Handling Large-Scale Unit Level Data Using STATA
Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology Roorkee
Lecture 28
Linear Regression Analysis in STATA-I

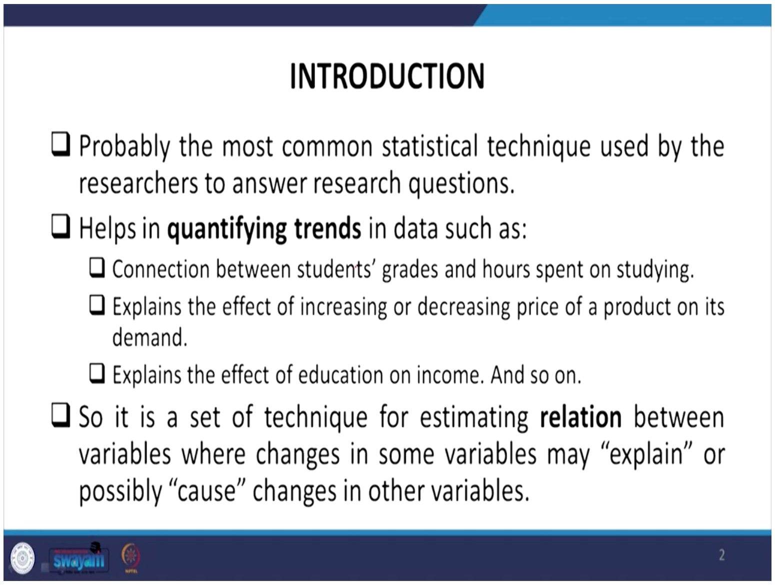
Welcome friends once again to the NPTEL MOOC module on Handling Unit Level Data Using STATA, there we mentioned handling large-scale unit level data. So, far we have discussed the data varieties, data collections and also we have equipped you with the latest know how of STATA that we have been handling in last two weeks particularly and today we will start a new chapter called Linear Regression Analysis using STATA.

In the last lecture if I just countdown few important points, last two lectures on identifying important factors. So, that is largely called factor analysis and we explained you to analyze the important factor, also their validation, their cut off points with sphericity measures, also through KMO level and its significance that we have discussed. Why am I analyzing this, recapitulating this, because those factor analysis guides us correctly to go for a better linear regression using STATA. Because in linear regression analysis some of the assumptions have to be validated and the assumptions we will analyze one by one.

As a backdrop to linear regression analysis, I just wanted to mention that three terms are very very important, one is linear, then regression, then analysis by any tools. So, why it is linear has to be validated, has to be understood correctly, hardly it requires 10-15 minutes to apply the STATA and get the coefficient or the results, but those results are going to be meaningless if you do not understand the crux of linear regression.

So, the foundation of all econometric packages, econometric tools or econometric analysis is none other than linear regression model and because of its important properties, we will be discussing the Gauss Markov properties as well, but I wanted to give you the background information of linear regression in my first two lectures.

(Refer Slide Time: 03:13)



INTRODUCTION

- □ Probably the most common statistical technique used by the researchers to answer research questions.
- □ Helps in **quantifying trends** in data such as:
  - □ Connection between students' grades and hours spent on studying.
  - □ Explains the effect of increasing or decreasing price of a product on its demand.
  - □ Explains the effect of education on income. And so on.
- □ So it is a set of technique for estimating **relation** between variables where changes in some variables may "explain" or possibly "cause" changes in other variables.

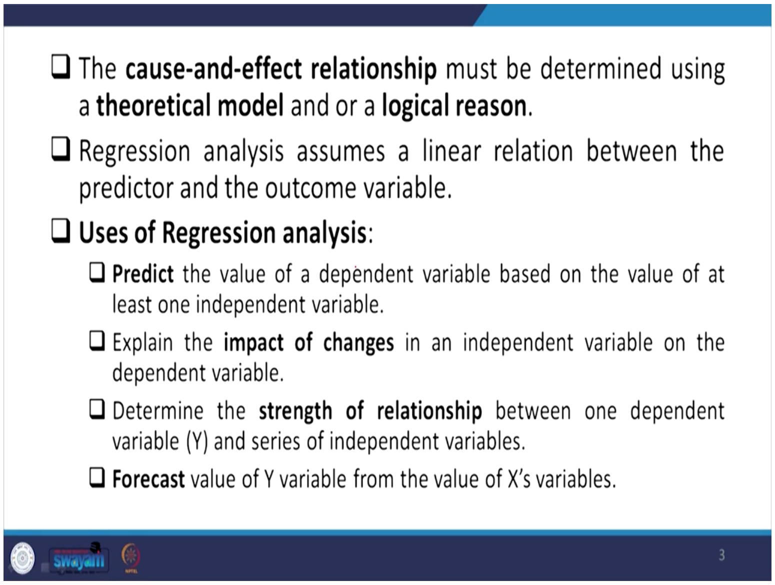
2

So, let me start by saying that the linear regression probably the most common statistical technique used by the researchers to answer research questions and it helps in quantifying trends in data such as connection between students' grade. If you take example of once in a class, so we can able to quantify, for example for students grades and hours spent by student for studying. So, how students are performing those those who spend more time in study. So, we can quantify their connection through the regression analysis.

This also helps in explaining the effect of increasing or decreasing price of a product on its demand. The impact of price on demand can also be explained. Those who want to have some market-based analyses for pricing strategy. So, another example of regression or the linear regression is to understand the effect of education on income and there are so many examples of such varieties, where the regression or the simple regression is going to be useful.

So, the regression is a set of technique for estimating a relation between variables where changes in some variable may explain or possibly cause changes in other variables. Basically, we are going to find out the relationship between one set of variable with another set of variable.

(Refer Slide Time: 04:54)



❑ The **cause-and-effect relationship** must be determined using a **theoretical model** and or a **logical reason**.

❑ Regression analysis assumes a linear relation between the predictor and the outcome variable.

❑ **Uses of Regression analysis**:

- ❑ **Predict** the value of a dependent variable based on the value of at least one independent variable.
- ❑ Explain the **impact of changes** in an independent variable on the dependent variable.
- ❑ Determine the **strength of relationship** between one dependent variable (Y) and series of independent variables.
- ❑ **Forecast** value of Y variable from the value of X's variables.

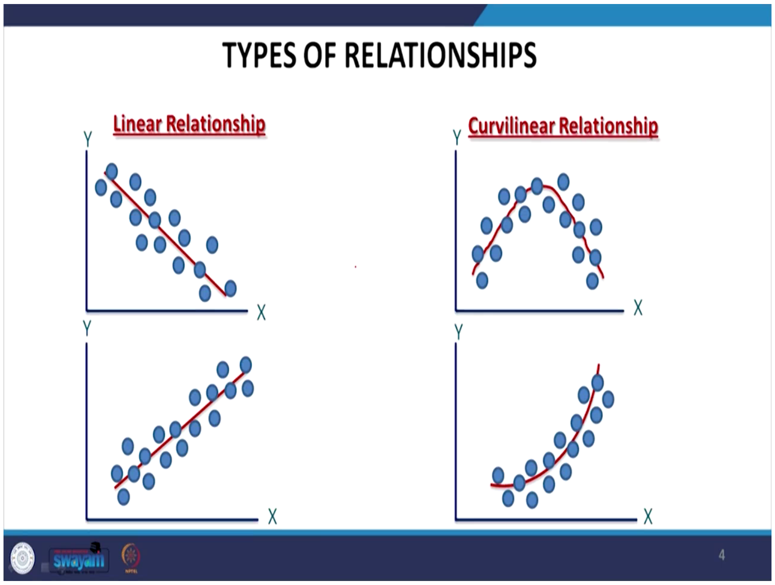
We might be little confused with correlation, but correlation only established the relationship. We are not just going to find the relationship rather we are going to find certain inferences out of it, some logical conclusion out of it and those are based on some theoretical models. Theoretical model we will explain steadily. So, regression analysis assumes a linear relation between the predictor and the outcome variable. So, predictor we wanted to say broadly speaking, the independent factor, independent variable with that of the outcome as the dependent variable.

We use regression analysis to predict the value of a dependent variable based on the value of at least one independent variable. To be noted, it may be one also, but then that model will be called bivariate kind. We are going to discuss our simple regression model. The regression analysis also explains the impact of changes in an independent variable on the dependent variable.

So, impact of changes when this is written then it also extends some other possible answers, but at this moment in this lecture, we are not clearly explaining the impact as such but we are finding out some possible impacts, but impacts cannot be rationally validated through the OLS or the simple regression model.

This regression analysis determines the strength of relationship between one dependent variable, as we are already cited the example, between one dependent variable that is Y we generally denote with Y and a series of independent variables usually denoted with X. So, this also helps in forecasting values of Y variable from the values of X variable. Broadly speaking, we can predict for some future variable.

(Refer Slide Time: 07:10)



TYPES OF RELATIONSHIPS

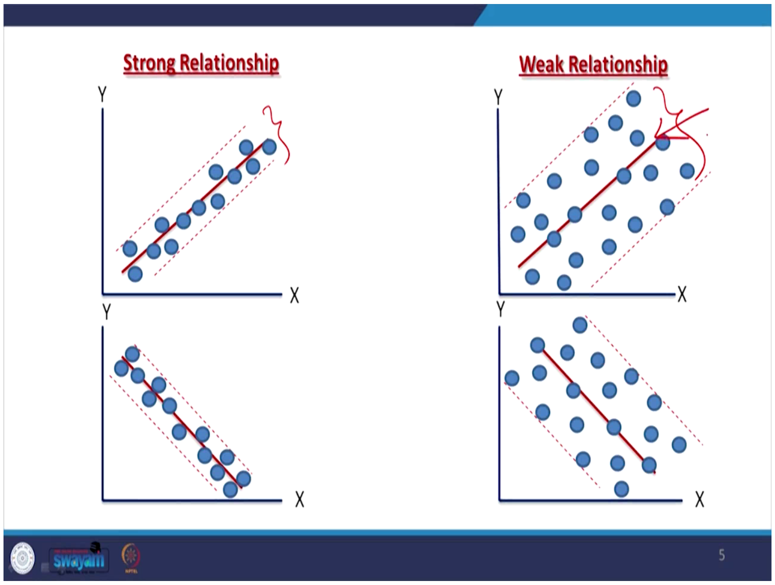
Linear Relationship

Curvilinear Relationship

Let us come to some basic fundamental aspects through the diagram that if there exists linear relationship between two variables, independent and dependent predictor or outcome variable, so on the left hand side two diagrams speak about linear relationship, because the trend of rise or fall is more or less confined within a linear line. A linear line can be predicted out of it or estimated out of it.

Whereas in case of the other two, the right hand side diagrams, the trend seems non-linear, its neither just straight line, neither straight line upward or downward both the aspects are visible it is upward and the moments are very dynamic. So, it is largely mentioned as curvilinear relationship. So, this is also called non-linear relationship. This is also called, sometimes we say, quadratic relationship if quadratic relationship is there defined and identified by our tools. To simplify further the relationship between two variables, if on the left hand side again, if it is confined within certain limit, upper and lower limit as highlighted by the red lines.

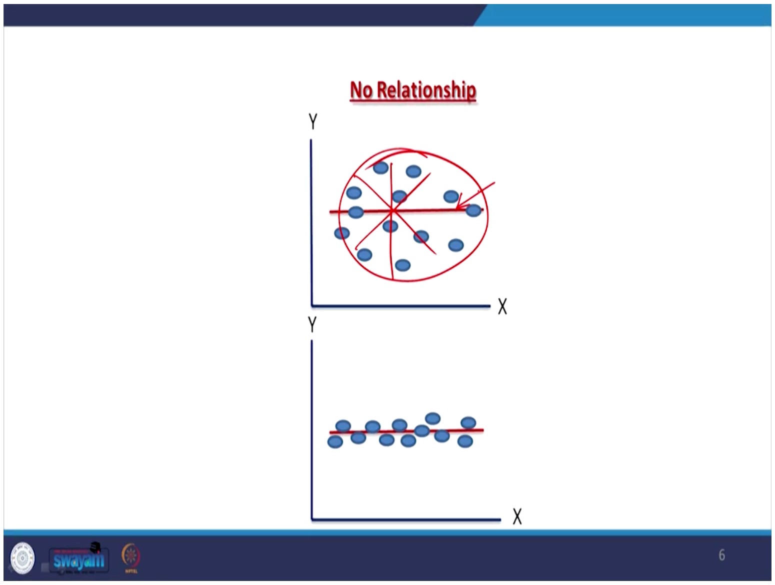
(Refer Slide Time: 08:36)



So, that we say they have strong relationship. But on the right hand side, the variability is much higher from the trend line that is the bold line, if the variability is much higher, but still captioned within a limit but the limit is defining as having weak relationship. In this case, the variability is very, very less, whereas in this case the variability is much higher from the trend line.

So, those who are new to econometrics model, I am trying to explain for you because for students, economics or management who have already applied linear regression model, it might be repetitive, but we have to do it for our larger audience. So, I am trying to clarify. Coming to the relationship between X and Y, when there are no clear trend defined and we are forcefully deriving a trend line, but the trend line, though it shows a trend line but it is not capturing all the variabilities correctly.

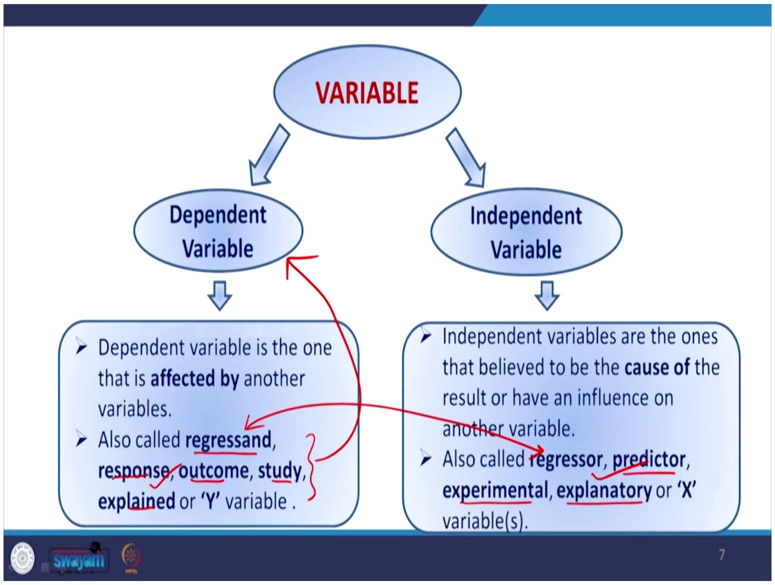
(Refer Slide Time: 09:51)


No Relationship

So, it seems there is no we can draw the trend, the trend line could be like this also, could be like this also, could be like this also, any possibility is there in this particular relationship. So, the computer might generate in one instance it might simply generate a trend line but that is not the final one. There is multiple trend line possible. So, it is not defined as having any relationship.

Whereas in the bottom diagram it clearly shows that the trend line does not have any variability. It has a trend line, but it is constant, isn't it? So, irrespective to the change in X, Y is not going to change. So, when Y is not going to change that means you are not going to find out any implications of X on Y. Whereas in case of the first diagram, the variability is very sporadic. So, our value is going to be very very close to 0. So, second one basically it says no trend. It is constant, more or less constant.

(Refer Slide Time: 11:00)



VARIABLE

Dependent Variable

- Dependent variable is the one that is **affected by** another variables.
- Also called **regressand**, **response**, **outcome**, **study**, **explained** or **'Y'** variable .

Independent Variable

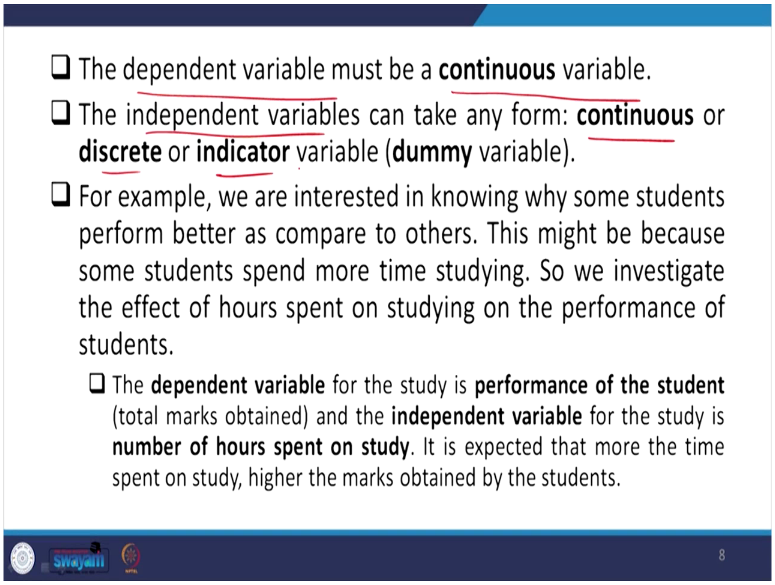
- Independent variables are the ones that believed to be the **cause of** the result or have an influence on another variable.
- Also called **regressor**, **predictor**, **experimental**, **explanatory** or **'X'** variable(s).

Coming to further clarification of the variables those are dependent and independent, though we all know the meaning of it but there is some important aspect we need to highlight. So, broadly it is dependent variable or independent variable. Dependent variable is the one that is affected by another variable or another set of variables. Also, this dependent variable is called in econometrics language we say regressand, response variable, outcome variable, study variable, explain variable or simply we take Y as the notation for it.

So let me come to this point once again, so do not be confused that if any time they are using regressand, response, outcome variable in some of the books, so do not get confused they simply mean this. Similarly, in this case, independent variables, these are the ones that believed to be the cause of the result or have an influence on another variable. This has influence on another variable.

This is also called in econometrics language regressor. This is called regressand and here it is called regressor, predictor, experimental variable or explanatory variable. Here we say explained variable, here we are saying explanatory variable and noted by X or Xs. A number of variables of X is there then X (i□n) is generally mentioned. We are going to discuss in our slides.

(Refer Slide Time: 12:45)



- ❑ The dependent variable must be a **continuous** variable.
- ❑ The independent variables can take any form: **continuous** or **discrete** or **indicator** variable (**dummy** variable).
- ❑ For example, we are interested in knowing why some students perform better as compare to others. This might be because some students spend more time studying. So we investigate the effect of hours spent on studying on the performance of students.
  - ❑ The **dependent variable** for the study is **performance of the student** (total marks obtained) and the **independent variable** for the study is **number of hours spent on study**. It is expected that more the time spent on study, higher the marks obtained by the students.

8

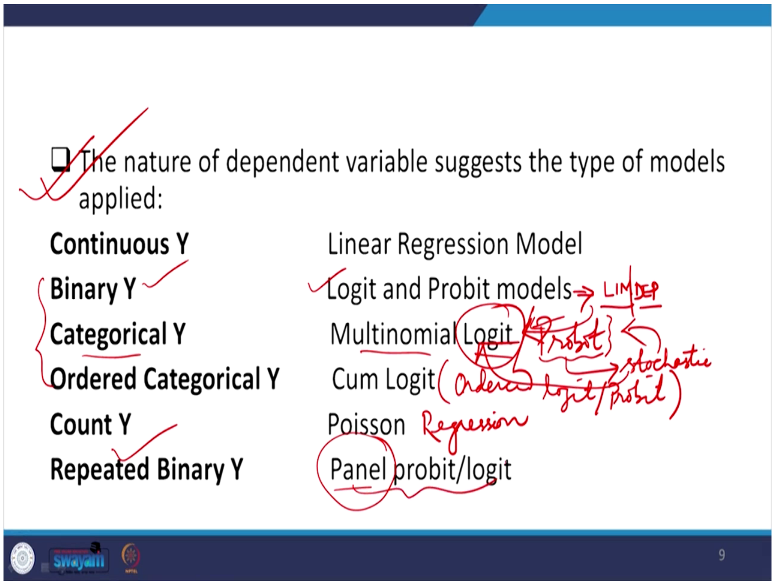
So, the dependent variable must be a continuous variable so far as our regression is concerned. We are going to discuss that if it is not continuous then there will be some advanced model. We will continue those discussions in our next week, not in this week. So, our next module is going to discuss or address non-continuous variable as well. But at this moment we will be emphasizing continuous variable for a simple regression analysis. The independent variables can take any form. So restriction is independent variable. I am just clarifying to everyone.

Restriction is very little, there are some possibilities, very little independent variable. It could be of any form, maybe continuous, maybe discrete, maybe indicator variable, indicator, simply a dummy variable. Yes, and no, it is not a continuous variable nor a discrete variable. Those clarifications we have already made in our previous lectures, regarding defining variables.

So, coming to the explanation with an example, suppose we are interested in knowing why some students perform better as compared to others. This might be because some students spent little more time in their study. isn't it? Better to investigate the effect of their relative more hours spent on studying as compared to others in terms of their performance as an outcome variable. So, for that our dependent variable in this particular example is our performance of the students in terms of grade or the marks obtained.

Whereas the independent variable for this particular one is hours spent for studying. So, we are now mentioning the quality hours. Quality hours we can discuss, but in our example it is only simply hour. So, higher the marks obtained by the students can easily be estimated through this approach.

(Refer Slide Time: 15:02)



□ The nature of dependent variable suggests the type of models applied:

Continuous Y	Linear Regression Model
Binary Y ✓	✓ Logit and Probit models → LIMDEP
Categorical Y	Multinomial Logit {Probit} → stochastic
Ordered Categorical Y	Cum Logit (Ordered logit/Probit)
Count Y	Poisson Regression
Repeated Binary Y	Panel probit/logit

This slide is very very interesting to note, because we are addressing some clear indication of the right model to be picked up given the data and their nature of data. The nature of dependent variables suggest some type of models applied and please take a note in my knowledge the most important aspect to be guided from our lecture of this today.

That if your Y that is the dependent variable or the explained variable, regressand, and that is continuous and you should apply linear regression model. But as I just said, if it is not continuous but it is in binary form, then 1 0 kind are there, then the dependent variable is no more continuous.

So, most appropriate model is logit or probit. Then these are also called, logit and probit models are also called LIMDEP. Why LIMDEP? In short I am using, but there is a package, there is a statistical package called LIMDEP, but I am not going by the package, I am simply trying to say limited dependent variable.

Whereas the dependent variable is limited by certain particular discrete numbers and those are in binary forms. If it is in categorical form, also LIMDEP up is applied, but that is also LIMDEP, limited dependent variable models are applied if it is categorical, not binary. So, in that case it will be multinomial logit or multinomial probit.

Question is whether logit or probit to be applied it depends upon the distribution function of the error term you have. If the error term distribution is stochastic, then probit is the best fit. Otherwise, if it is non-symmetric then logit is the best fit. We will explain in our next week lectures on limited dependent variable models. So, we will clarify further details to it.

Now coming to ordered categorical dependent variable, when it is ordered, not necessarily it is binary, not necessarily it is categorical, but ordered form. That means like I have given you the example of standard education case, even in standard of living also, the lower standard of living then middle and upper, those are clearly categorical but ordered. In that case, we use cum logit model or that is sometimes we said ordered logit model. If you search it in STATA or even in Google, ordered logit or probit model, so those clarifications will be coming to your screen. So, in short it is called cum logit model.

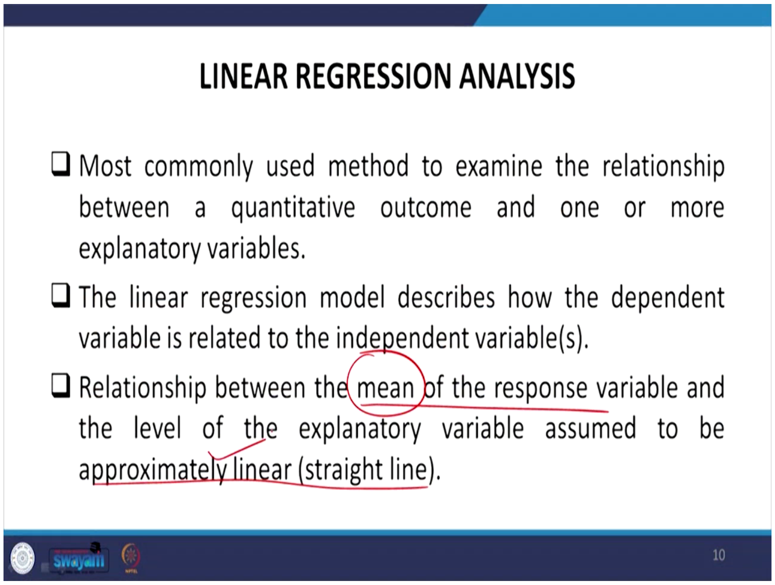
There are some other classifications, other important, I have discussed these in one category. Another aspect I am going to emphasize here, if your Y is count data, there is no continuity, only count so many important points of counting is possible it does not have variability in its own data, particular pockets of counting is there, if it is count data, then Poisson regression is the best fit.

So, Poisson regression is the best fit and you must apply accordingly. But right now, we are not explaining everything. We will open the description a little more later, but not exactly all the instruments we are going to use, because of time paucity, because of confining the analysis within the limit of our module.

So, if we have repeated binary dependent variable, your dependent variable you have in particular time period that is, let it be in binary or even any other form if that gets repeated in another time period, usually we refer to as panel.

So, panel having a time component, since it is a time component in different time period there is a possibility of repetition of that numbers of the dependent variable, so in that case we will be applying panel probit or panel logit depending upon the distribution of the data. So, these are I think a very good starting point. This slide I suggest everyone please try to logically remember and for better application. let us make a move to the some set of other arguments, other understanding of linear regression analysis.

(Refer Slide Time: 20:24)



LINEAR REGRESSION ANALYSIS

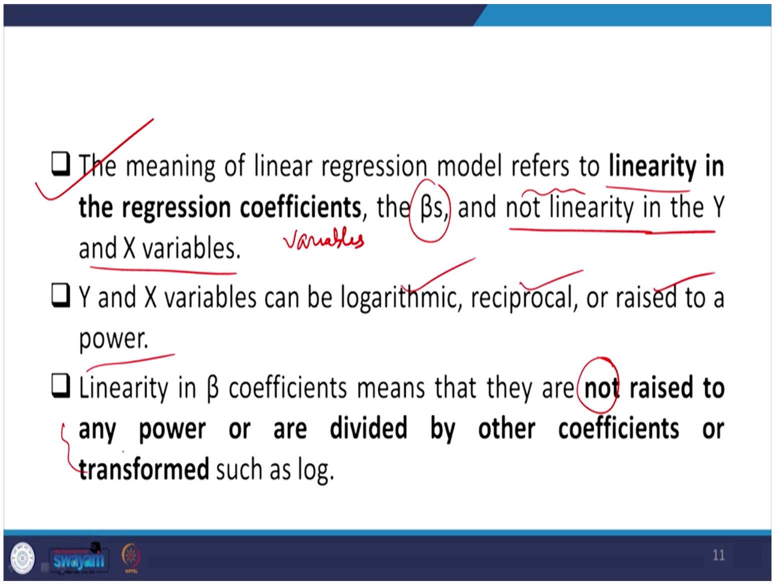
- Most commonly used method to examine the relationship between a quantitative outcome and one or more explanatory variables.
- The linear regression model describes how the dependent variable is related to the independent variable(s).
- Relationship between the mean of the response variable and the level of the explanatory variable assumed to be approximately linear (straight line).

10

This is most commonly used method to examine the relationship between a quantitative outcome and one or more explanatory variables. We have already discussed couple of minutes back. The linear regression model really describes how the dependent variable is related to the independent variable. So, this relationship between the mean of the response variable and the level of explanatory variable assumed to be approximately linear or straight line that is important.

The mean response is important, because where to compare. The trend value always takes the mean. From the mean response variable and the level of this explanatory variable that really assumes a linear approximation that is usually a straight line. We have already discussed.

(Refer Slide Time: 21:14)



- ❑ The meaning of linear regression model refers to **linearity in the regression coefficients**, the βs, and not linearity in the Y and X variables. Variables
- ❑ Y and X variables can be logarithmic, reciprocal, or raised to a power.
- ❑ Linearity in β coefficients means that they are **not raised to any power or are divided by other coefficients or transformed** such as log.

11

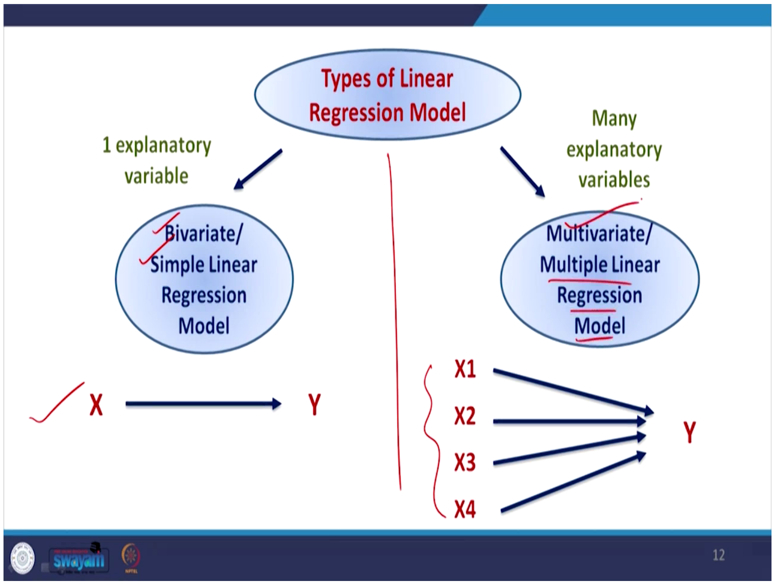
The meaning of linear regression model refers to linearity in the regression coefficient. This is another most important aspect must be understood and usually this is asked in most of the interviews to the candidate who appearing for faculty position, appearing for PhD selection. Why am I saying, linearity in what, your linearity in the regression coefficient, not with the regression or not with the variables.

You have to clarifying very clearly that not necessarily it should be linear, not necessarily your model should be linear with the variables, variable might be in square form, maybe non-linear form, but your coefficient that is, we are going to discuss what is called coefficient, what are the variables in the equation form in our next slide, I am just trying to give you the background that the linearity in the regression coefficient that is the betas we are emphasizing and not linearity in the Y and X variable.

So, not linearity which I just wanted to say Y and X variable it may be linear may be non-linear it does not matter. It only matters that the coefficient which we are going to estimate, our estimation should be linearly established. So, if that is in beta square and beta cube, it is very difficult to estimate. So, Y and X variable can be either, Y or X variable as I just said can be non-linear, we wanted to emphasize could be in logarithmic transformation, maybe reciprocal or maybe raise to any power.

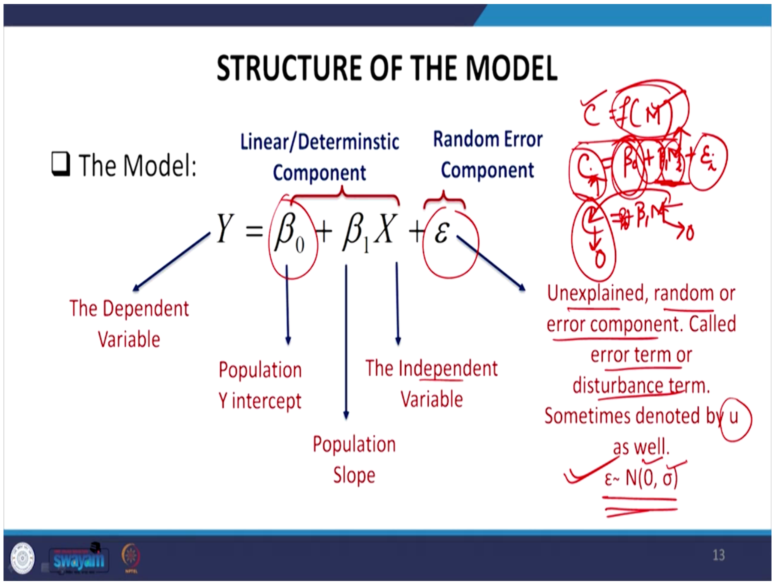
So, linearity in beta coefficients means that they are not raised to any power and/or divided by other coefficients or transformed such as log. So, basically it should not be in this form. So, not is the word that means we should not, our beta coefficient should be linear enough.

(Refer Slide Time: 23:25)


Types of Linear Regression Model
1 explanatory variable
Many explanatory variables
Bivariate/ Simple Linear Regression Model
Multivariate/ Multiple Linear Regression Model
X
Y
X1
X2
X3
X4
Y
12

So, regarding the types of linear regression model, we have some clarification here. You might be a little confused with this. I am just dividing into two. One is when our explanatory variable is only one with the dependent variable, if it is one that is either called bivariate model or it is called simple regression model. It is like this to this. X and its relationship with Y. Whereas when there are more than X that is a set of X variables are there that is called multiple regression model or also called multivariate analysis.

(Refer Slide Time: 24:07)



STRUCTURE OF THE MODEL

❑ The Model:

Linear/Determinstic Component

Random Error Component

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The Dependent Variable

Population Y intercept

Population Slope

The Independent Variable

Unexplained, random or error component. Called error term or disturbance term. Sometimes denoted by u as well.

$\varepsilon \sim N(0, \sigma)$

$C = f(M)$

$C_i = \beta_0 + \beta_1 M + \varepsilon_i$

$C \neq \beta_1 M \to 0$

Let us come to the understanding of a model of a simple linear structure of the model, which says that, if we can write it down like Y is equal to beta not plus beta 1 X plus epsilon. Epsilon we referred to the random error term and beta not is the intercept, the population intercept and beta 1 is called the slope of the variable X that is the independent variable and Y is the dependent variable we start with.

Sometimes you might be having some question later to beta not, why we are supposed to write down. We will again clarify from our diagram. Our diagrams are going to establish a better understanding. Just to mention you the role of intercept with an example that we have a dataset of consumption expenditure as a relationship with income.

So, suppose my dependent variable is expenditure on consumption items as a function of income of the consumer per capita income of the consumer. So, my income is in the right hand side and C is a function of Y, suppose Y stands for money income, C is a function of M.

How should I write down? I should write down here as C, when I am writing simply C is a function of M. I will write down like here as maybe beta not plus beta 1 M and then epsilon i. If I just go by i every time, then we have to write down i. That stands for so many individual responses.

What I wanted to refer in case of this function that I wanted to clarify this beta not term. M with the risen of M, higher the income of the consumer, higher the expenditure is expected. That exist a positive relationship, so beta one is expected to be positive. The estimated value of beta 1 is going to be positive while understanding consumption expenditure with reference to money income.

But suppose income of the person is 0, suppose by any reason some unexplained factors there are some exigencies in the economy or with the person that the person could not be able to earn any money. So, does that mean, the equation could have been like C is equal to only beta 1 M.

Suppose I am writing down this instead of beta not, what is the danger we are going to find? So, when this is equal to 0, C should tend to 0. But in reality this is not equal to 0. Even if M income, the person does not have income, the person might get loan to get its consumption fulfilled.

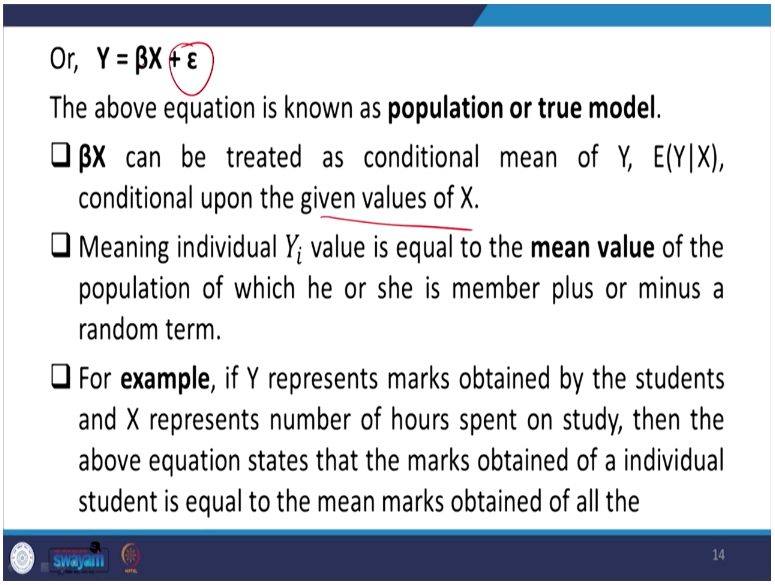
So, even if M stands at 0, though the consumer has to consume something for his or her livelihood. Even if this part is 0, so there is a positive consumption, there must be an autonomous consumption. This is called autonomous which is not affected by your M.

So, beta not has to be mentioned. Beta naught stands for an autonomous spending by a particular consumer and epsilon every time stands for some error, like whatever we expect with a 1 unit change in M may not be exactly predicting to the change in consumption. There are some fluctuations. The fluctuations are identified by the epsilon. We are going to explain everything in our diagram.

So, error term is called unexplained, random, error component. This is also called error term or disturbance term, sometimes denoted by u as well and it has to follow a distribution that is one of the important assumption and every time we are going to recall or remind you during our lectures from now onwards for all the models that how does your error term explain.

If the error term itself distributing correctly, it follows a standard normal distribution with 0 mean, a sigma that is standard deviation, if that is there, then we are referring to the error term, a standard normal distribution. If that is true that means our whole data is standard normal, because if error is fluctuating correctly with a normality that means your data is also following a normal distribution. That is one of the important crux point we wanted to highlight. We will discuss everything eventually.

(Refer Slide Time: 29:32)



Or, $Y = \beta X + \varepsilon$

The above equation is known as **population or true model**.

- ❑ **βX** can be treated as conditional mean of Y, E(Y|X), conditional upon the given values of X.
- ❑ Meaning individual Y_i value is equal to the **mean value** of the population of which he or she is member plus or minus a random term.
- ❑ For **example**, if Y represents marks obtained by the students and X represents number of hours spent on study, then the above equation states that the marks obtained of a individual student is equal to the mean marks obtained of all the

Coming to the same simplified model beta X plus epsilon, the above equation is also known as population or true model. Beta X can be treated as a conditional mean of Y. What do you mean by beta X, how we interpret beta X, this is called conditional mean of Y. That is expected value of Y conditional upon the given values of X.

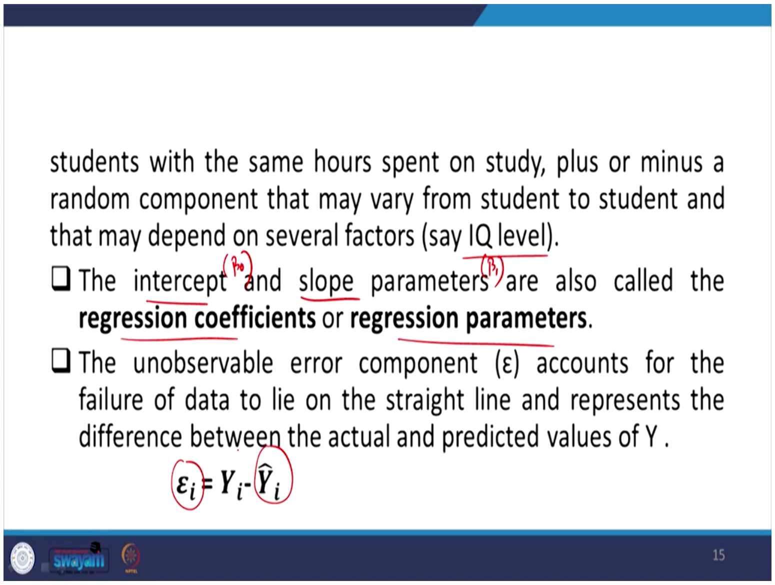
Generally, how we define dy, if I simply take dy/dx, we get a beta value. That beta value is corresponding to a particular unit change in X. Why I am saying unit change in X, because beta is not just defined in isolation, it is due to Y given X. So, that has to be understood. This is a conditional mean, not just an individual mean.

So, the meaning individual Y_i, what does this mean, individual Y_i value is equal to the mean value of the population of which he or she is member plus or minus random term. So, error term could be plus or minus that has to be understood. In our example, Y represents marks obtained by the students and X represents the number of hours spent, the example we have already cited.

In that case, the above equation states that the marks obtained of an individual student is equal to the mean marks obtained of all the students with the same hours spent on study or plus and minus by a random component. That may vary from student to student and that may depend on several factors, say other factors not just the hours important, like the IQ level of the person along with that if you are dealing with multiple regression models.

So, just bivariate analysis is not enough that is the reason why there are lots of criticisms to the bivariate models. We have to include other factors as well.

(Refer Slide Time: 31:36)



students with the same hours spent on study, plus or minus a random component that may vary from student to student and that may depend on several factors (say IQ level).

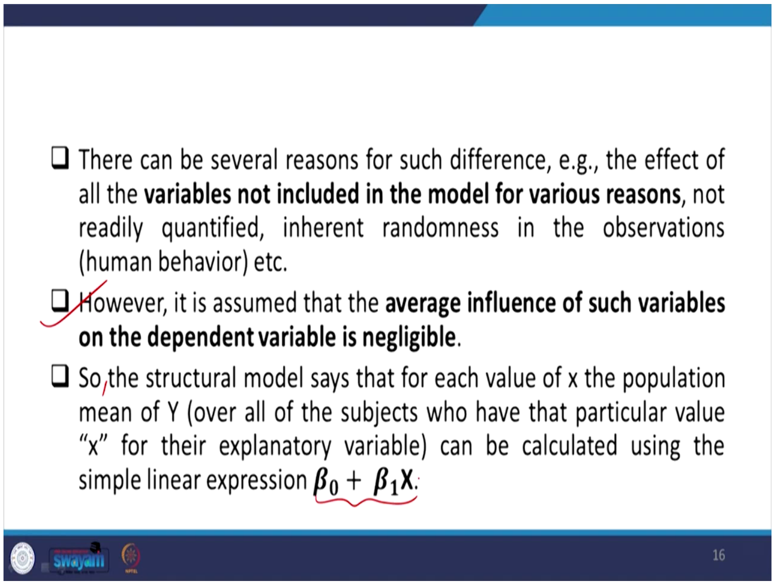
- ❑ The intercept (β_0) and slope parameters (β_1) are also called the **regression coefficients** or **regression parameters**.
- ❑ The unobservable error component (ε) accounts for the failure of data to lie on the straight line and represents the difference between the actual and predicted values of Y .

$$\varepsilon_i = Y_i - \hat{Y}_i$$

The intercept and slope parameters, intercept, which I have started emphasizing and slope parameter that is beta, intercept is beta not, so slope parameters are also called regression coefficients or regression parameters. This is going to be used throughout our lectures. The unobservable error component that is epsilon accounts for the failure of data to lie on the straight line and represents the difference between the actual and predicted values of Y.

So, the error term which we wanted to mention if it deviates from a standard line or standard trend line that is also called the straight line that is simply calculated between the changes between the population that is Y the dependent variable minus the estimated value of that Y_i. Y hat we wanted to mention, hat stands for estimated value. That is basically beta X. The estimated value we wanted to find out through the beta X. So, beta not plus beta one X in our model can be calculated.

(Refer Slide Time: 32:48)

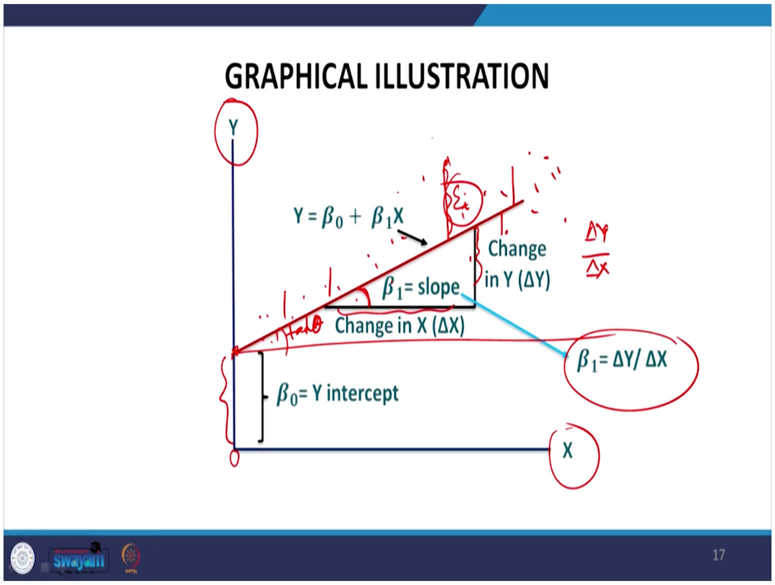


- □ There can be several reasons for such difference, e.g., the effect of all the **variables not included in the model for various reasons**, not readily quantified, inherent randomness in the observations (human behavior) etc.
- □ However, it is assumed that the **average influence of such variables on the dependent variable is negligible**.
- □ So, the structural model says that for each value of x the population mean of Y (over all of the subjects who have that particular value "x" for their explanatory variable) can be calculated using the simple linear expression $\boldsymbol{\beta_0 + \beta_1 X}$.

So, there are several reasons for such differences in terms of epsilon that is the effect of all the variables not including the model for various reasons, not readily quantified sometimes some variables and there is some inherent randomness also in the observations like human behavior they respond in a wrong manner, which is little different than that of the average numbers. So, however, it is good to assume that the average influence of such variables on the dependent variable is negligible.

So, if that impact is not negligible then it will create disturbances to the model. So the average disturbances should be negligible that is the one of the important assumptions to be made in all the models. So, that is why it distributed with 0. That is why I said it is a standard normal distribution with a mean of that error term is 0. So, the structural model says that for each value of X the population mean of Y that is overall the subjects who have that particular value of that X for their explanatory variable can be calculated using the simple linear expression with the help of this, beta not plus beta 1 X.

(Refer Slide Time: 34:14)



GRAPHICAL ILLUSTRATION

From this diagram we will explain that what is all those terms. When I say I have already given an example of intercept, it must have some intercept between the relations of X and Y, Y is our dependent variable. So, it has to be starting with an intercept, because even if X is 0 there must be some positive portion that is also called intercept term that is also called autonomous variable because it is not affected by the change in X.

So, once the X gets changed, it always follows certain standard parameter. With a certain particular rate it follows with an average change. So, that parameter is beta 1, we will explain by a slope. That slope can be also explained with this. So, this is equivalent to this. Any line perpendicular, parallel to the X axis this degree that slope can be explained with tan theta. So, can be explained perpendicular by base that is simply, this is your tan theta, this divided by base B. So, basically this is delta Y by delta X, perpendicular by base defines the slope. We need to define in a 90-degree line.

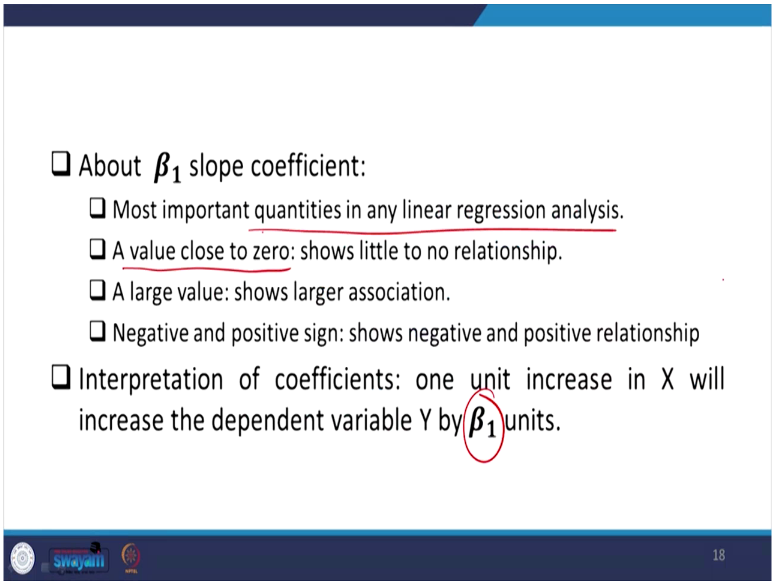
So, what is our beta 1 then, beta 1 is nothing but the slope, slope of that particular line. So, slope is indicating the angle through tan theta we will define and slope is derived through delta Y by delta X. Question arises this line is not always a straight line.

There are many important frequencies maybe like this, like there are many responses. This only gives an average trend line. So average trend line, but assumption is that if this is, there are these points ups and down, the vertical difference between these two and in all other points any other vertical difference that those are called error terms, called epsilon. So, epsilon i, why am I writing i, because there are a number of epsilon, not just one epsilon to a corresponding X, to all those epsilon has to be averaged with 0.

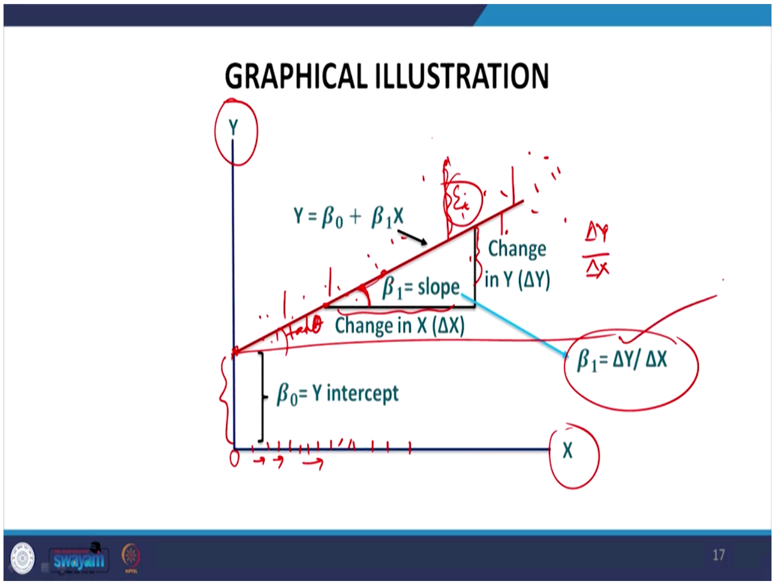
The average impact of this epsilon has to be minimum, minimum was possible. If that is there, that means if the average is zero that means we are following a perfect trend line and there is a relationship between X and Y and then only we can postulate number of interpretation of X and Y, because we have a clear average value and our regression equation explains these things very correctly.

So, I will be explaining another couple of things right now, maybe till this I will be explaining, then we will move into a next class. So at this moment I have explained, so most of the concept I have already delivered and discussed with you.

(Refer Slide Time: 37:20)



- ❑ About β_1 slope coefficient:
  - ❑ Most important quantities in any linear regression analysis.
  - ❑ A value close to zero: shows little to no relationship.
  - ❑ A large value: shows larger association.
  - ❑ Negative and positive sign: shows negative and positive relationship
- ❑ Interpretation of coefficients: one unit increase in X will increase the dependent variable Y by β_1 units.

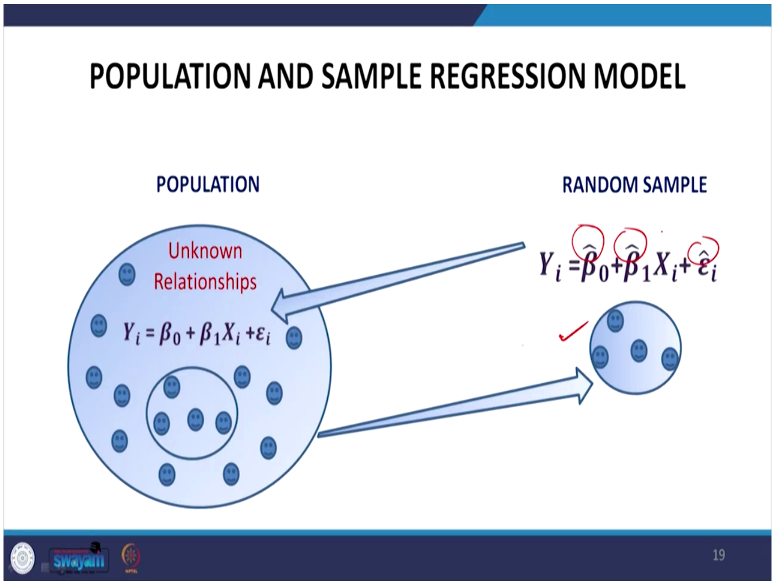


GRAPHICAL ILLUSTRATION

So, about beta 1 coefficient, slope coefficient I said that it is most important quantities in any linear regression analysis. A value close to 0, what does it mean, that means it shows little to no relationship. There is no relationship. How does it happen? When there are many error terms, if the trend line is very zigzag, the error term is not having any average, 0 error value. So, that means it will be close to 0. That means it might have no trend line, no clear angle, no slope is defined. A large value shows a larger association between Y and X.

So, negative and positive sign shows negative and positive relationship accordingly. The interpretation of this coefficient is that with one-unit change, every time as I said, with one unit change every time fraction unit there are many changes like this, like this, like this, every 1-unit change, what is the impact on the Y term. So, every 1-unit change what is the impact on the dependent variable can be explained and the entirety is explained by beta 1, because that is the slope we have already clarified.

(Refer Slide Time: 38:46)



POPULATION AND SAMPLE REGRESSION MODEL

POPULATION

RANDOM SAMPLE

Unknown Relationships

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

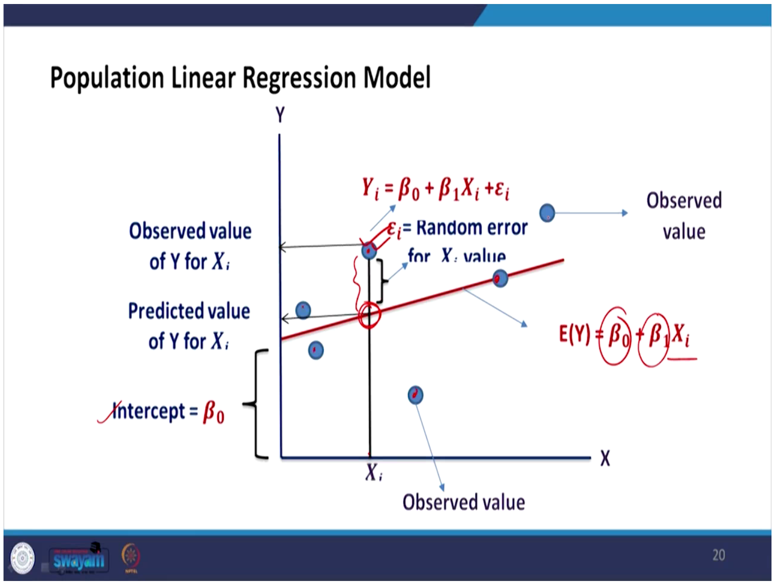
$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$

✓

Coming to the population, so far we are explaining the population parameters. We have to explain sample statistics. In case of parameter we referred to population. Statistics we referred to sample. In that case, we are supposed to deal with the estimated value, because it is not possible to go by all population for the regression model. We have to take some sample. So, like here in the left hand panel there are sample, the population is given, which is randomly distributed and we do not know the relationship. We have to test by some sample.

Suppose the sample is taken here. This is the sample picked up, a random sample is taken and we have to take so many samples from the population, so that our estimation does not vary much that means our estimation is good. That sample, out of the population now we are deliberately mentioning a hat term, because we are going to estimate the sample, error estimation as well.

(Refer Slide Time: 39:53)



Population Linear Regression Model

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

ε_i= Random error for X_i value

Observed value of Y for X_i

Predicted value of Y for X_i

Intercept = β_0

$E(Y) = \beta_0 + \beta_1 X_i$

Observed value

Observed value

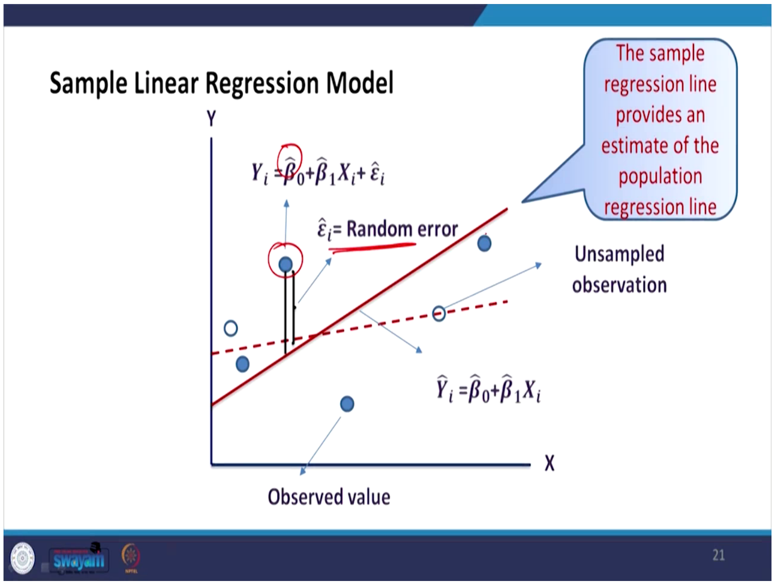
Population linear regression model which we have explained, but what I am trying to find out is that, intercept I have already said, from the linear line as highlighted in the red line, trend line that is simply mentioning the expected value of Y were expecting what is the value of Y through the two parameters here that is beta 1 and beta not with the help of changing X.

e_i is the error term we already discussed. These are called, in this model, since we have already started discussing about sample, we are saying these are called observed value. So the observed values and now try to understand one point here.

Suppose I picked up this particular point, what does this mean, that this is indicating the value of Y at this point and this also indicated a value of X. In that particular line from this particular point till this, we reach at a predicted value. Predicted value stands at the red, red line is a predicted line. So, the predicted value of Y for X_i is indicated at this line, at this particular point. So, that is all about our predicted value.

So, then above to that what is this all about, that is basically the change from the predicted to other variable. The actual value is here. So, the gap between these two is nothing but called error term that is also called random error for the given value X. Similarly, there are many points, maybe this point, maybe this point, maybe this point, there are many possibilities, so random error term could be estimated.

(Refer Slide Time: 41:53)



Sample Linear Regression Model

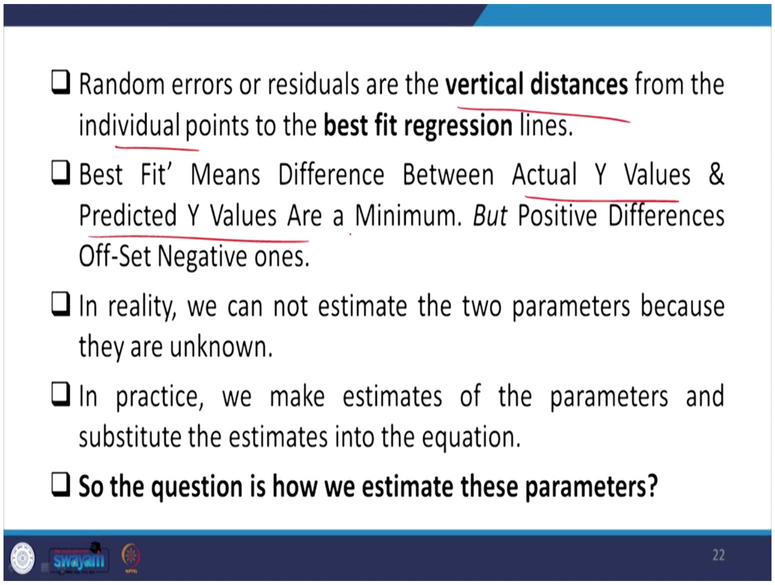
Y
$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$
$\hat{\varepsilon}_i$= Random error
Unsampled observation
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
X
Observed value
21

The sample regression line provides an estimate of the population regression line

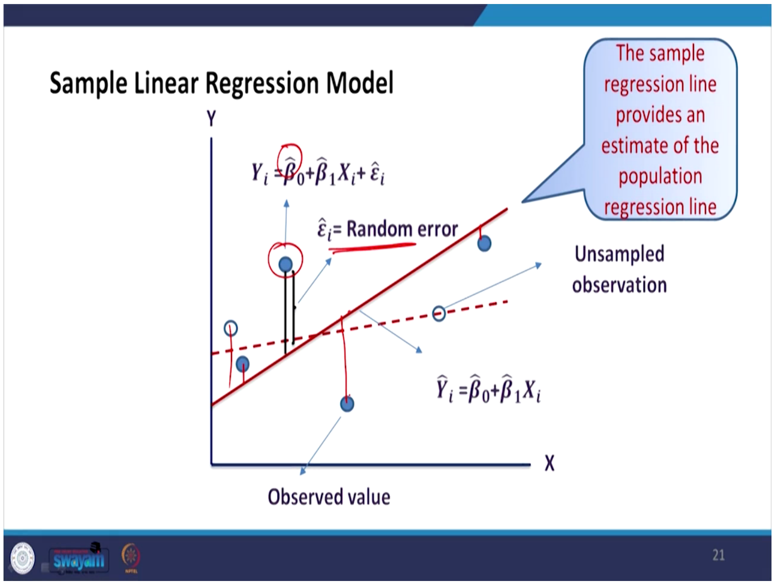
It is also interesting to note that the particular point which we wanted to estimate from the, how deviation is there that is purely identified by the random epsilon term. But our estimation follows with hat, beta hat, beta not hat with beta 1 hat X_i plus epsilon X_i, so through that we can able to estimate.

In this case, what is written here, look at carefully, the simple regression line provides an estimate of the population regression line. So, we are basically trying to estimate the population regression line. We cannot simply say that this is the final estimation. We need to validate it by different reasons. Similarly, there are unsampled observation also, unsampled observation as well that has to be emphasized. So, some interpretations are here mentioned.

(Refer Slide Time: 43:09)



- ❑ Random errors or residuals are the **vertical distances** from the individual points to the **best fit regression** lines.
- ❑ Best Fit' Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative ones.
- ❑ In reality, we can not estimate the two parameters because they are unknown.
- ❑ In practice, we make estimates of the parameters and substitute the estimates into the equation.
- ❑ **So the question is how we estimate these parameters?**



Sample Linear Regression Model

The sample regression line provides an estimate of the population regression line

Y
$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$
$\hat{\varepsilon}_i$= Random error
Unsampled observation
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
X
Observed value

So, random errors or the residuals are the vertical distance, which I already explained from the beginning, the vertical distances from the individual points. So from the trend line, from the individual points like from the standard line if you are considering this, the vertical differences are the residuals into estimate and if it is the residuals are minimum then we can find out a best fit regression lines. So, the best fit means difference between actual value and the predicted values are minimum but positive differences offset negative ones.

In reality, we cannot estimate the two parameters because they are unknown. In practice, we can estimate of the parameters and substitute the estimates into equations. We can also convert it to equation with the parameters. So, the equation is how we estimate these parameters are very very important. So this is all the background information about the linear regression and in our next class we are going to start with ordinary square method and also we will be emphasizing the best fit line with R, R squared, adjusted R squared and so forth, many things are going to be explained. So, let me close here. Thank you.