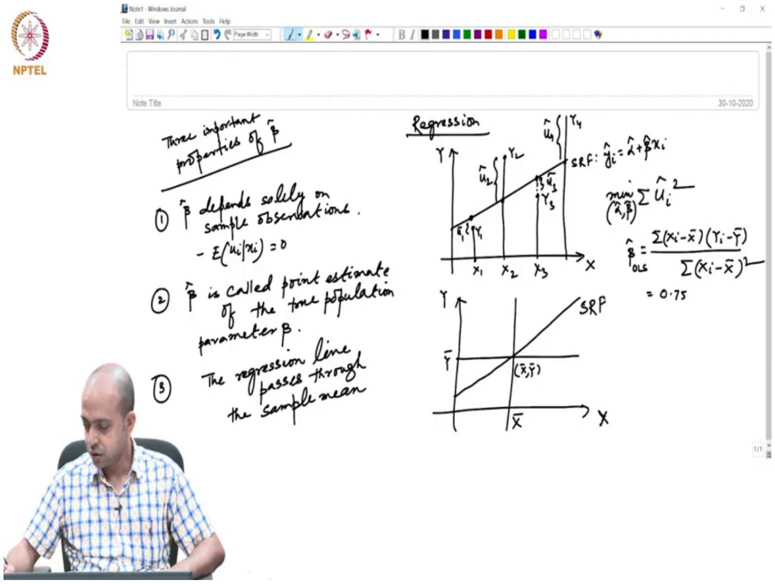
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 11
Classical Linear Regression Model Part-5

Welcome once again. We were discussing about regression technique yesterday and we were trying to understand using a simple diagram (())(0:27). We have also discussed the OLS (ordinary least square) technique to estimate the population parameter alpha and beta. Today we will take one data set and then try to estimate the model using a statistical software called Stata. But before we estimate the model we will quickly recap what we were discussing yesterday.

(Refer Slide Time: 01:00)



Three important properties of $\hat{\beta}$

① $\hat{\beta}$ depends solely on sample observations.
- $E(u_i|x_i) = 0$

② $\hat{\beta}$ is called point estimate of the true population parameter β.

③ The regression line passes through the sample mean

Regression

SRF: $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$

$\min_{(\hat{\alpha},\hat{\beta})} \sum \hat{u}_i^2$

$\hat{\beta}_{OLS} = \frac{\sum (x_i-\bar{x})(Y_i-\bar{Y})}{\sum (x_i-\bar{x})^2}$

$= 0.75$

So, this is basically the regression we are discussing. This is our x in this axis here we are measuring y. This is our regression line SRF (sample regression function) which is basically $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$. For any given value of x let us say, this is x_1 then your observed y_i is let us say y_1 but your model says this is your predicted value so that means this gives you the difference between this y_i and $\hat{y}_i$ which is your error term $\hat{u}_i$. Similarly, for x_2 let us say your observed y is here but your predicted value is here on the line. So, that means this gives you $\hat{u}_2$ which is y_2 - $\hat{y}_2$. So, then for x_3 let us say again this is your observed y_3 but your predicted value is here.

So, that means this gives you $\hat{u}_3$ and let us say for again x_4 this is your observed y which is y_4 but your predicted value is here.

So, that means this gives you $\hat{u}_4$ so when your actual y is basically greater than your predicted value then y_i - $\hat{y}_i$ is positive. That means the predicted value of the error term is positive but when your observed value is lower than what you have predicted, then your error term predicted data value is negative. Here we are trying to minimize the ui hat square with respect to alpha hat and beta hat and then if you solve this we get beta hat which is basically summation xi minus x bar times yi minus y bar divided by xi minus xbar whole square. This is your beta hat OLS because we have estimated this by minimizing the sum of predicted error square. That is why it is called ordinary least square technique. Now, there are 3 important properties about this beta hat.

First of all from the formula beta hat equals to xi minus xbar times yi minus y bar divided by summation of xi minus xbar the whole square. We can understand that beta hat depends solely on sample observation. That means if you know x and y and if you have data on x and y, then without knowing anything you can actually estimate your β because this is the $\hat{\beta}$ formula. Only one assumption what we are making here is that expectation of u_i given x_i is equal to 0. This is the assumption we need for the estimation purpose.

So, once again I repeat that econometric analysis involve two stages. First stage is estimation and second stage is inference making. So, as far as estimation of $\hat{\beta}$ is concerned from the formula what we have derived, we can easily understand that it does not require any other assumption or anything else apart from knowing the x_i and y_i. So the only assumption about the error term that is required is expectation of u_i given x_i should be equal to 0 and we are minimizing $\Sigma\hat{u}_i^2$ sum of because sum of ui is 0 since this line is the average line and we say that expectation of ui or mean value of ui is actually 0. So, this is the assumption that is required for estimating beta hat. So, that means the assumption that ui follows a normal distribution is not required for estimation purpose. Without knowing the distributional properties of ui also we can estimate $\hat{\beta}$ distributional properties are required for the second stage of inference making.

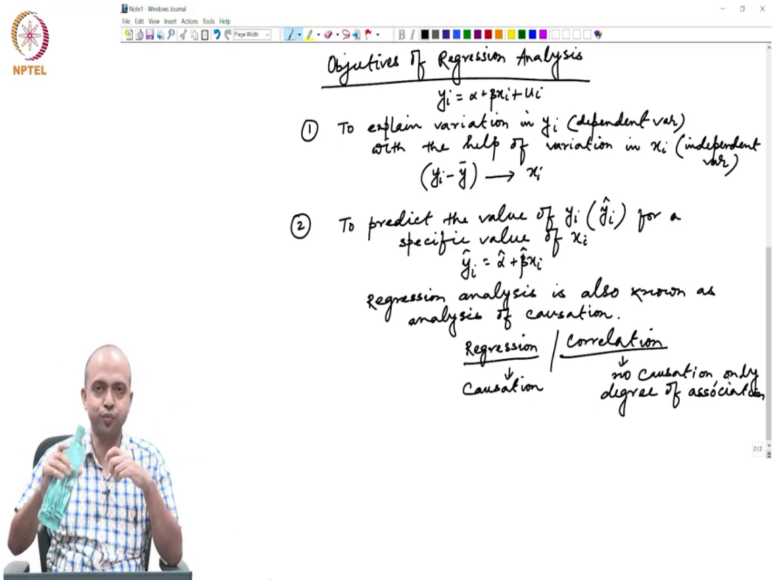
So, a $\hat{\beta}$ depends solely on the sample observation. Secondly $\hat{\beta}$ is called point estimate of the true population parameter β because if you solve this you will arrive at a specific value let us say, 0.75. So, that means if you know x_i and y_i and if you solve this using this formula then you will arrive at a specific value of $\hat{\beta}$. Let us, assume that the value is 0.75. Since we are arriving at a particular value of $\hat{\beta}$ using this technique this $\hat{\beta}$ so derived is called point estimates of the true population parameter β. Now, this point estimate is used for inference making. That means in our entire econometric analysis we are trying to infer or guess something about the true population parameter β. That means we are trying to understand the consumption behavior of the population using a sample. From the sample will get $\hat{\beta}$ and we are trying to use that $\hat{\beta}$ to guess or infer something about β.

So, when $\hat{\beta}$ gives you a specific value like 0.75 to guess something about the true population parameter it involves some amount of risk because from the given population as I told we can derive so many. In one sample I got 0.75 and there is no guarantee that in the next sample also I will get a value like 0.75 or 0.80 or 0.65 which is significantly greater than 0 but less than 1. Because, this gives only one specific value and that is the risk of using point estimate for inference making even though it is easy to calculate. Instead of using this point estimate we will discuss and estimate other types of estimates of beta which is called interval estimation.

So, instead of giving a specific value we will say that this is the range in which the beta hat may lie so that is called interval estimation. We are not discussing interval estimation right now. We will discuss interval estimation when we will discuss hypothesis testing but for the time being you have to keep in mind that this is point estimate and the risk of using point estimate for inference making is that there is no guarantee that in all other samples derived from the same population your point estimate of β will arrive like a value which is significantly different from 0 but less than 1. This is property number 2. This is called point estimate.

Property number 3 says that the regression line passes through the sample mean. That means if this is your x and this is y, let us say this is $\bar{x}$ and $\bar{y}$ point. So, your estimated regression line must pass through the sample mean $\bar{x}$ and $\bar{y}$. This is also an important property of the sample regression function that you must keep in mind and this property would be useful later on. Now will go back to the regression and we will learn regression. We will use a data set and we will regress y_i on x_i to get your $\hat{\beta}$.

(Refer Slide Time: 13:44)



Objectives of Regression Analysis

$y_i = \alpha + \beta x_i + u_i$

① To explain variation in y_i (dependent var) with the help of variation in x_i (independent var)

$(y_i - \bar{y}) \longrightarrow x_i$

② To predict the value of y_i ($\hat{y}_i$) for a specific value of x_i

$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Regression analysis is also known as analysis of causation.

Regression | Correlation

↓ Causation

↓ no causation only degree of association

But before that I will remind you about the objectives of a regression analysis. So, when you are estimating this type of model where $y_i = \alpha + \beta x_i + u_i$, the first objective of the regression analysis as I have already mentioned is that regression analysis is also known as dependence analysis that means we are trying to analyze how this y_i is dependent on x_i. So, that means alternatively we can say that the objective of regression analysis is to explain the variation in y with the help of variation in x. Thus the first objective is to explain variation in y_i which is the dependent variable with the help of variation in x_i which is the independent variable.

Now, when I am saying variation that means this is basically variance I am talking about and variance is always defined as a deviation from the mean. That means you can say that objective of the regression analysis is to explain the variation in y around its mean value that is variation in y_i around its mean value. That means y_i - $\bar{y}$ is the variation or variance which we are trying to explain by the variation in x. That means with the help of the explanatory variable that is the first objective. Second objective is to predict the value of y_i which is called $\hat{y}_i$ for a specific value of x_i. That means you will estimate this type of model where $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$. So, if you know your

$\hat{\alpha}$ and $\hat{\beta}$, after estimation then you can predict y_i when you put a specific value of x. So, these are the two objectives of regression analysis and I would also like to remind you that regression analysis is also known as analysis of causation. So, that means I would like to estimate some causal relationship between y_i and x_i. So, if your beta hat turns out to be significant that means different from 0, I can say that x actually causes y. That means income can explain the variation in consumption and that is why it is also known as a causal relationship.

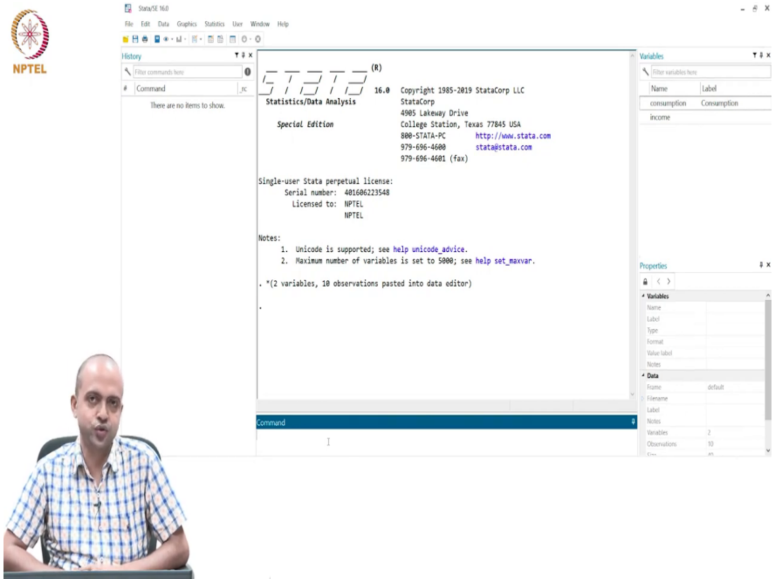
But, econometric analysis per se will not tell you the direction of causation. That means here you have specified y_i as a dependent variable and x_i as an independent variable. That is why when it is significant I am saying that the causation runs from x to y but it is you who is specifying y as the dependent variable based on your own knowledge or existing theory.

So, that means this analysis is not telling anything about the causation and what should be your dependent and independent variable. You have to keep in mind that it is you, it is the researcher who is specifying the dependent variable and independent variable and after estimation by looking at the significance of this coefficient $\hat{\beta}$ you are telling that the causal relationship is established. You only hypothesize the causal relationship and it is the econometric analysis that tells whether the causal relationship what we have hypothesized earlier is established or not. So, this is known as analysis of causation.

One related concept if you might have heard is called correlation. There are two related concepts. One is called causation or regression and then the other is called correlation. So, regression analysis is causal relationship or causation regression which is is also known as causation. But in correlation there is no causation. There is only degree of association. For example, if I look at the smoking behavior of individuals and cancer patients then correlation will tell you whether smoking behavior has some kind of association with cancer but I do not know whether smoking is actually leading to cancer or not. Another example let us say, I have collected a student's score in mathematics and statistics and then the correlation analysis of mathematics score and statistics score can tell you whether there is some kind of degree of association between the math score and stat score but I do not know whether math score is leading to stat score or stat score is leading to math score.

So that means while regression is called a causal data analysis, correlation can only tell you the degree of association and nothing more than that. This is the difference between regression and correlation. Now I introduce the statistical software and I will now show you how to estimate the model for which you might be eagerly waiting.

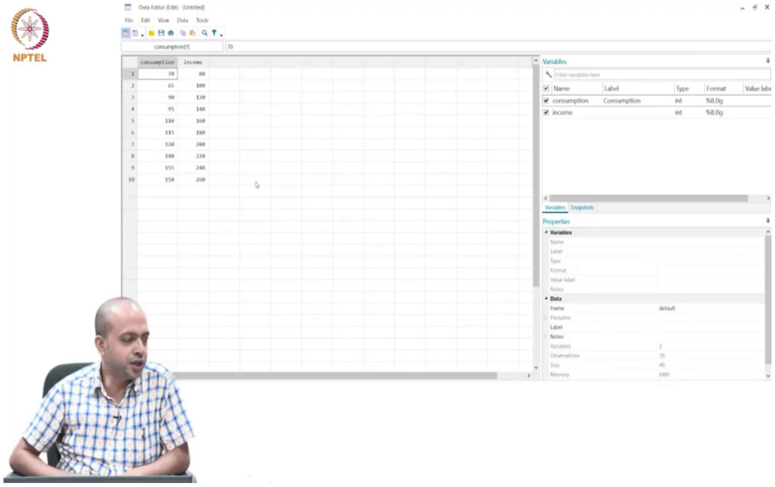
(Refer Slide Time: 23:09)



```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   16.0   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  401606223548
         Licensed to:  NPTEL
                       NPTEL

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

. *(2 variables, 10 observations pasted into data editor)

.
```

Look this particular software known as Stata. This is the software that I am going to use throughout my course. So you have to get an access to this particular software. Of course the estimation is possible in other software's also like R or Mat-lab or SPSS or any other software but the only thing is that I am going to demonstrate throughout this course of basic econometrics using this particular software. So, it would be really helpful if you can manage this software otherwise you have to estimate using the other software's for which I am not going to give you the demonstration. So, the alternative software's for this estimation purpose use SPSS or R Mat-lab and eviews. Mat-lab is of course not a statistical software but with that also you can do it. So, since I am going to use this software it would be helpful for you if you can manage to get an access to this data.

(Refer Slide Time: 24:48)

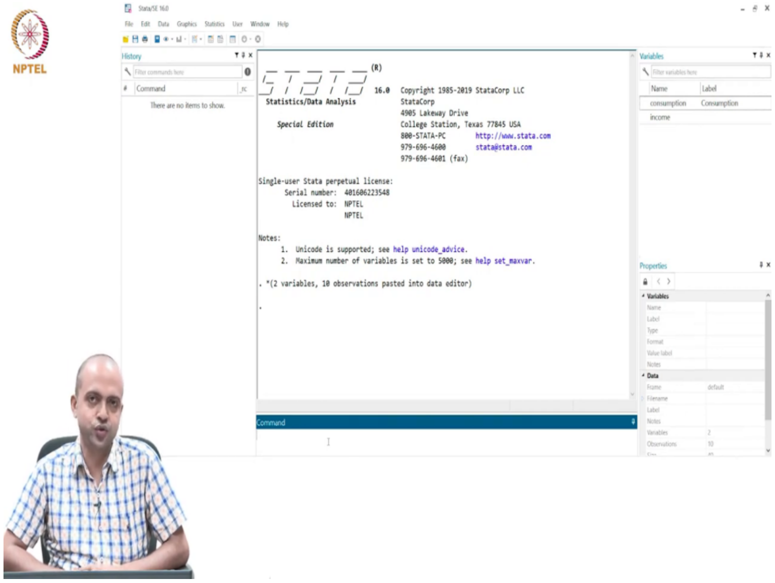


	consumption	income
1	70	80
2	65	100
3	90	120
4	95	140
5	110	160
6	115	180
7	120	200
8	140	220
9	155	240
10	150	260

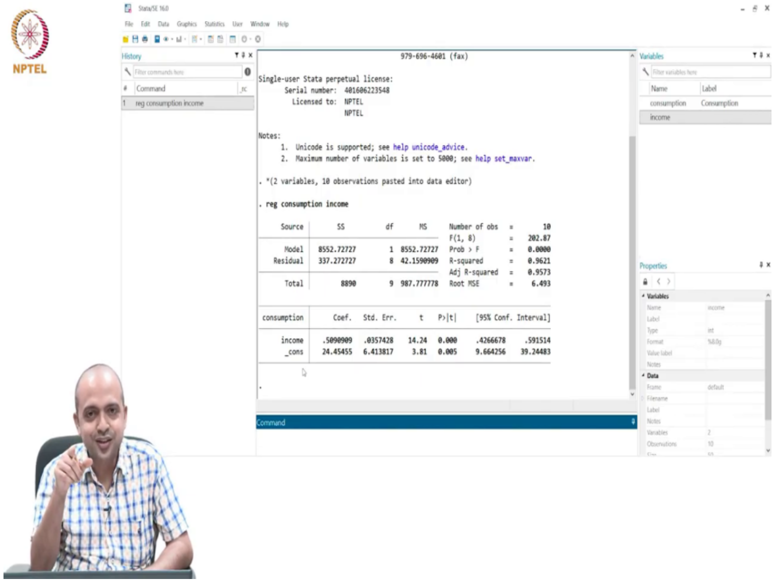
☑	Name	Label	Type	Format	Value labe
☑	consumption	Consumption	int	%8.0g	
☑	income		int	%8.0g	

Now I will show you my data source that is the data that we are going to estimate. This is a hypothetical data sample on consumption and income for some 10 families. I have a simple data on income and consumption. If you have data on 100 families or 1000 families the technique used is same. So if you learn the technique of estimation using this small software you can use it for a large data set also. There is absolutely no problem for that and I am going to share this particular data set also with you so that at home if you have access to Stata then you can actually estimate the statistical model.

(Refer Slide Time: 25:29)



```
                                     (R)
  ___  ____  ____  ____  ____
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   16.0   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  401606223548
         Licensed to:  NPTEL
                       NPTEL

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

. *(2 variables, 10 observations pasted into data editor)

.
```



```
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  401606223548
         Licensed to:  NPTEL
                       NPTEL

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

. *(2 variables, 10 observations pasted into data editor)

. reg consumption income

      Source |       SS           df       MS      Number of obs   =        10
-------------+----------------------------------   F(1, 8)         =    202.87
       Model |  8552.72727         1  8552.72727   Prob > F        =    0.0000
    Residual |  337.272727         8  42.1590909   R-squared       =    0.9621
-------------+----------------------------------   Adj R-squared   =    0.9573
       Total |        8890         9  987.777778   Root MSE        =     6.493

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .5090909   .0357428    14.24   0.000     .4266678     .591514
       _cons |   24.45455   6.413817     3.81   0.005     9.664256    39.24483
------------------------------------------------------------------------------

.
```

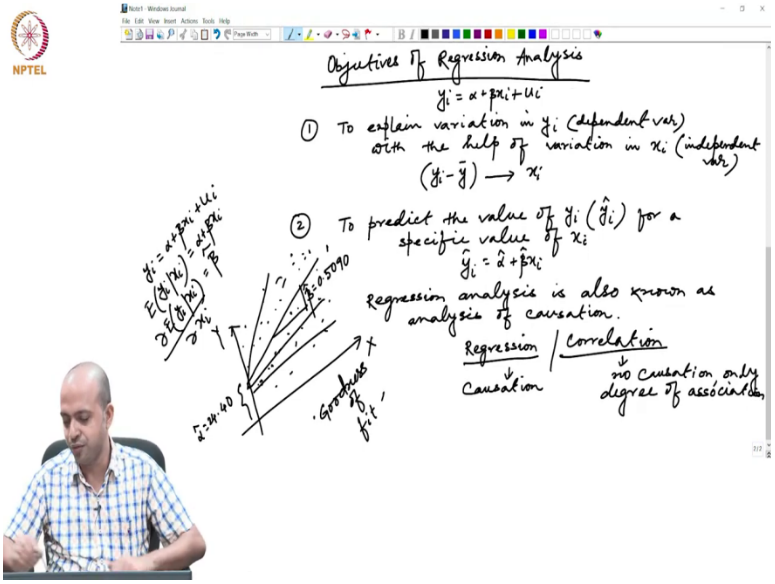
So there is this window called command window. Whatever you would like to estimate, you have to specify a command for that and all these commands are easily available online. That is why I said earlier that estimating a model is not at all difficult. It is very simple with the advent of statistical software but what is more important is to know the theoretical steps of econometrics so that after estimation you can actually interpret the coefficients. That means for interpreting the coefficients you need to know the theory and for command and other things you can simply type in google and it will give all sorts of commands that is required to perform any type of statistical analysis using Stata.

So, it would be good if you can remember or you can write down this simple command for estimation. So, the command that I am going to write is ‘reg’ for regression. Then you have to give one space and then you have to specify your dependent variable. In this example the dependent variable is consumption. So, if you double click on consumption it will automatically be added. You do not have to type here rather you should avoid typing because while typing if you make even small mistakes then Stata will not take that command because the variable is already defined in this way. So, this is how I have given the command for estimating the model. So I have again given ‘reg’ command then after that my dependent variable and then my independent variable and now if you put enter then Stata will immediately estimate the model and give you the result. Now, I have got so much of the result but I do not know how to interpret.

Now, look at step by step. This table they have mentioned consumption which is basically your dependent variable then they have mentioned coefficient. I am not going about the standard error and other things for the time being because those things I have not discussed and you would not be able to understand also. So, after coefficient they have given your independent variable which is income and the coefficient of income is 0.50. So, that means the model that we have specified is $y_i = \alpha + \beta x_i + u_i$. The estimate of β that means $\hat{\beta}$ is 0.509 and $\hat{\alpha}$ is actually 24.45.

So we know that Stata is immediately giving you the estimate of α and β. The constant term is 24.45 and the β coefficient is 0.5090. Now, the question is how do you interpret this coefficient. The interpretation of this $\hat{\beta}$ is very simple. It says that for a unit change in income or when income changes by 1 unit your consumption changes by 0.5090 units on an average. Again I am telling you this ‘on an average’ is very important because this interpretation is valid only for the average individual. That means if the individual is having income with sample mean then for that individual when income changes by one unit from the sample average then your β that is your consumption changes by 0.5090 units and that is also from the average because first of all you need to take expectation of y given x_i otherwise you cannot remove your u_i.

(Refer Slide Time: 30:14)



Objutives of Regression Analysis

$y_i = \alpha + \beta x_i + u_i$

① To explain variation in y_i (dependent var) with the help of variation in x_i (independent var)

$(y_i - \bar{y}) \longrightarrow x_i$

② To predict the value of y_i $(\hat{y}_i)$ for a specific value of x_i

$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Regression analysis is also known as analysis of causation.

Regression | Correlation

Regression $\downarrow$ Causation

Correlation $\downarrow$ no causation only degree of association

$y_i = \alpha + \beta x_i + u_i$
$E(y_i|x_i) = \alpha + \beta x_i$
$\frac{\partial E(y_i|x_i)}{\partial x_i} = \beta$
Y
X
$\hat{\beta} = 0.5090$
$\hat{\alpha} = 24.40$
'Goodness of fit'
NPTEL
Note1 - Windows Journal
File Edit View Insert Actions Tools Help

So, I will just show you what actually you are getting. When your $y_i = \alpha + \beta x_i + u_i$ first thing you do is to take expectation of y_i given x_i and that is equal to $\alpha + \beta x_i$ because expectation of ui is equal to 0. So, that means from this point 2445 and 0.50 I can easily draw the diagram also this is x and this is y and your regression line. $\hat{\alpha}$ is equal to 24.40 and this is actually your $\hat{\beta}$ which is 0.5090.

So, $\hat{\beta}$ actually gives you slope of your sample regression function and the other name of $\hat{\beta}$ as I told you earlier is MPC or marginal propensity to consume. When your income changes by 1 unit the change in your consumption on an average is called marginal propensity to consume and that is called $\hat{\beta}$ which is 0.5090.

So, from the output what Stata has given only this much we can understand but there are so many other things that are given by Stata. Let us now try to understand those things. Basically you have a raw data on your y and x so that means this is let us say scatter plot and you have fitted a line to represent your data. But I do not know how good is this line to represent your data. So, that means I need to know something about the goodness of fit for the model what you have estimated. If you plot your raw data of x and y in excel or any software and ask the software to give you a scatter plot this is the scatter plot which is the relationship between x and y.

Now this line basically is the regression line which is going to represent that scatter plot. So once you fit a line, I can fit any line instead of this. So, once you fit a line to represent the raw data how good is this line?

So to know that I need to know the goodness of fit. Let us now try to understand the goodness of fit using a simple diagram. We will first understand that and then we will go back to the Stata result and interpret.