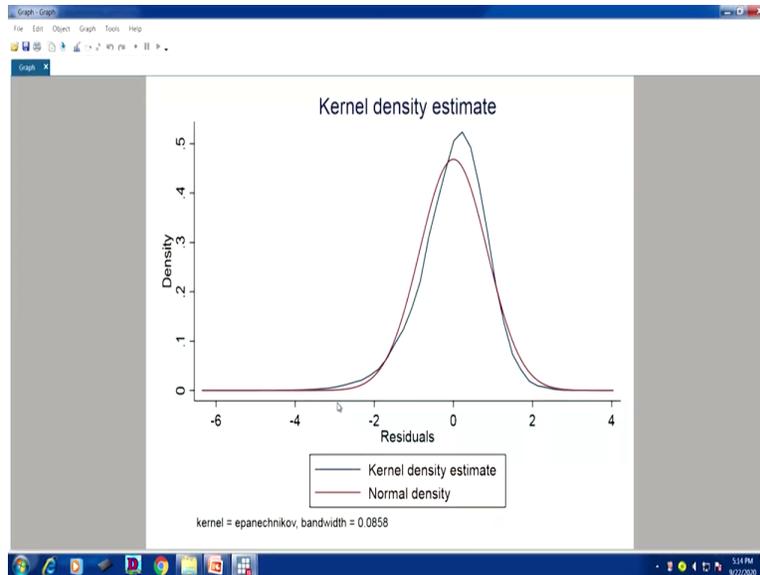
Variables 49

Observations 3839

Size 580K

Memory 64M

Sorted by



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)

- ❑ A kernel density plot produces a kind of smooth diagram for the residuals, the option normal overlays a normal distribution to compare. Here residuals seem to follow a normal distribution.
- ❑ **Standardize normal probability plot (pnorm)** checks for non-normality in the middle range of residuals.
Command: `pnorm r, title("normal check")`

Stata 11.1 - O:\regression\practic01.d

File Edit Data Graphics Statistics User Window Help

Review

Command

```

1 use 'O:\regression\practic01.d'
2 desc _percapita01a age of_
3 sum _percapita01a age of_
4 scatterplot _percapita01a_199
5 graph twoway (scatter) (_per_
6 _percapita01a age of_
7 percent) (_percapita01a age_
8 histogram) (_percapita01a_
9 histogram) (_percapita01a_
10 density) (_percapita01a_
11 search) (estab)
12 regress (_percapita01a tot_
13 _estab) (_percapita01a tot_
14 _estab) (_percapita01a tot_
15 _estab) (_percapita01a tot_
16 _estab) (_percapita01a tot_
17 _estab) (_percapita01a tot_
18 _estab) (_percapita01a tot_
19 _estab) (_percapita01a tot_
20 _estab) (_percapita01a tot_
21 _estab) (_percapita01a tot_
22 _estab) (_percapita01a tot_
23 _estab) (_percapita01a tot_
24 _estab) (_percapita01a tot_

```

Variables

Name	Label
total_worker	average number of...
inc_maxact	major inc activity
activity_group	
loan_outstan...	Amount outstand...
min_GVA	
max_GVA	
normal_GVA	
age_of_enter...	
percapita01a	
percapita01c	
_est_model1	example1 from ex...
_est_model2	example2 from ex...
_est_model3	example3 from ex...

Properties

Variables

Data

Filename: reg.practic01.d

Label: %10.0g

Variables: 49

Observations: 16,809

Size: 1,804

Memory: 64M

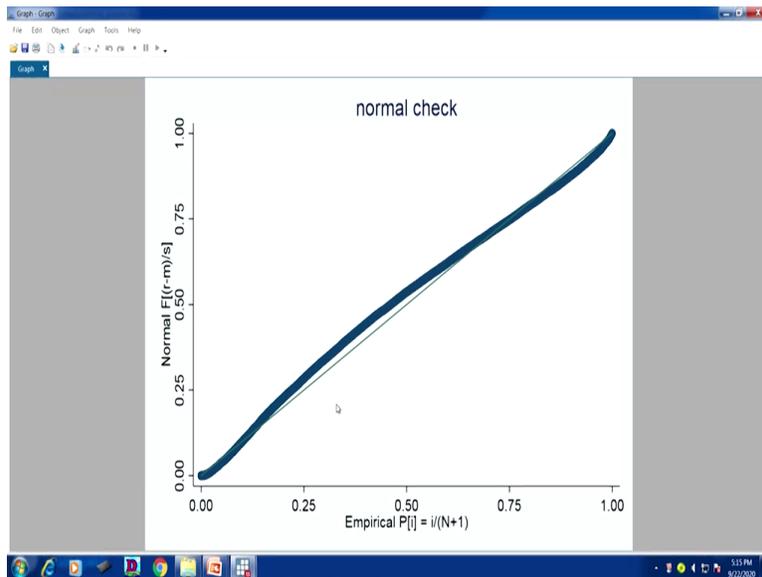
Sorted by:

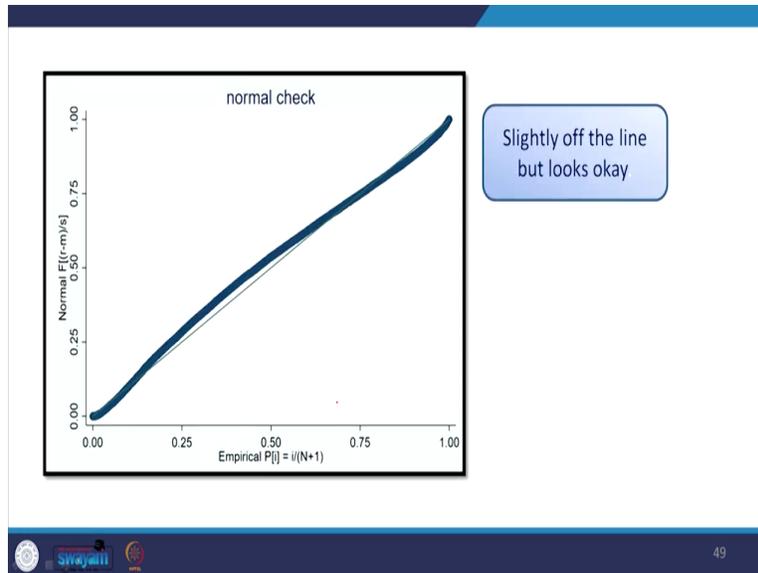
Command

```

- predict _percapita01a_hat
- graph twoway (scatter) (_percapita01a_hat) (_percapita01a) (lfit) (_percapita01a_hat) (_percapita01a), title("Model Check")
- predict r_essid
- density r_normal
- pnorm r_essid, normal
- pnorm r_essid, normal check

```





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.

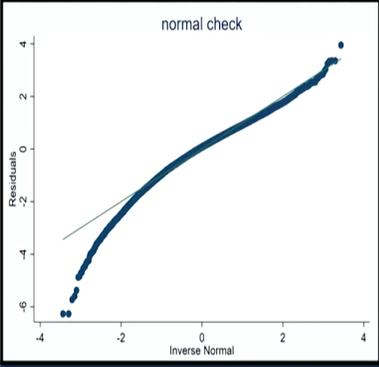
So, what is important here to note that the standardized normal probability plot that is `pnorm` checks for non-normality in the middle range of the residuals. So, for that you simply go by the command `pnorm r` and if you go by a title like this we can explain it correctly. So, it gives us the plot 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)

□ **Quintile-normal plots (qnorm)** check for non-normality in the extremes of the data (tails). It plots quintiles of residuals vs quintiles of a normal distribution. Tails are a bit off the normal.

Command:
`qnorm r, title("normal check")`

Left Tail is a bit off the normal. It means residual is symmetric with fat tails.



50

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)

Check for multicollinearity:

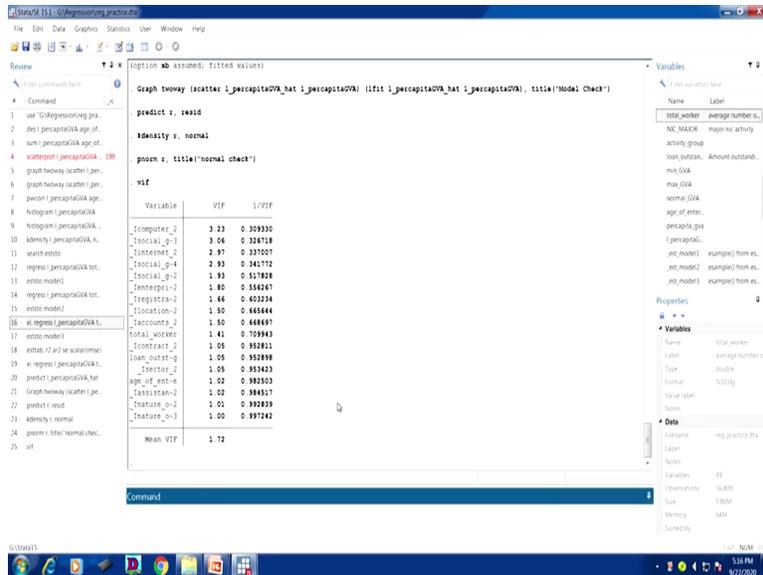
The Stata command to check for multicollinearity is `vif` (variance inflation factor). Right after running the regression type:

```
. vif
```

Variable	VIF	1/VIF
_lcomputer_2	3.23	0.309330
_lsocial_g-3	3.06	0.326718
_linternet_2	2.97	0.337007
_lsocial_g-4	2.93	0.341772
_lsocial_g-2	1.93	0.517828
_lenterpri-2	1.80	0.556267
_lregistra-2	1.66	0.603234
_llocation-2	1.50	0.665644
_laccounts_2	1.50	0.66697
total_worker	1.41	0.709943
_lcontract_2	1.05	0.952811
loan_outst-g	1.05	0.952898
_lsector_2	1.05	0.953423
age_of_ent-e	1.02	0.982503
_lassistan-2	1.02	0.984517
_lnature_o-2	1.01	0.992839
_lnature_o-3	1.00	0.997242
Mean VIF	1.72	

A VIF > 10 or a 1/VIF < 0.10 indicates trouble

51



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 σ^2 or the variance of the error term varies. But our assumption is that there should be homoscedasticity.

(Refer Slide Time: 56:53)

Testing for homoscedasticity:

- Breusch-Pagan test detects heteroscedasticity in the model. The null hypothesis is that the residuals are homoscedastic.

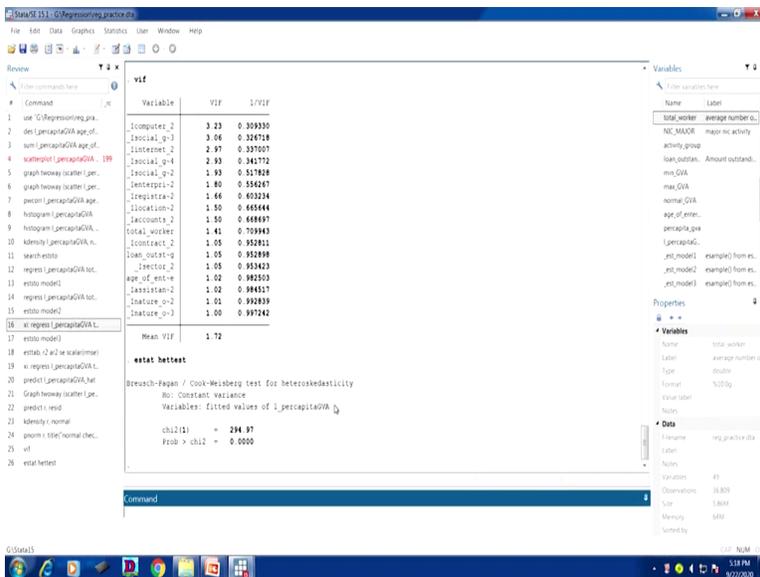
estat hettest

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of l_per capitaGVA

      chi2(1)      =    294.97
      Prob > chi2   =    0.0000
```

In our model there is a presence of heteroscedasticity, as we reject the null hypothesis of homoscedasticity. Although we have suggested in previous slides to use robust option to control for heteroscedasticity



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.