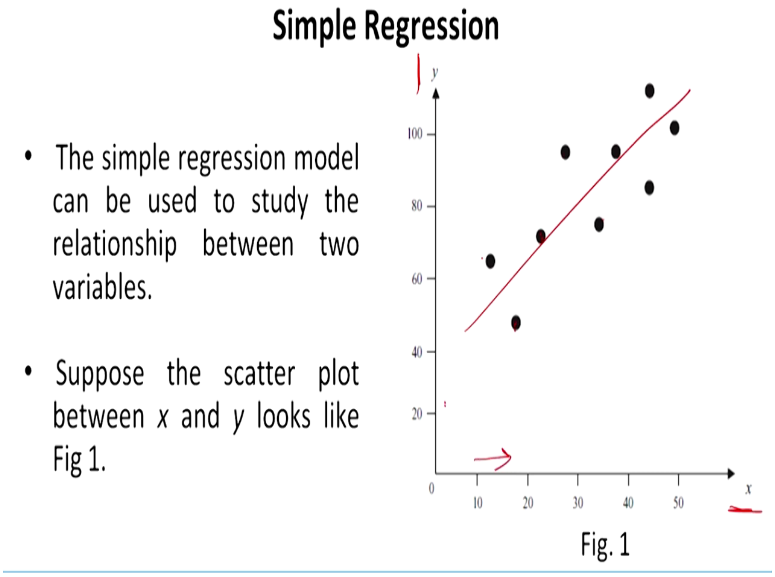
Econometric Modelling
Professor Sujata Kar
Department of Management Studies
Indian Institute of Technology - Roorkee
Lecture 06
Simple Regression – 1

Hello and welcome to module 6 of econometric modelling. This is the first module in part -2 where we present the overview of the classical linear regression module. So, there are 2 modules assigned to the concept of simple regression, modules 6 and 7. So, this is the first module on simple regression. Regression as has already been mentioned that this is the main technique in econometrics and we actually use different modifications and different types and varieties of regression methods only besides certain other methodologies of course.

But this prepares the basic or the background of the majority of the econometric methods and that is why we begin with simple regression. Simple regressions specifically refer to the situation where we have only 2 variables to deal with, one is a dependent variable, one is an independent variable.

Now regression technique, let me tell you, at the beginning that regression technique the name has come from the fact that we actually regress. So we already have observed certain observations or observed certain events, collected data on them so when we already have data on the observed events or facts then we try to find out any relationship between 1 or 2 or more variables. Since we are regressing, going back, in order to explore the relationship between 2 or more variables, it is called a regression technique.

(Refer Time Slide 2:23)



Simple Regression

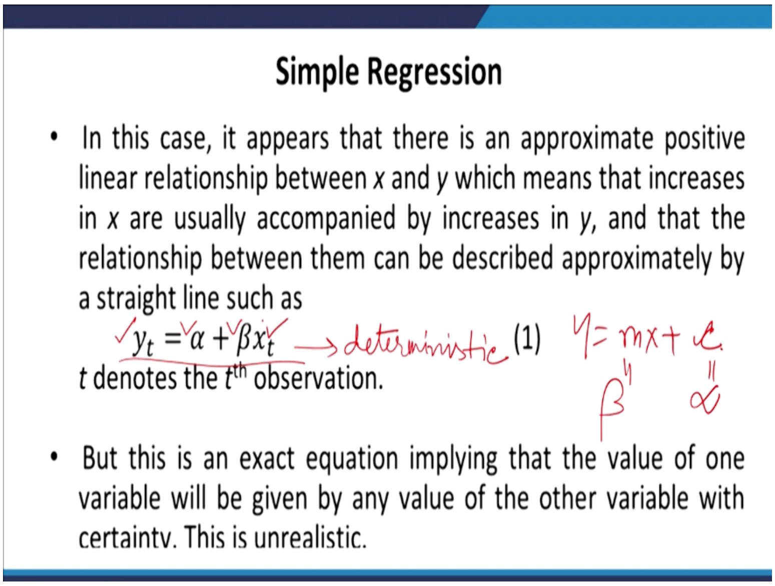
- The simple regression model can be used to study the relationship between two variables.
- Suppose the scatter plot between *x* and *y* looks like Fig 1.

Fig. 1

Now, how it is done between 2 variables is the primary concern of this module and the next one. So, the simple regression model can be used to study the relationship between only 2 variables. Suppose the scatter plot between 2 variables like x and y, looks like this. So, this is the scatter plot where we have values of x plotted against values of y where we are measuring x on the horizontal axis and y on the vertical axis.

Now, these are the scatter plots and what we try to do in regression is, try to explore a linear relationship between x and y. So, if we try to explore a linear relationship between x and y, then this diagram at least shows that there could be possibly a positive relationship between x and y which implies that when x increases y also increases or vice versa, so there is a positive relationship.

(Refer Slide Time: 3:20)



Simple Regression

- In this case, it appears that there is an approximate positive linear relationship between *x* and *y* which means that increases in *x* are usually accompanied by increases in *y*, and that the relationship between them can be described approximately by a straight line such as

  $\checkmark \underline{y_t = {}^{\checkmark}\alpha + {}^{\checkmark}\beta x_t^{\checkmark}} \longrightarrow \text{deterministic}$ (1) $\quad y = \underset{\beta}{\underset{\parallel}{m}}x + \underset{\alpha}{\underset{\parallel}{c}}$

  t denotes the t^{th} observation.

- But this is an exact equation implying that the value of one variable will be given by any value of the other variable with certainty. This is unrealistic.

So, in this case, it appears that there is an approximate positive linear relationship between x and y which means that, increases in x are usually accompanied by increases in y and that the relationship between them can be described approximately by a straight line such as

$$y_t = \alpha + \beta x_t \qquad (1)$$

This is actually the same as any equation for a straight line so straight-line equations are $y = mx + c$, so here c is equivalent to α and m is equivalent to β.

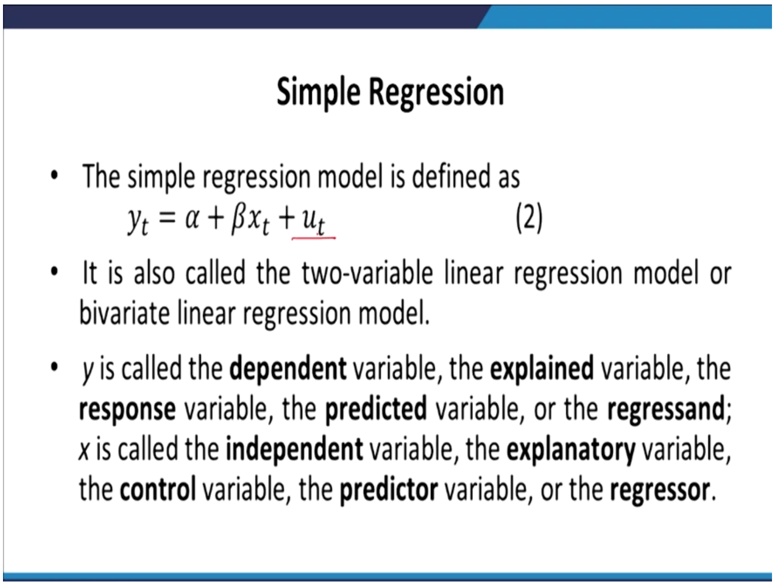
So, I am simply coming up with a straight-line relationship between x and y. Now here t denotes t'th observation. We can also replace t with i, so if I write $y_i = \alpha + \beta x_i$ then it would be the same thing, would refer to a cross-sectional data and t refers to a time series data, but this is an exact equation implying that the value of one variable will be given by any value of the other variable with certainty and this is something unrealistic.

Unrealistic, because in the beginning if you remember, we stated that we work with random variables and random variables have certain randomness or uncertainties associated with them. So, in this case, if this is a relationship then this can be called a deterministic relationship. Deterministic relationship means it does not have any randomness, any uncertainty, we can determine the value of y_t for given values of α, β , and x_t with 100 percent accuracy or certainty.

So, this is something unrealistic when we go for a prediction of economic variables or even if you do not go for prediction of economic variables, if we go for modelling of economic variables, we will find that it is not possible to come up with predictions or estimations of economic variables with 100 percent certainty. I tried to model inflation as explained by a large number of other economic variables.

Now whatever be it, if I try to predict inflation for the upcoming periods, it is not possible or most often not possible to come up with a prediction of inflation with 100 percent accuracy because there is certain randomness associated with movements in prices or inflation figures.

(Refer Slide Time 6:03)



Simple Regression

- The simple regression model is defined as

$$y_t = \alpha + \beta x_t + \underline{u_t} \qquad (2)$$

- It is also called the two-variable linear regression model or bivariate linear regression model.
- *y* is called the **dependent** variable, the **explained** variable, the **response** variable, the **predicted** variable, or the **regressand**; *x* is called the **independent** variable, the **explanatory** variable, the **control** variable, the **predictor** variable, or the **regressor**.

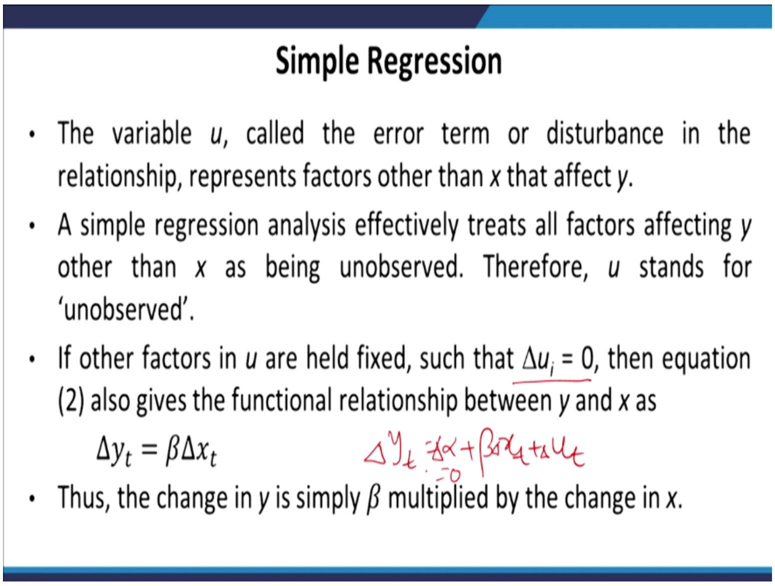
So, that is how we have this simple regression model which is defined as

$$y_t = \alpha + \beta x_t + u \qquad (2)$$

So, this randomness is introduced here, it is also called the two-variable linear regression model or bivariate linear equation model. y is called the dependent variable or the explained variable or the response variable or the predictive variable or the regressand. So, these are alternative names used for y, and for x we have these names which are independent variable, explanatory variable, controlled variable, predictive variable, or the regressor.

So, what we are doing is that we are trying to find out how x_t impacts y_t or changes in x_t impact y_t. We would come to the specific interpretation a little later but one thing to be noted here first is that y always refers to the dependent variable, the variable and x refers to the independent variable or explanatory variable that is the variable used to explain changes in y.

(Refer Slide Time: 7:19)



Simple Regression

- The variable u, called the error term or disturbance in the relationship, represents factors other than x that affect y.
- A simple regression analysis effectively treats all factors affecting y other than x as being unobserved. Therefore, u stands for 'unobserved'.
- If other factors in u are held fixed, such that $\underline{\Delta u_i = 0}$, then equation (2) also gives the functional relationship between y and x as

  $\Delta y_t = \beta \Delta x_t$ $\Delta y_t = \underset{=0}{\Delta\alpha} + \beta\Delta x_t + \Delta u_t$
- Thus, the change in y is simply β multiplied by the change in x.

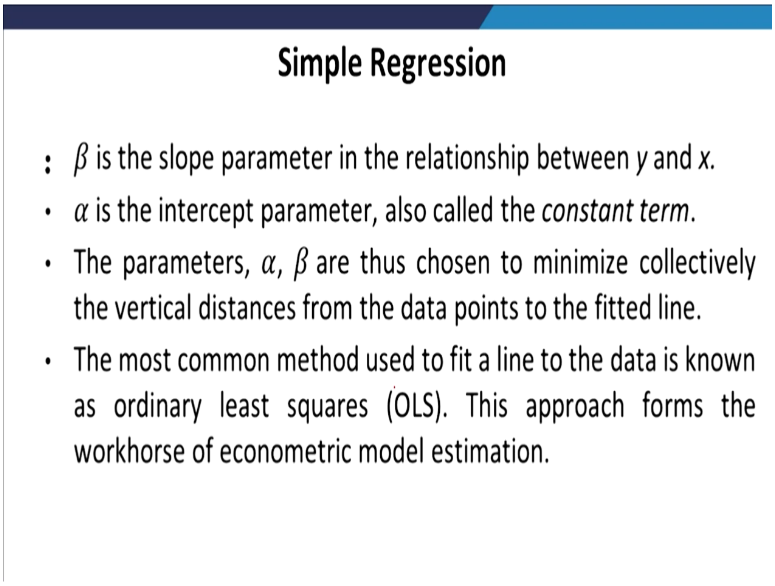
We introduce the randomness by including u, the variable u called the error term or disturbance in the relationship represents factors other than x that affect y. So, in any model, whatever big it is, whatever close you try to make it to the reality, some kind of randomness always remains there as a result of which most of it, we are not able to come up with 100 percent accuracy and all those randomnesses could be, it can be completely a random factor, it can be certain factors which are not included in the model.

And because of which all those components are being clubbed into the unobserved components denoted by u. If other factors in u are held fixed such that $\Delta u_i = 0$ then equation 2, also gives the functional relationship between y and x as $\Delta y_t = \beta\Delta x_t$, this is simply because if we measure changes then we have (refer slide time 8:39)

We are assuming delta $\Delta u_t = 0$ or $\Delta u_i = 0$ and that is why $\Delta y_t = \beta\Delta x_t$ and $\Delta\alpha$ is of course 0 because as I have mentioned that α is a constant, it cannot change thus the change in y is simply $\beta\Delta x_t$ that is actually a very crucial thing in understanding the interpretation of regression parameters and here is another thing we have mentioned and which are other things held fixed.

So, this is a common concept in economics where we say ceteris paribus so other things are held constant, we are assuming that when other things are not changing then how a change in x is going to impact y or how y is going to change in response to a change in x only and no other factors or variables.

(Refer Slide Time: 9:46)



Simple Regression

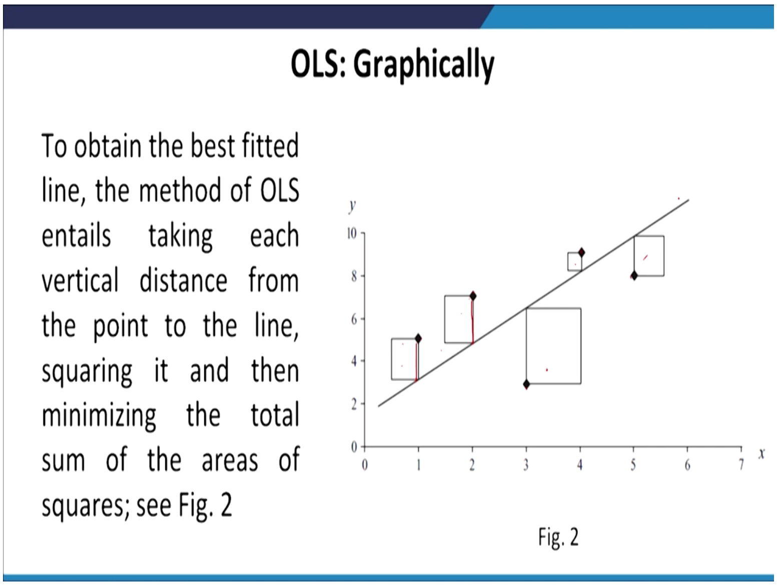
- β is the slope parameter in the relationship between y and x.
- α is the intercept parameter, also called the *constant term*.
- The parameters, α, β are thus chosen to minimize collectively the vertical distances from the data points to the fitted line.
- The most common method used to fit a line to the data is known as ordinary least squares (OLS). This approach forms the workhorse of econometric model estimation.

Having said that, in our simple regression beta is the slow parameter in the relationship between y and x which is equivalent to m in typical linear equations like $y = mx + c$, α is the intercept parameter also called constant term, the parameters α and β are thus chosen to minimize collectively the vertical distance from the data points to the fitted line.

So, when I showed you the scatter plot, I fitted the line through it. It can be any line but then we do not go for any line. That is why we need a statistical method, a specific method that gives us the best-fitted line and how do we choose that best-fitted line?

The best-fitted line is chosen by minimizing the collective vertical distance from the data point to the fitted line. The most common method used to fit a line to the data is known as ordinary least squares or OLS in short. This approach forms the workhorse of econometric model estimation.

(Refer Slide Time: 11:02)



OLS: Graphically

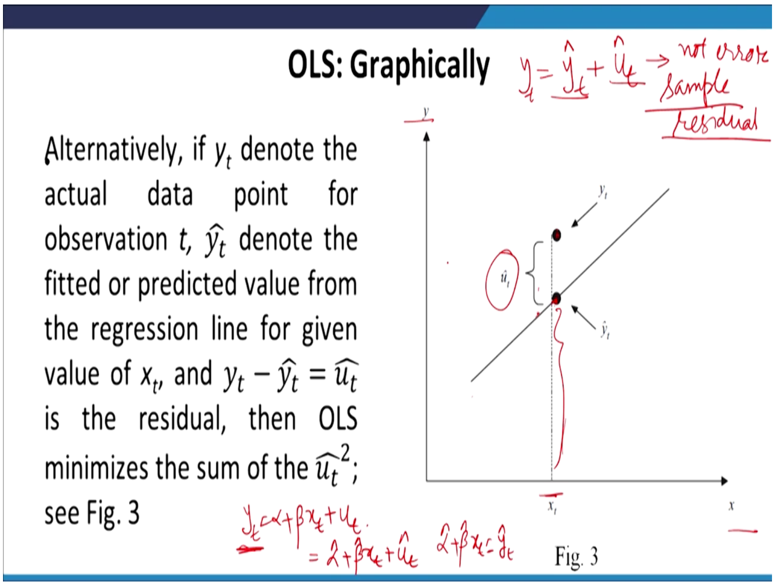
To obtain the best fitted line, the method of OLS entails taking each vertical distance from the point to the line, squaring it and then minimizing the total sum of the areas of squares; see Fig. 2

Fig. 2

So, from now onwards you would use the term OLS, so first of all let me explain OLS graphically then we will do it mathematically. To obtain the best-fitted line the method of OLS entails taking each vertical distance from the point to the line squaring it and then minimizing the total sum of the areas of the squares.

So, this is figure 2, you can see these individual points. Suppose there are only 5 points so in this scatter plot, there are only 5 points. For the sake of exposition, we consider only 5 points, and then the vertical distance is considered, it is squared and these squares are then summed up. The line which gives us the minimum of this sum is the best-fitted line as given by the method of ordinary least squares or OLS.

(Refer Slide Time: 12:03)



OLS: Graphically $y_t = \hat{y}_t + \hat{u}_t \rightarrow$ not error Sample residual

Alternatively, if y_t denote the actual data point for observation t, $\hat{y}_t$ denote the fitted or predicted value from the regression line for given value of x_t, and $y_t - \hat{y}_t = \hat{u}_t$ is the residual, then OLS minimizes the sum of the $\hat{u}_t^2$; see Fig. 3

$y_t = \alpha + \beta x_t + u_t.$

$= \hat{\alpha} + \hat{\beta} x_t + \hat{u}_t \quad \hat{\alpha} + \hat{\beta} x_t = \hat{y}_t$

Fig. 3

Alternatively, if y_t denotes the actual data point for observation t and $\hat{y}_t$ denotes the fitted or predicted value from the regression line for a given value of x_t and $y_t - \hat{y}_t = \hat{u}_t$ is the residual then OLS minimizes the sum of $\hat{u}_t^2$. So, here we show it graphically, again we are measuring x on the horizontal axis and y on the vertical axis and this is my fitted line. Now I am not having all those 5 scattered points, we only have considered one point, this is my fitted line and this is my actual observation and this distance is given by $\hat{u}_t$, this is $\hat{y}_t$.

Now here I need to tell you what I have mentioned (refer slide time 13:00). Now when we actually estimate it, then we will be having something like (refer slide time 13:15), essentially what we are trying to do is that this random variable is divided into 2 components, one is the component which is estimated or estimable, the other component is not estimable.

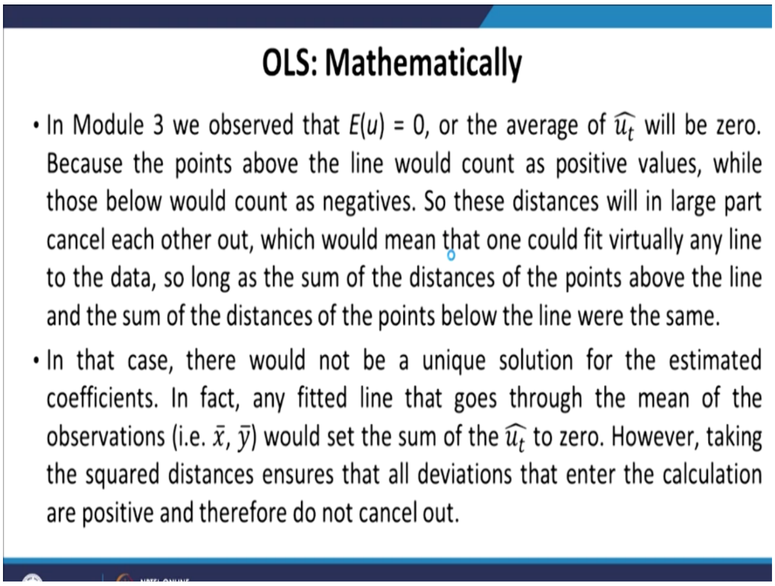
So, if you remember previously, we had discussed the 2 components fixed component and random component for a variable x the fixed component was its mean, and the variable component or the random component was the unobserved term or the error term. Now here in this case we have y where $\hat{y}_t$ is the estimable or observable component and what remains unobserved?

Whatever remains unobserved is denoted by $\hat{u}_t$. Now we are using $\hat{u}_t$ and not ut, ut is not used here because if you remember u actually correspondence to the population but when we go for regression analysis most often, we work with the sample, so I have picked up a sample and that sample gives us an estimate of y_t which is equivalent of $\hat{y}_t$.

Now whatever remains is very specific to that sample and we do not call it an error, we call it to sample residual and we denote it by $\hat{u}_t$. So, now note that $\hat{u}_t$ is not the error term, it is basically a sample residual, there is a difference between these two things, $\hat{u}_t$ is sample residual so it is very specific to the sample whereas u is specific to or u corresponds to the population unobserved or error term.

So, you can see that, this is my actual observation, I could estimate corresponding to this observation only up to this much so basically x explains this much variation in y_t and corresponding to this value of x, we have this estimated value of $\hat{y}_t$ and whatever remains unexplained, given this sample is called $\hat{u}_t$, fair enough? So, this is what OLS is graphically.

(Refer Slide Time: 15:47)



OLS: Mathematically

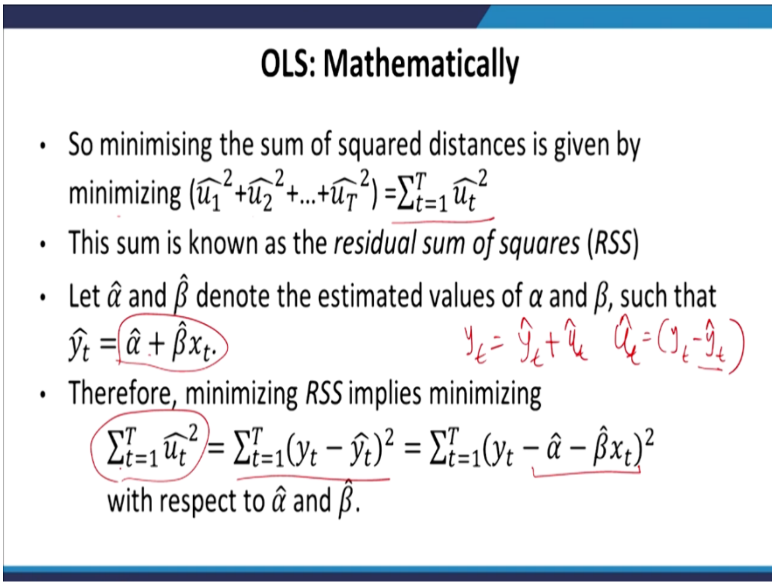
- In Module 3 we observed that $E(u) = 0$, or the average of $\widehat{u_t}$ will be zero. Because the points above the line would count as positive values, while those below would count as negatives. So these distances will in large part cancel each other out, which would mean that one could fit virtually any line to the data, so long as the sum of the distances of the points above the line and the sum of the distances of the points below the line were the same.
- In that case, there would not be a unique solution for the estimated coefficients. In fact, any fitted line that goes through the mean of the observations (i.e. $\bar{x}$, $\bar{y}$) would set the sum of the $\widehat{u_t}$ to zero. However, taking the squared distances ensures that all deviations that enter the calculation are positive and therefore do not cancel out.

So, in module 3, we observed that the expected value of u is equal to 0 so in the population the mean value or the population mean is 0 or the average of $\hat{u}_t$ will also be 0 so when we come to the sample counterpart it is also expected to be 0 because the points above the line would count as positive values while those below would count as negative values and these distances will in large part cancel each other out.

Which would mean that one could fit virtually any line to the data so long as the sum of the distances of the points above the line and some of the distances of the points below the line were the same so they cancel each other out and the summation of $\hat{u}_t$ becomes 0.

In that case, there would not be a unique solution for their estimated coefficients. In fact, any fitted line that goes through the mean of the observations that is $\bar{x}$ and $\bar{y}$ would set the sum of $\hat{u}_t$ to 0. However, taking the square distances ensures that all deviations that enter the calculation are positive and therefore do not cancel out and that leaves us the scope of finding out a fitted line which basically minimizes these squared distances.

(Refer Slide Time: 17:12)



OLS: Mathematically

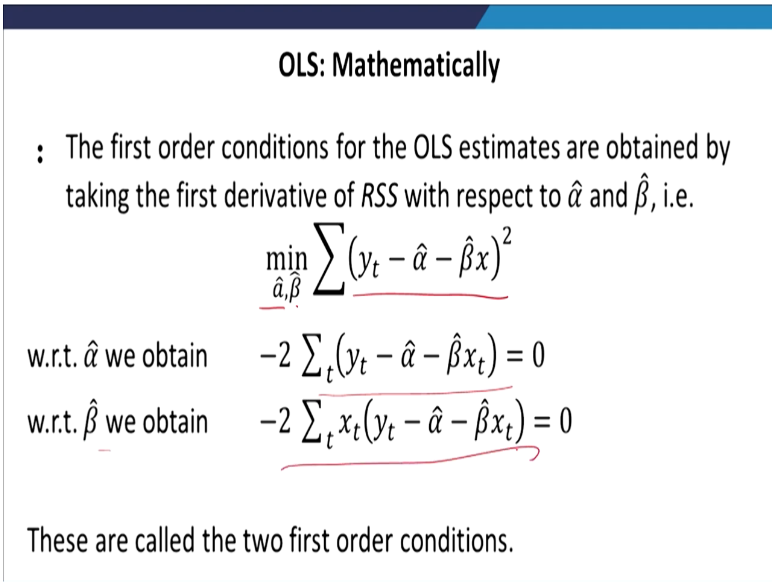
- So minimising the sum of squared distances is given by minimizing $(\widehat{u_1}^2+\widehat{u_2}^2+...+\widehat{u_T}^2) = \underline{\sum_{t=1}^{T}\widehat{u_t}^2}$
- This sum is known as the *residual sum of squares* (*RSS*)
- Let $\hat{\alpha}$ and $\hat{\beta}$ denote the estimated values of α and β, such that $\hat{y}_t = \boxed{\hat{\alpha} + \hat{\beta}x_t.}$ $y_t = \hat{y}_t + \hat{u}_t$ $\hat{u}_t = (y_t - \underline{\hat{y}_t})$
- Therefore, minimizing *RSS* implies minimizing $\boxed{\sum_{t=1}^{T}\widehat{u_t}^2} = \underline{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} = \sum_{t=1}^{T}(y_t \underline{- \hat{\alpha} - \hat{\beta}x_t})^2$ with respect to $\hat{\alpha}$ and $\hat{\beta}$.

So, minimizing the sum of square distances is given by this expression, minimizing $(\hat{u}_1^2 + \hat{u}_2^2 + \ldots + \hat{u}_T^2) = \sum_{t=1}^{T} \hat{u}_t^2$. The sum is known as the residual sum of squares or in short RSS, this is going to be also used in later modules as well. So, this is a very important concept to be remembered.

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the estimated values of α and β, such that $\hat{y}_t = \hat{\alpha} + \hat{\beta}x_t$, that I have just tried to explain, this is the estimable part of y therefore minimizing RSS implies minimizing this sum so we are minimizing the sum of squared deviations of the actual value from the fitted value and which is equal to $\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 = \sum_{t=1}^{T}(y_t - \hat{\alpha} - \hat{\beta}x_t)^2$, this is also because, if you see (refer slide time 18:24)

So, of course, (refer slide time 18:34). And we are going to minimize this with respect to $\hat{\alpha}$ and $\hat{\beta}$. So, we choose $\hat{\alpha}$ and $\hat{\beta}$in such a manner that this sum of square deviation or sum of square residuals, samples residuals is minimized.

(Refer Slide Time: 19:08)



OLS: Mathematically

- The first order conditions for the OLS estimates are obtained by taking the first derivative of *RSS* with respect to $\hat{\alpha}$ and $\hat{\beta}$, i.e.

$$\min_{\hat{\alpha},\hat{\beta}} \sum (y_t - \hat{\alpha} - \hat{\beta}x)^2$$

w.r.t. $\hat{\alpha}$ we obtain $\quad -2\sum_t (y_t - \hat{\alpha} - \hat{\beta}x_t) = 0$

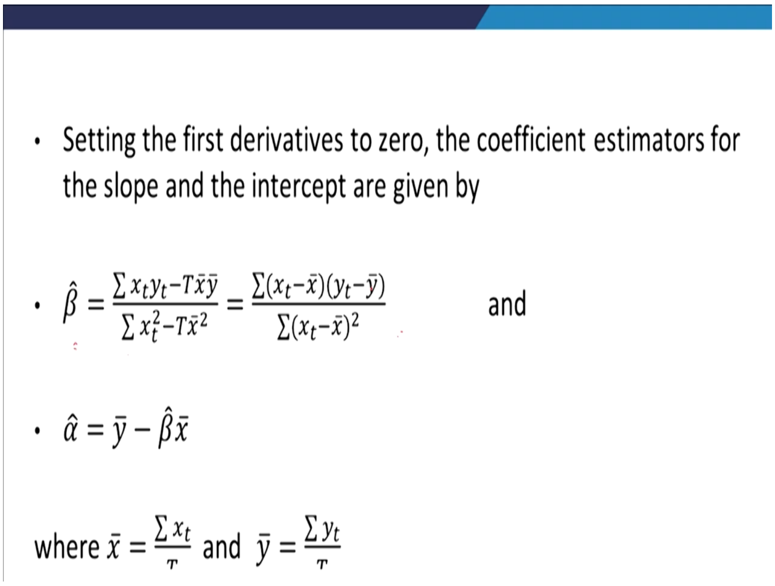
w.r.t. $\hat{\beta}$ we obtain $\quad -2\sum_t x_t(y_t - \hat{\alpha} - \hat{\beta}x_t) = 0$

These are called the two first order conditions.

And this is written as the first-order conditions for the OLS estimates are obtained by taking the first derivative of RSS with respect to $\hat{\alpha}$ and $\hat{\beta}$ so we are minimizing that expression which we just obtained that is summation ut hat square and then we have plugged in the values we are minimizing with respect to $\hat{\alpha}$ and $\hat{\beta}$.

So, with respect to $\hat{\alpha}$ that is taking the first derivative with respect to $\hat{\alpha}$, we have this expression and by taking the first derivative with respect to $\hat{\beta}$, we had this expression and these are called the two first-order conditions (refer slide time 19:46), these are simple first derivative, the first with respect to $\hat{\alpha}$ so we have 2 here minus here and the same expression and here we have additional xt here.

(Refer Slide Time: 20:05)



- Setting the first derivatives to zero, the coefficient estimators for the slope and the intercept are given by
- $\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2}$ and
- $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

where $\bar{x} = \frac{\sum x_t}{T}$ and $\bar{y} = \frac{\sum y_t}{T}$

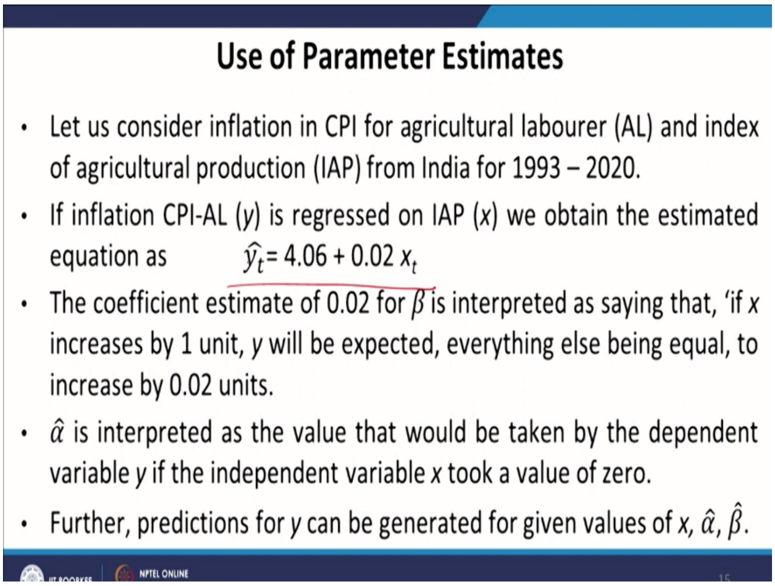
And next, we solve for these 2 expressions and what we arrive at is the expression for $\hat{\alpha}$ and $\hat{\beta}$, here I am not getting into the derivations of these expressions as such the derivations are available with any standard textbooks including the one that has been referred to here. But the derivations become easier when we do it in the context of multiple regression analysis and we use matrix analysis. So, there I will be proving the derivation that how we arrive at the value of $\hat{\alpha}$ and $\hat{\beta}$. For the time being, I do not get into the cumbersome derivation of $\hat{\alpha}$ and $\hat{\beta}$.

Setting the first derivatives to 0, the coefficient estimators for the slope and the intercept are given by $\hat{\beta}$ are the same expressions, i.e two alternative expressions of $\hat{\beta}$ where we have.

$$\hat{\beta} = \frac{\Sigma(x_t - \bar{x})(y_t - \bar{y})}{\Sigma(x_t - \bar{x})^2}. \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \text{where } \bar{x} = \frac{\Sigma x_t}{T} \quad \text{and} \quad \bar{y} = \frac{\Sigma y_t}{T}$$

.

(Refer slide Time: 21:32)



Use of Parameter Estimates

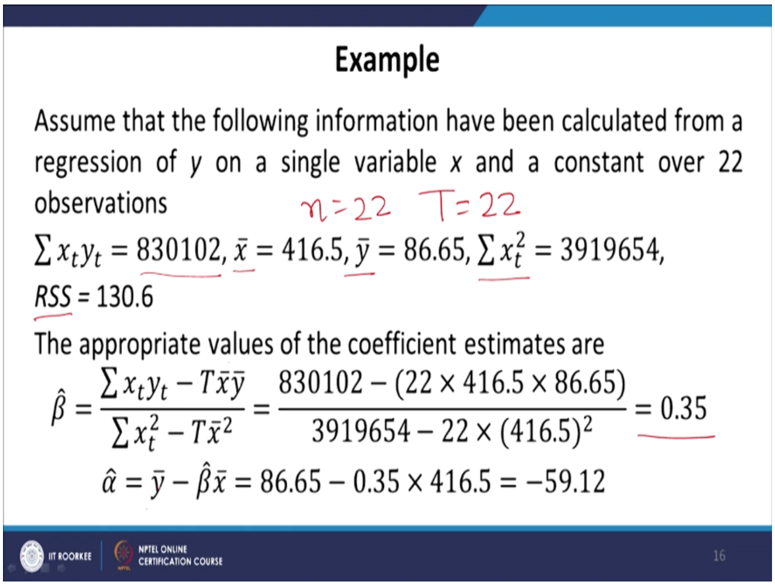
- Let us consider inflation in CPI for agricultural labourer (AL) and index of agricultural production (IAP) from India for 1993 – 2020.
- If inflation CPI-AL (y) is regressed on IAP (x) we obtain the estimated equation as $\hat{y}_t = 4.06 + 0.02\, x_t$
- The coefficient estimate of 0.02 for β is interpreted as saying that, 'if x increases by 1 unit, y will be expected, everything else being equal, to increase by 0.02 units.
- $\hat{\alpha}$ is interpreted as the value that would be taken by the dependent variable y if the independent variable x took a value of zero.
- Further, predictions for y can be generated for given values of x, $\hat{\alpha}$, $\hat{\beta}$.

Now let us consider an example, in order to explain how the parameters estimates are actually interpreted or what their uses are. So, in CPI, we consider CPI for agricultural laborers that is Consumer Price Index for agriculture laborer, one such index available in the Indian context and index of agriculture production, IAP, from India for 1993 to 2020. Now inflation is CPI-AL is our dependent variable, denoted by y, this is regressed on IAP which is our independent variable denoted by x.

So, basically, we are trying to find out the impact of agricultural production on inflation in CPI for agricultural laborers. Now following the usual OLS procedure we actually arrive at this equation where we have $\hat{y}_t$= 4.06 + 0.02 $x_{t,}$ we are writing $\hat{y}_t$ here, because we have not mentioned the error term here, so this is the estimable or estimated component of the variable yt.

The coefficient estimate of 0.02 for β is interpreted as, saying that if x increases by 1 unit, y will be expected, everything else being equal, to increase by 0.02 units. $\hat{\alpha}$ is interpreted as the value that will be taken by the dependent variable if the independent variable x took a value of 0. Further predictions for y can be generated for given values of x, $\hat{\alpha}$, and $\hat{\beta}$.

(Refer Slide Time: 23:22)



Example

Assume that the following information have been calculated from a regression of y on a single variable x and a constant over 22 observations

$n = 22 \quad T = 22$

$\sum x_t y_t = 830102, \bar{x} = 416.5, \bar{y} = 86.65, \sum x_t^2 = 3919654,$

$RSS = 130.6$

The appropriate values of the coefficient estimates are

$$\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} = \frac{830102 - (22 \times 416.5 \times 86.65)}{3919654 - 22 \times (416.5)^2} = 0.35$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 86.65 - 0.35 \times 416.5 = -59.12$$

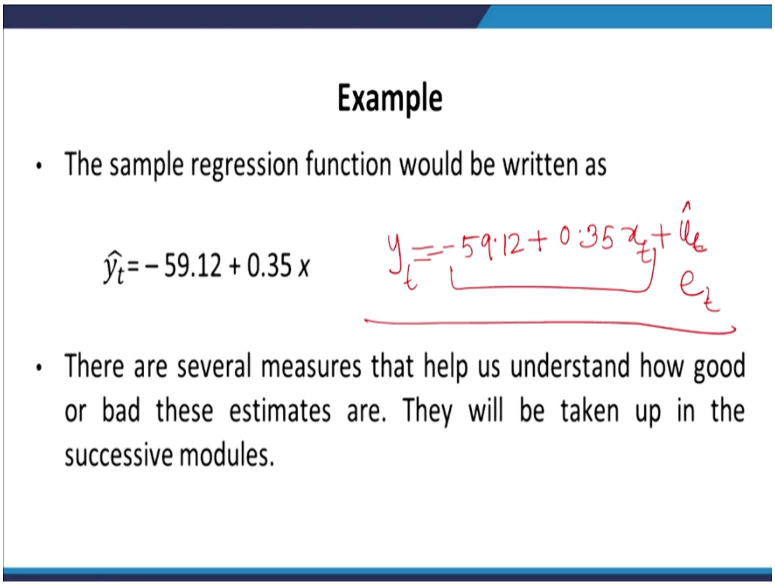
Assume that the following information has been calculated from a regression of y on a single variable x and a constant over 22 observations, so when I say over 22 observations this implies that n is equal to 22 or in case we have time-series data then T is equal to 22. So, the information that is given to us is

$$\sum x_t y_t = 830102, \bar{x} = 416.5, \bar{y} = 86.65, \sum x_t^2 = 3919654,$$

$RSS = 130.6$

The appropriate values of the coefficient estimates are, so we simply plug in these values into this expression for β and arrive at a figure which is 0.35 and similarly by plugging in the values of y bar and beta hat x bar into the expression for α, we arrive at a value of minus 59.12 (refer slide time 24:20).

(Refer Slide Time: 24:29)



Example

- The sample regression function would be written as

$$\hat{y}_t = -59.12 + 0.35\,x \qquad y_t = \underbrace{-59.12 + 0.35\,x_t} + \underset{e_t}{\hat{u}_t}$$

- There are several measures that help us understand how good or bad these estimates are. They will be taken up in the successive modules.

Once we have this information then we can construct the sample regression function which will be written as $\hat{y}_t = -59.12 + 0.35\,x$, there are several measures that help us understand how good or bad these estimates are but they will be taken up in the successive modules but for the time being we will simply try to explain how we arrive at these different estimates.

So, once I write this $\hat{y}_t = -59.12 + 0.35\,x$ which is basically the estimated component of the yt, my formal or final regression equation becomes (refer slide time 25:20). Or this is the sample, at times the sample residual is also denoted by e so we can also write (refer slide time 25:40) so this is my final regression function, where this is my estimable component or estimated component, and this is the unobserved component or the sample residual.

But as such these estimates, do they, not themselves talk about how good the model is? How much the variable x is able to explain the variations in y so the goodness of fit and further interpretations of the parameters, excreta will be taken up in successive modules. So, these are the references I have followed in order to explain the simple regression, in the next module also I will be taking up further discussion on simple regression itself. Thank you!