Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis

Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Linear Regression Analysis
Lecture - 40
Simple Linear Regression Analysis
Basic Concepts and Least Squares Estimation

Hello friends, welcome to the course Essentials of Data Science with R Software - 2 where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. In this module on Linear Regression Analysis, we are going to begin with a new chapter on Simple Linear Regression Analysis.

So, in the last lecture we started a topic on linear regression analysis and I had tried my best to give you an overview of what is regression analysis, and what is our aim, and what we want to do. So, now in this lecture, I am going to start the topic on simple linear regression analysis.

And after this the next chapter will be multiple linear regression analysis. Well, that is important for you to understand what is the difference between the two and why I am doing this chapter first. Well, in the case of simple linear regression analysis we consider only one independent variable that is there is only one variable, which is affecting the outcome.

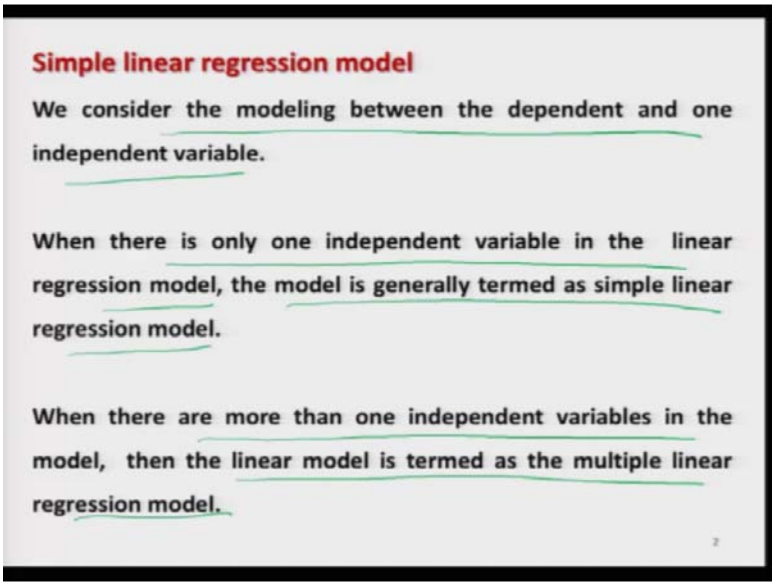
Well, this may not really hold in practice because in practice there are usually more than one independent variables which are affecting the output variable, but through this chapter, I will try to give you the basic concepts, and when you are trying to do in a simple framework with one independent and one dependent variable then it is easier for me to convey the concepts more easily.

And whatever the concepts we are going to learn in this chapter they will simply be extended to the case of multiple linear regression model, but the problem will be that in the case of simple linear regression model I will be able to show you the things more clearly graphically as well as analytically.

And, in the case of multiple linear regression model when we have more than one independent variable then it will be difficult for me to show you each and everything graphically, but whatever the interpretation, whatever the concepts we are going to develop in this chapter they will simply be extended to the multiple linear regression model.

So, this is the reason that why this is very important for all of us to understand and first learn the simple linear regression analysis, right. So, now in this lecture I am trying to give you different concepts, different definitions, and I will try to give you an idea that how the model is developed. And in the case of multiple linear regression model we will be dealing with the more realistic model, ok. So, let us begin our lecture with this slide.

(Refer Slide Time: 03:07)



Simple linear regression model

We consider the modeling between the dependent and one independent variable.

When there is only one independent variable in the linear regression model, the model is generally termed as simple linear regression model.

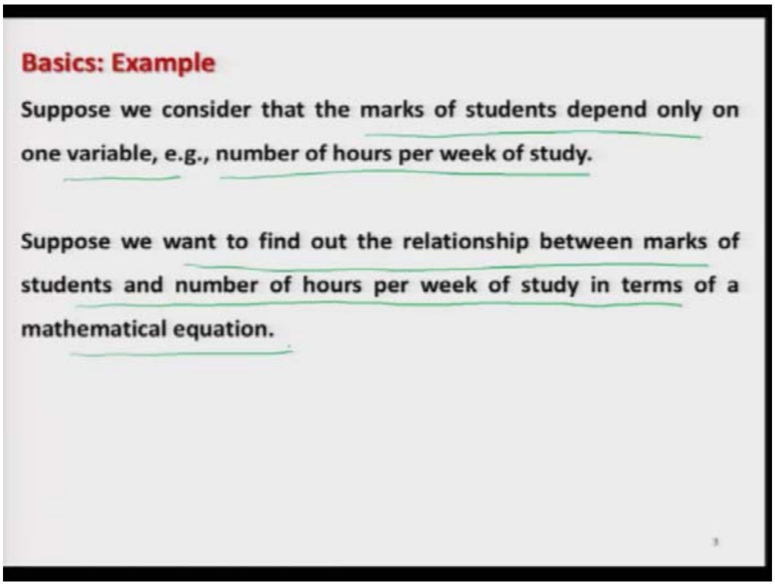
When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.

2

So, in the case of a simple linear regression modeling as I said we consider the modeling between the dependent and only one independent variable, right. And when we are considering and when there is only one independent variable in the linear regression model, then the model is generally called as simple linear regression model. And when there are more than one independent variables in the model then the linear model is termed as multiple linear regression model.

So, this is the basic difference between the simple and multiple linear regression model. There is another terminology and you should be careful about that thing that is called multivariate linear regression model. So, in case of multivariate linear regression model there are more than one multiple linear regression model, which are considered together, but we are not going to consider it here, but that is important for you to know, ok.

(Refer Slide Time: 04:08)



Basics: Example

Suppose we consider that the marks of students depend only on one variable, e.g., number of hours per week of study.

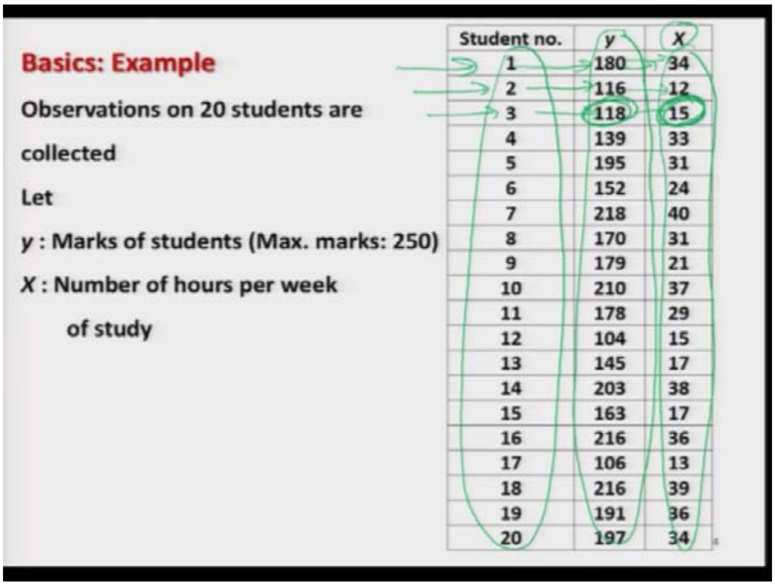
Suppose we want to find out the relationship between marks of students and number of hours per week of study in terms of a mathematical equation.

So, now let me take a simple example and through that example I will try to give you different types of interpretations and various types of concept. So, in the last lecture I had taken an example where we have collected the marks of the student on three different variable, but now in this chapter I am trying to simplify that example and I will be taking the same example, but in the setup of simple linear regression model.

So, we assume that here that we have a data set and where we have considered the marks of the student depend only on one variable and we have considered the variable that was number of hours per week of study. So, we know that the number of hours of per week of study that is going to affect the marks obtained in the examination.

And now our objective is this that we want to find out the relationship between the marks of the student and the number of hours per week of study in terms of a mathematical equation, in terms of a statistical model. So, that is our basic objective, right.

(Refer Slide Time: 05:24)



Basics: Example

Observations on 20 students are collected

Let

y : Marks of students (Max. marks: 250)

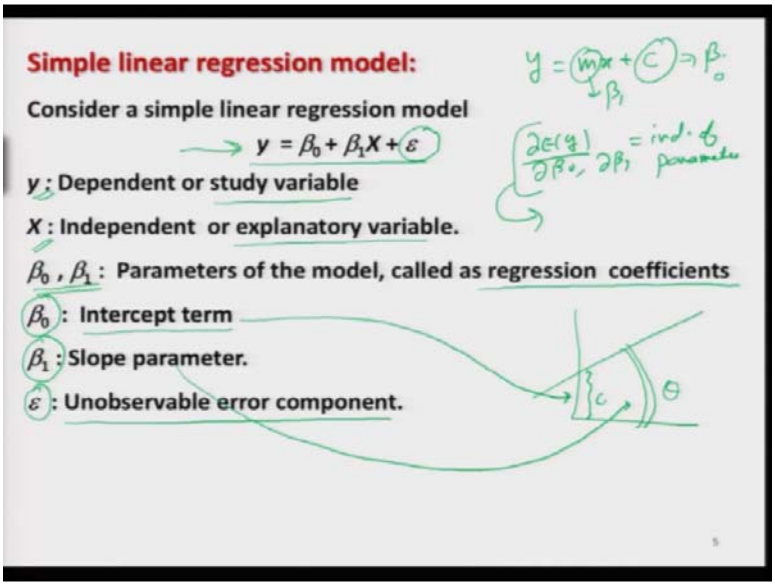
X : Number of hours per week of study

Student no.	*y*	*X*
1	180	34
2	116	12
3	118	15
4	139	33
5	195	31
6	152	24
7	218	40
8	170	31
9	179	21
10	210	37
11	178	29
12	104	15
13	145	17
14	203	38
15	163	17
16	216	36
17	106	13
18	216	39
19	191	36
20	197	34

So, now this is the same data set that we considered earlier, but now you can see here that we have 20 students over here, and we have collected the marks of those 20 student and the corresponding value of X, which is indicating the number of hours per week, which a student has studied. So, these are the corresponding observations.

So, the interpretation goes like this, suppose if I take the student number 1. So, student number 1 has got 180 marks and the student has studied for 34 hours in a week. Similarly, the student number 2 the student has got 116 marks and he has studied 12 hours in a week. Similarly, student number 3 the student has got 118 marks and the student has studied 15 hours a week. So, this is our setup.

(Refer Slide Time: 06:23)



Simple linear regression model:

Consider a simple linear regression model

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

y : Dependent or study variable

X : Independent or explanatory variable.

β_0, β_1 : Parameters of the model, called as regression coefficients

β_0 : Intercept term

β_1 : Slope parameter.

ε : Unobservable error component.

$y = mx + c \rightarrow \beta_0$
β_1
$\frac{\partial E(y)}{\partial \beta_0}, \partial \beta_1$ = ind. of parameter
c
θ

So, now we consider here a linear regression model and this is actually a linear model because, if you remember we had defined that how to call a model to be a linear model. That it mean; that means, the expected value of y or the average value of y is considered and then you try to find out the partial derivatives of expected value of y with respect to each of the parameter, which are β_0 β_1, here.

And, if the partial derivative of say expected value of y with respect to β_0 and with respect to β_1 this comes out to be independent of say these parameters, then this model is called as linear model. And remember one thing when I am talking of the parameters here, parameters are essentially related to the regression parameter which are β_0, β_1.

And similarly in the case of multiple linear regression model the regression parameters will be in terms of β. Why I am telling you here? Because, you will see later on that we will also assume a probability distribution on the random error component and that will have its own parameters.

So, the linearity of the model is defined with respect to the regression parameters. And if you try to see this model which I am denoting here by $y = \beta_0 + \beta_1$ X it is something like y = mx + c, which I had discussed in the last lecture. So, what I have done that this c has been replaced by β term, β_0 and m is replaced by the term β_1, and the interpretation of β_0 and β_1 that is similar to the interpretation of the parameters m and c.

If you remember, if you try to plot a simple equation, linear equation, then the intercept term on the y axis that is indicated by c and the slope of the parameter that is indicated by m that is that tan of this $\theta = m$. So, you will see at a later stage that we will have a similar interpretation for β_0 and β_1 also, but this model or the equation, which I have written here this is comparable with $y = m x + c$, right.

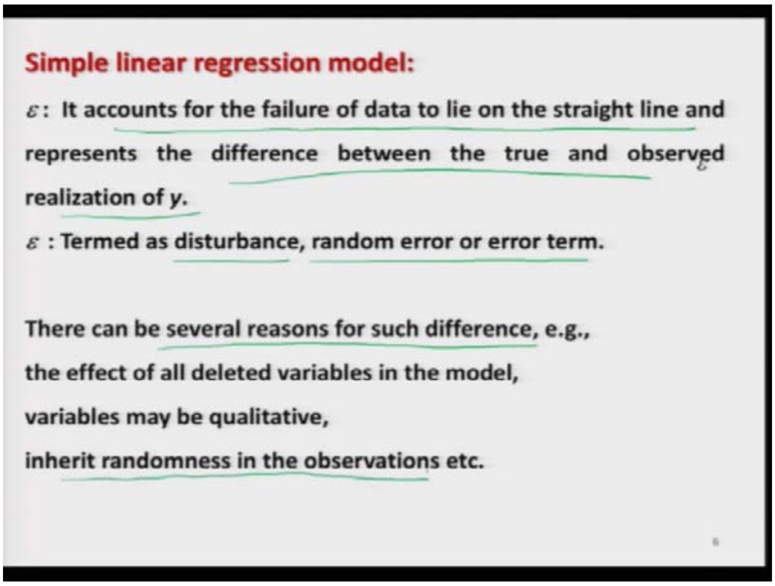
So, in this model which is called here as a simple linear regression model, y is our dependent variable or this is also called as the study variable. And there is only one independent variable, which is denoted as X and this is also called as explanatory variable.

And in the books if you try to see this y and X have been given different types of names also, but I am concentrating on these two terminologies only, and here this β_0 and β_1 these are the parameters of the model which are called as in general regression parameters or regression coefficients.

And out of this these two parameters this β_0 is the intercept term and β_1 is the slope parameter. So, as the name suggests intercept term this intercept term is going to indicate the part on the y axis, which is intercepted by the line and β_1 is going to indicate the slope of that parameter.

And here we are adding here one more term ε, this ε is an unobservable error component because when you are trying to collect the observations in practice this observation will not be exactly following this model, but there will be small deviations and those deviations are coming because of those factor, which are beyond our control and all and the effect of all such factors is being indicated here by the term ε, ok.

(Refer Slide Time: 10:45)



Simple linear regression model:

ε: It accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y.

ε : Termed as disturbance, random error or error term.

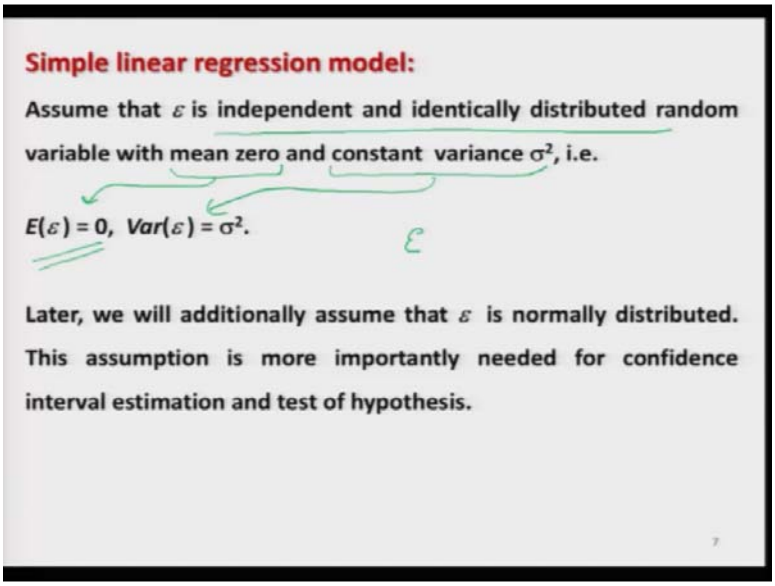
There can be several reasons for such difference, e.g.,
the effect of all deleted variables in the model,
variables may be qualitative,
inherit randomness in the observations etc.

So, and ε has a very important role. Actually in simple words if I try to explain it then ε accounts for the failure of data to lie on the straight line and it represents the difference between the true and observed values of y that we are going to obtain, what is this thing I will try to show you with some graphical also.

And this is actually also termed as disturbance term and this is actually a random error component and this is also called as random error or say random error term. And as we have discussed there can be various reason, there can be several reasons for such differences in the observed value and the true value of the response variable, right.

For example, if you try to consider the in case of multiple linear regression model then it makes better sense that there are some variables which are affecting the output, but we are unable to take observations on them because, they are qualitative, they are not really observable and there will be some inherit randomness in the observations, ok.

(Refer Slide Time: 12:07)



Simple linear regression model:

Assume that ε is independent and identically distributed random variable with mean zero and constant variance σ^2, i.e.

$E(\varepsilon) = 0, \; Var(\varepsilon) = \sigma^2.$

ε

Later, we will additionally assume that ε is normally distributed. This assumption is more importantly needed for confidence interval estimation and test of hypothesis.

So, we assume here, that these random errors are iid that, that means, they are independently and identically distributed since they are random variable. So, we assume that they have got mean zero. So, we assume that here expected value of ε is 0 and we assume that the variance of ε is σ^2, which is constant.

The reason that why we try to take the mean 0 is that when you are trying to observe the observations in real experiment, so since we are not really going to observe the exact values of the y. So, some values are will have the random error in the positive direction and some observation will have random errors in the negative direction.

So, in totality if you try to consider the set of all the observation we can expect that the random error in the population is almost 0, and that we expect the same thing to happen in the sample also. At this moment I am not associating any probability distribution with ε, I am not assuming that whether they are normally distributed or binomially distributed, but I am simply assume that they have got the mean 0 and variance σ^2.

At a later stage when we are going to develop the test of hypothesis and confidence interval then, we will need an assumption on the probability distribution of ε. At this moment we are aiming to estimate the values of regression coefficient and σ^2.

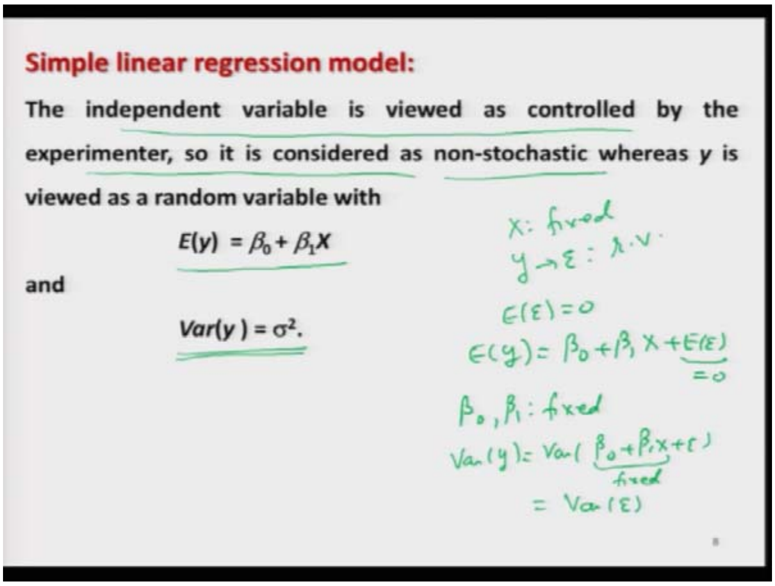
So, as long as we are only doing the estimation process through the principle of least square which we are going to use first we do not need any assumption on the probability distribution. On the other hand, there is another technique which is a maximum likelihood estimation.

So, in case if you are going to choose the maximum likelihood estimation for estimating the regression coefficient; and this variance then in that case, right from the beginning, right from this slide you will need to assume the a probability distribution for this random error component.

So, this you have to be clear that we are going to use here two types of approaches one is principle of least square and principle of maximum likelihood. The principle of least squared do not require any assumption on the probability with density function of ε as far as estimation is concerned, but the maximum likelihood estimation will require the assumption of probability with density function, right from the beginning.

But, after that after estimation when we are trying to construct the confidence interval or test of hypothesis, then in both the cases either the estimators have been obtained by least square or by maximum likelihood, we will need the assumption of some probability density function for ε, ok.

(Refer Slide Time: 15:51)



Simple linear regression model:

The independent variable is viewed as controlled by the experimenter, so it is considered as non-stochastic whereas *y* is viewed as a random variable with

$$E(y) = \beta_0 + \beta_1 X$$

and

$$Var(y) = \sigma^2.$$

x: fixed

$y \to \varepsilon$: r.v.

$E(\varepsilon) = 0$

$E(y) = \beta_0 + \beta_1 x + \underbrace{E(\varepsilon)}_{=0}$

β_0, β_1 : fixed

$Var(y) = Var(\underbrace{\beta_0 + \beta_1 x}_{fixed} + \varepsilon)$

$= Var(\varepsilon)$

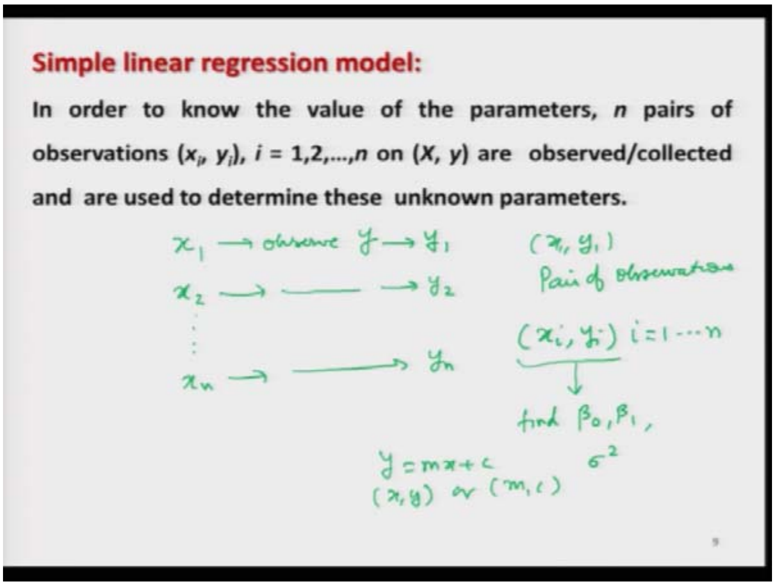
So, now since we have assumed that expected value of ε is 0 and we also have assumed that the independent variable is fixed. It is fixed by because, it is controlled by the experimenter, well in case if the experimenter makes a mistake in giving the value of the input variable means you cannot consider it to be random that is a mistake that is an intentional mistake.

So, that is why this X is going to be considered as fixed or non stochastic. And since y here depends on ε so, this y becomes random variable. And since we have assumed the E(ε) to be 0 then I can write down here E(y) = β_0 + β_1 X + expected value of ε and so, this becomes quantity becomes here 0.

And also note that we are also assuming that β_0 and β_1 are fixed. And as we discussed in the introductory part this β_0 and β_1 sometime can be random also in that case we have a random coefficient model and this explanatory variable or the independent variable can also be random, and in that case the model is called as random regressor model, right. But, anyway we are going to consider here the setup where X β_0, β_1 they all are fixed.

So, in this case E(y) = β_0 + β_1 X and variance of y is given by σ^2. Why? Because, you can see here variance of y will become variance of β_0 + β_1 X + ε and since you are considering this quantity to be fixed. So, that is why this will be same as variance of ε because variance of a fixed quantity, variance of a constant is 0, ok.

(Refer Slide Time: 17:46)



Simple linear regression model:

In order to know the value of the parameters, *n* pairs of observations (x_i, y_i), $i = 1,2,\ldots,n$ on (X, y) are observed/collected and are used to determine these unknown parameters.

$x_1 \rightarrow$ observe $y \rightarrow y_1$ (x_1, y_1)

Pair of observations

$x_2 \rightarrow$ —— $\rightarrow y_2$

$\vdots$

$x_n \rightarrow$ —— $\rightarrow y_n$

$\underbrace{(x_i, y_i)}$ $i=1\cdots n$

$\downarrow$

find β_0, β_1, σ^2

$y = mx + c$

(x, y) or (m, c)

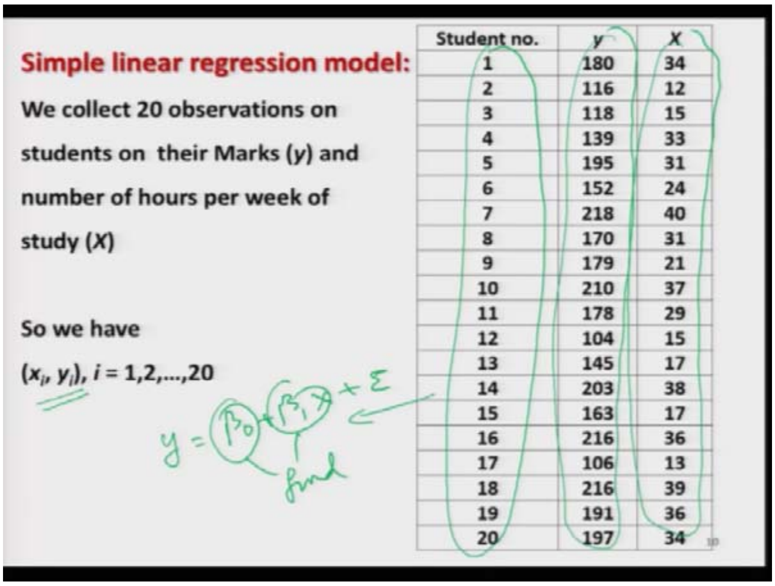
So, now what is the setup that we start conducting the experiment, how? We try to choose a value of x say x_1, and then we try to observe the value of the response variable y which is denoted as say y_1. So, x_1 , y_1 this is going to be the going to be a pair of observations. Why this is pair of observation? Because, x_1 and y_1 both are inter related to each other.

When you are giving the value of x_1 only then you are going to get the value of y_1.So, then we try to repeat this experiment and we choose another value of independent variable say x_2 we give it to the experiment and we then we try to observe y; and then we obtain another value of the response variable say here y_2, and we try to repeat this experiment suppose a small n number of times and we have here the value of response variable as y_N.

So, essentially we will have a small n number of observation which are denoted as x_i, y_i on the independent and dependent variable. And now, what we have; what we have to do? Our objective is this that using these observations we have to find the value of regression coefficient β_0, β_1 and another parameter σ^2, right.

And if you can recall that in the earlier lecture I had given you an example I had shown and I was trying to explain that if in this equation $y = mx + c$, if you know x or y or the second option is if you know m and c then what will happen. So, this is exactly on the same lines that if you note the value of the parameter m and c then, you know the entire line all the properties of the line, where it is intersecting it on the x axis, y axis for a given value of x, what will be the value of y and each and everything. But just by knowing the value of x and y you will not come to know say information about the this linear model.

(Refer Slide Time: 20:17)



Simple linear regression model:

We collect 20 observations on students on their Marks (y) and number of hours per week of study (X)

So we have

$(x_i, y_i), i = 1,2,\ldots,20$

$$y = \boxed{\beta_0} + \boxed{\beta_1 X} + \varepsilon$$

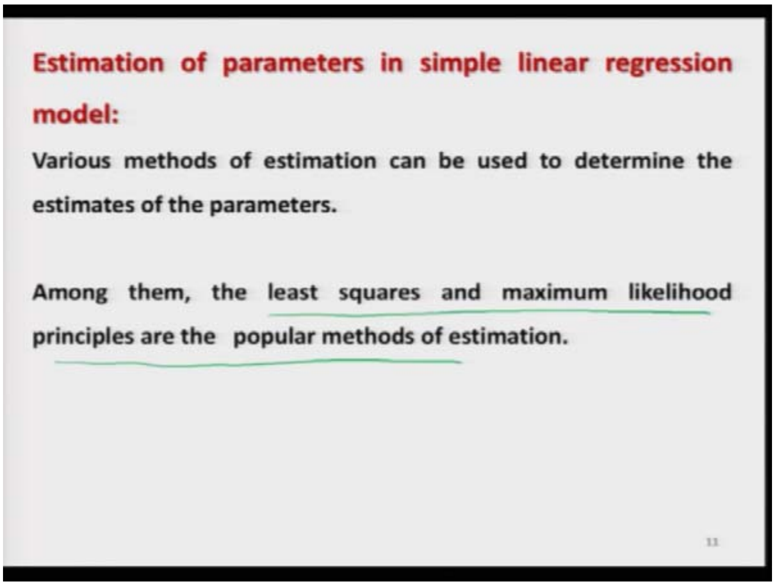
find

Student no.	y	X
1	180	34
2	116	12
3	118	15
4	139	33
5	195	31
6	152	24
7	218	40
8	170	31
9	179	21
10	210	37
11	178	29
12	104	15
13	145	17
14	203	38
15	163	17
16	216	36
17	106	13
18	216	39
19	191	36
20	197	34

So, now we have so we have conducted the experiment and we have taken here the 20 students and then we have obtained, we have collected the data on the marks, which they have obtained and we also ask the student that how many hours the person has studied, right.

So, this is how we have collected the 20 observations on x_i and y_i. So, this is our setup now, this setup we will assume that this is going to follow the model $\beta_0 + \beta_1 X + \varepsilon$ and based on that we will try to find the values of β_0 and β_1 in the first step, and after that we will try to find out the value of σ^2.

(Refer Slide Time: 21:06)



Estimation of parameters in simple linear regression model:

Various methods of estimation can be used to determine the estimates of the parameters.

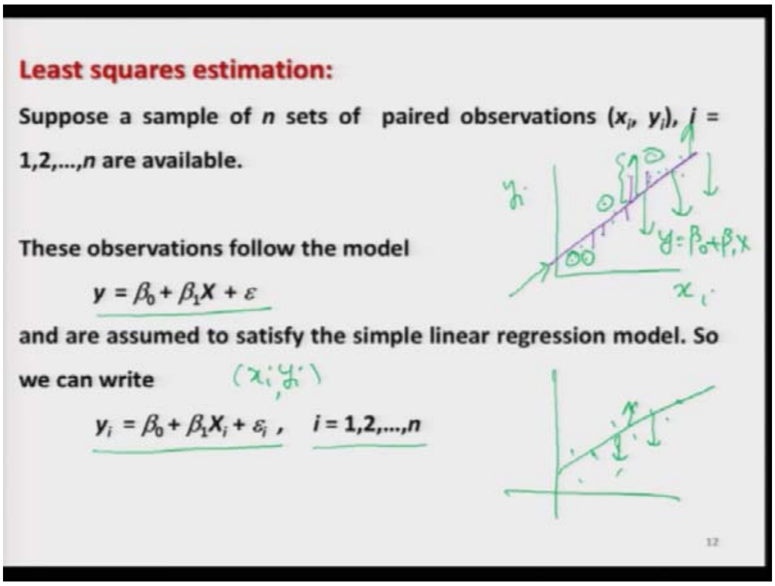
Among them, the least squares and maximum likelihood principles are the popular methods of estimation.

11

Now, in order to find out or in order to estimate the regression coefficient β0 and β1 are different types of estimation methods can be used, principle of least square, method of maximum likelihood and there are some other things also.

So, that we are not going to consider here, they are something like orthogonal regression, inverse regression there are several other techniques also. And among all these techniques the least square estimation and maximum likelihood estimation are the popular methods for estimation, right.

(Refer Slide Time: 21:48)



Least squares estimation:

Suppose a sample of *n* sets of paired observations (x_i, y_i), i = 1,2,...,*n* are available.

These observations follow the model

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

and are assumed to satisfy the simple linear regression model. So we can write

$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,2,\ldots,n$

y_i
$y = \beta_0 + \beta_1 x$
x_i
(x_i, y_i)
12

So, first we try to consider the least square estimation and our objective is this using the principle of least square we will try to find out the value of β_0 and β_1. What is your principle of least square? Suppose I have got the values of say here x_i and y_i and I try to plot this paired observation. So, suppose I am plotting them here like this.

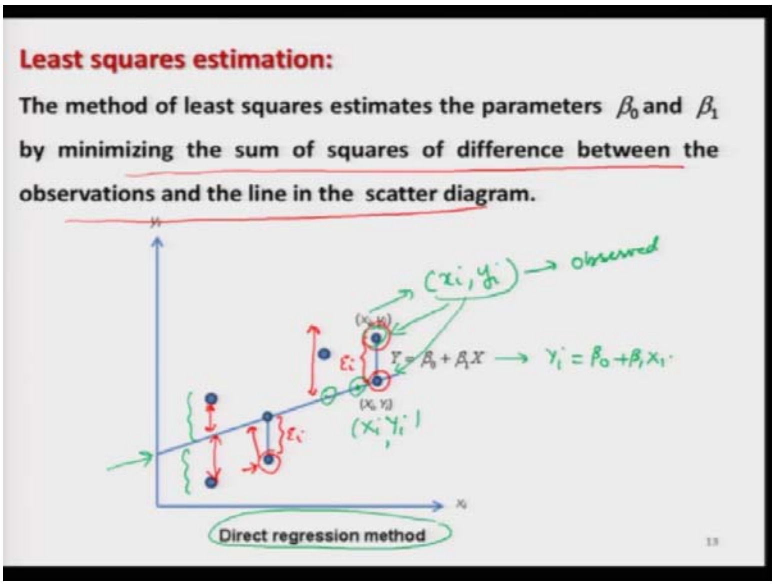
Now, what is my objective is this that I want to find here a straight line. So, the principle of least square says that please try to find out a line like this one in such a way which is passing through with the maximum number of points and in which the random error, which is happening here like this, which I am denoting by this vertical line this error is as minimum as possible, right. I will try to explain you in more detail as we move further.

So, now we assume that all these x_i and y_i's they are following the model $y = \beta_0 + \beta_1 X + \varepsilon$. So, this means, every pair of observation will also follow the same model, right. So, I can write for a given pair of observation x_i and y_i the simple linear regression model will also satisfy this equation and this equation is given here, that $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, right.

So, all this observation which you have written here they are going to follow the line $y = \beta_0 + \beta_1 X$ and this line, purple line will be your here $\beta_0 + \beta_1 X$. So, you can see here you are assuming this line purple line, but your observations are not exactly lying on the line some of them are lying on the line, but some of the observations are above the line and some observation are below the line, right.

So, this is what I meant that if you try to see that there is random variation in the values of y, and whatever is this random variation that is indicating the presence of ε_i's. And that is what I said that some points will be lying above the line and some points will be lying below the line; like this over here I can draw. So, some points are lying above the line and some points are lying below the line. So, in general I will say that the in general the average error is 0, right.

(Refer Slide Time: 24:50)



Least squares estimation:

The method of least squares estimates the parameters β_0 and β_1 by minimizing the sum of squares of difference between the observations and the line in the scatter diagram.

y_i
x_i
$(x_i, y_i) \rightarrow$ observed
(x_i, y_i)
ε_i
$Y = \beta_0 + \beta_1 X \longrightarrow Y_i = \beta_0 + \beta_1 X_i$
ε_i
(X_i, Y_i)
(X_i, Y_i)
Direct regression method
13

So, as I tell you as I explain you briefly now I try to explain it more formally, suppose this is your here line. Y = β0 + β1 capital X. So, all the observations, which are obtained they are assumed that they will follow this line. So, all those points which are lying on this line over here they are indicated by X_i and Y_i capital X_i, capital Y_i.

Now, in practice you are getting the observation which are not exactly lying on the line, but suppose let me take this observation. This is lying above the line, and this is the observation say x_i a small x_i small y_i which you have observed. So, I can say this observation small x_i and small y_i that is expected to lie on this line, right.

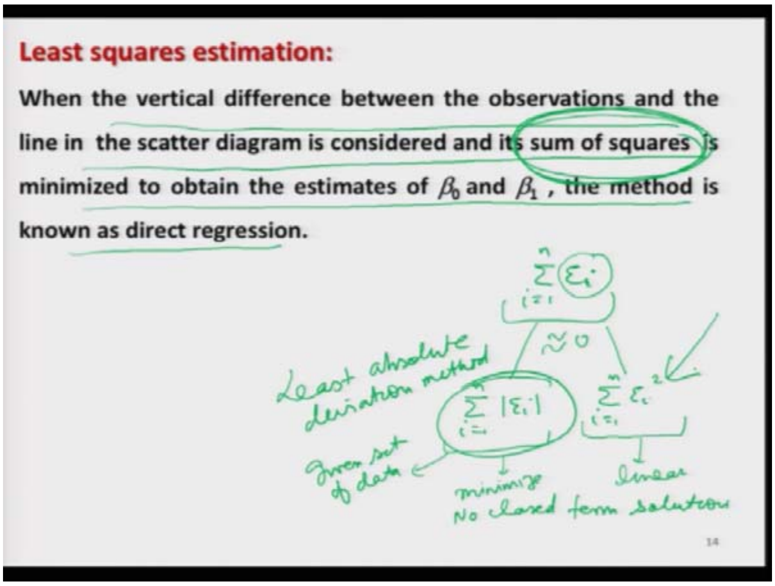
So, there is here some error here. Similarly, if you try to take here another observation here this observation is expected to lie on this line, but it is lying somewhere here so, there is some random error into it, right. So, these are your here ε_i. I will try to show you more clearly.

So, the principle of least square says that you try to find out the values of β_0 and β_1; such that the random errors are as minimum as possible and the lines and the line is passing through with the maximum number of points.

So, what we try to do? We try to minimize the sum of squares of the difference between the observation and the line; that means, this observation and this observation. We will try to minimize this distance and this distance is for here each and every observation.

So, and the values of β0 and β1 that you will obtain from there they will be called as least squares estimation. And in particularly when you are trying to minimize these errors in the vertical direction here, you can see here in this direction then this method is called as direct regression method that is alternative name. And you will see in the literature that there are different types of regression method, inverse regression, orthogonal regression etc.

(Refer Slide Time: 27:48)



Least squares estimation:

When the vertical difference between the observations and the line in the scatter diagram is considered and its sum of squares is minimized to obtain the estimates of β_0 and β_1, the method is known as direct regression.

$\sum_{i=1}^{n} \varepsilon_i$
≈ 0
Least absolute deviation method
$\sum_{i=1} |\varepsilon_i|$
$\sum_{i=1}^{n} \varepsilon_i^2$
Given set of data
minimize
No closed form solution
linear
14

So, but here we are going to concentrate on the this vertical difference or the direct regression method, right. So, now I can reformulate it in a very simple sentence that, when the vertical difference between the observations and the line in the scatter diagram is considered and the sum of squares is minimized to obtain the estimates of β0 and β1, the method is known as direct regression.

At this moment there comes a question, that why I am writing here sum of squares? Why cannot I simply minimize the sum of this random errors i goes from 1 to n? Here there is a problem there are actually several problem. First problem is this sometime this ε_i is going to take a positive value and sometime it is going to take a negative value when it is measured with respect to the line.

So, it is possible that the sum of all these random errors might be very close to 0 and that may indicate that the random variation in the data is very very less, right. So, that would indicate that as if the data is containing very small amount of error and it will become difficult to justify it, if this is not really indicating the correct scenario because the data has error.

So, what we try to do? That we have here two options that I can consider here the absolute values because I want to get rid of the sign or the second option here is $\sum_{i=1}^{n} \varepsilon_i^2$.

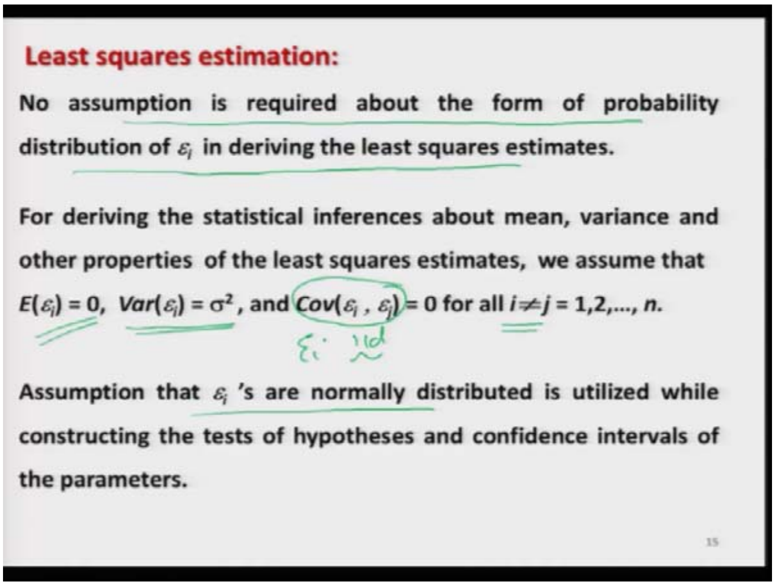
Now, in the next slide I will try to show you that how to minimize it. So, when we are trying to minimize it then using the principle of maxima and minima if you try to differentiate this part this will give you a linear equation, which can easily be solved, right.

But in this case the this error can also be minimized, but in this case there is no closed form solution, right. So, that is also considered, but in this case the solution can be obtained only for a given set of data. For a given set of data you try to use any optimization technique and try to find out the value of β_0 and β_1.

So, depending on the optimization method that you will use in this case the values of the β_0 and β_1 they are going to be different and this method is called as least absolute deviation method, right.

So, we are not going to consider here although this is also popular and I will try to take up that issue at a later stage when we want to choose some important variable, but not now at this moment we are going to concentrate on the sum of squares, ok.

(Refer Slide Time: 31:17)



Least squares estimation:

No assumption is required about the form of probability distribution of ε_i in deriving the least squares estimates.

For deriving the statistical inferences about mean, variance and other properties of the least squares estimates, we assume that $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$, and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j = 1, 2, \ldots, n$.

$\varepsilon_i \overset{iid}{\sim}$

Assumption that ε_i 's are normally distributed is utilized while constructing the tests of hypotheses and confidence intervals of the parameters.

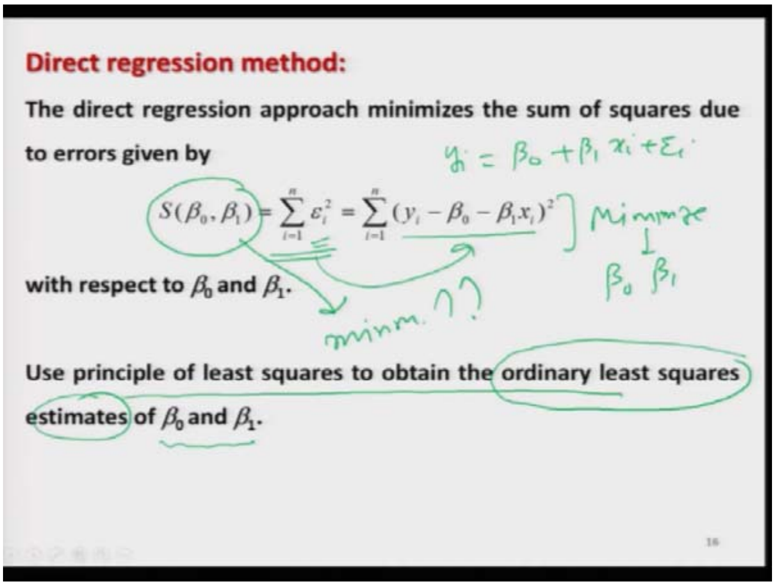
15

So, now what we try to do? That we are not assuming any assumption about the form of the probability distribution of this ε i and we are simply assuming that ε_i's are iid, iid means they have been identically they are identically and independently distributed.

Since they are independent so, for the covariance between ε_i and ε_j for i not equal to j will be 0 and we already have assumed that expected value of $\varepsilon_i = 0$ and variance of ε_i is equal to σ^2. Actually these assumptions on the properties of random variable will not be required to estimate the parameters, but they will be required when we try to establish the properties of the estimator, right.

And as I said the assumption that we need this ε_i's to be for example, normally distributed that will be utilized when we are trying to construct the test of hypothesis and confidence interval for the parameters, right.

(Refer Slide Time: 32:27)



Direct regression method:

The direct regression approach minimizes the sum of squares due to errors given by

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$S(\beta_0,\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \Bigg] \text{Minimize} \atop \beta_0 \ \beta_1$$

with respect to β_0 and β_1.

minm. ??

Use principle of least squares to obtain the ordinary least squares estimates of β_0 and β_1.

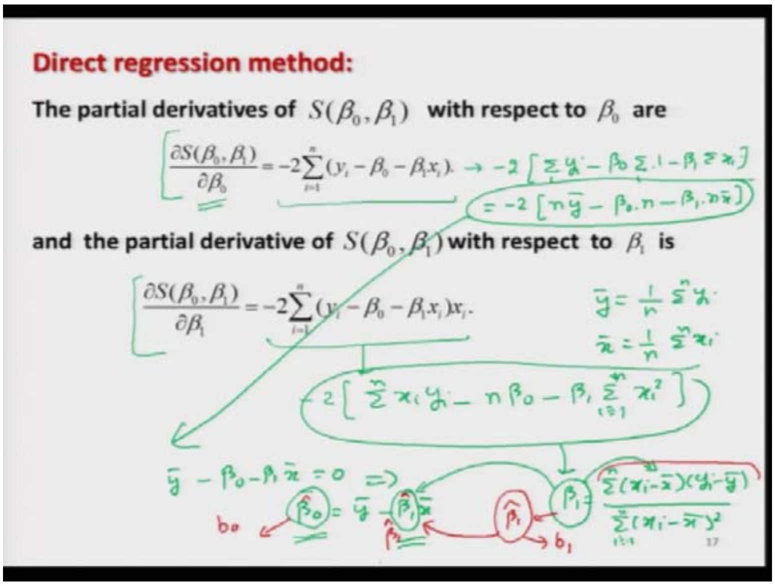
16

So, now let me try to first explain you the method of direct regression. So, in the direct regression what I am trying to say? That you try to consider the sum of squares of the random errors which is $\sum_{i=1}^{n} \varepsilon_i^2$. And since you have written that $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

So, this ε_i is y_i minus β_0 minus β_1 x_i and so, I can rewrite this thing in terms of β_0 and β_1. Now, after this you simply have to use some optimization principle so, that you can minimize this function and obtain the values of β_0 and β_1. So, one of the most simple principle to optimize that we know is the principle of maxima and minima.

So, employing the principle of maxima and minima and using the principle of least squares that we will try to obtain the estimates of β_0 and β_1. So, now what we do? We try to consider this function S β_0, β_1 and we try to differentiate it partially differentiate it with respect to the parameters β_0 and β_1.

(Refer Slide Time: 33:40)



Direct regression method:

The partial derivatives of $S(\beta_0, \beta_1)$ with respect to β_0 are

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i). \rightarrow -2\left[\sum y_i - \beta_0 \sum 1 - \beta_1 \sum x_i\right]$$

$$= -2\left[n\bar{y} - \beta_0 . n - \beta_1 . n\bar{x}\right]$$

and the partial derivative of $S(\beta_0, \beta_1)$ with respect to β_1 is

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i.$$

$$\bar{y} = \frac{1}{n}\sum^{n} y_i$$

$$\bar{x} = \frac{1}{n}\sum^{n} x_i$$

$$2\left[\sum^{n} x_i y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i^2\right]$$

$$\bar{y} - \beta_0 - \beta_1\bar{x} = 0 \Rightarrow \beta_0 = \bar{y} - \beta_1\bar{x}$$

$$b_0 \leftarrow \hat{\beta}_0 \qquad \hat{\beta}_1 \rightarrow b_1$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

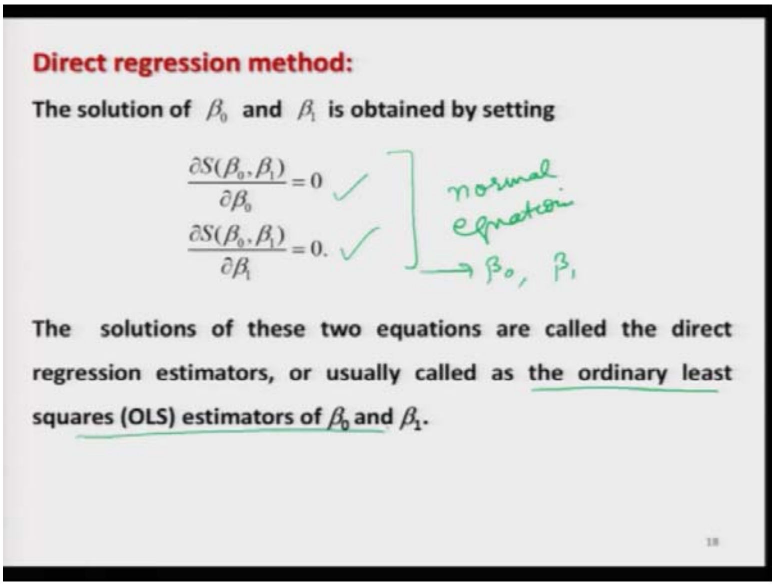
17

So, when I try to partially differentiate the function S with respect to β0 we get here this equation, and when we try to differentiate this function S with respect to β1 then we get here this equation. Now, what we have to do? We simply have to solve these two equation and we have to obtain the value of β0 and β1.

For example, if you try to see here that is very simple, but if I try to show you here this equation can be written as here $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$, right. So, this can be written as minus 2 and $\sum_{i=1}^{n} y_i = n\bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ that is the sample mean of y_i.

And similarly we can define the sample mean of the values of x_i here like this, right. So, this I can write down here say $n\bar{y}$ - β0n - β1n$\bar{x}$. And similarly this equation can be solved as summation here minus 2 and this will become here $\sum_{i=1}^{n} x_i y_i$ - n β0 -β1 $\sum_{i=1}^{n} x_i^2$, right. So, in case if you try to solve this equation and this equation then, what we obtain here?

(Refer Slide Time: 35:41)



Direct regression method:

The solution of β_0 and β_1 is obtained by setting

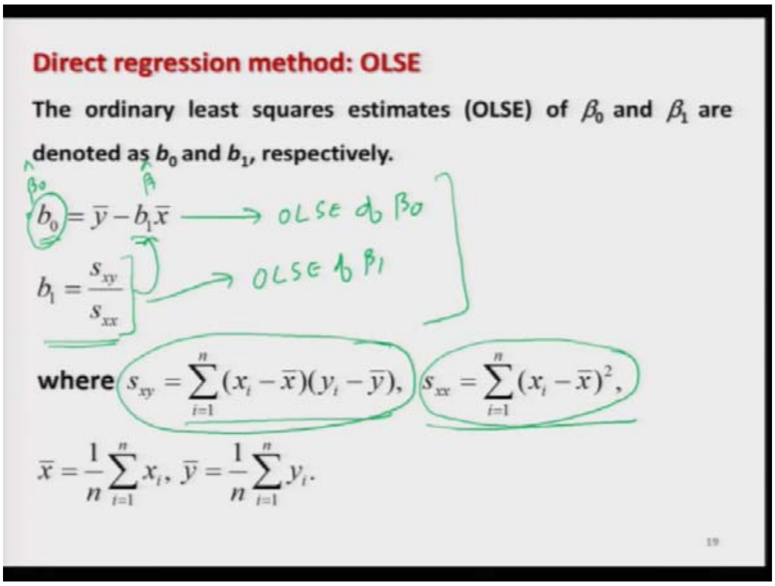
$$\left.\begin{aligned}\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} &= 0 \checkmark\\ \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} &= 0. \checkmark\end{aligned}\right] \text{normal equation} \longrightarrow \beta_0,\ \beta_1$$

The solutions of these two equations are called the direct regression estimators, or usually called as the ordinary least squares (OLS) estimators of β_0 and β_1.

18

So, if you simply try to put these two partial derivatives equal to 0, that is the principle of maxima minima and try to solve it you will get the values of β0 and β1. So, this is these two equations are called as normal equations. And whatever you are going to obtain after solving this thing these will be two values of β0 and β1, and they will be called as ordinary least square estimates of β0 and β1.

(Refer Slide Time: 36:24)



Direct regression method: OLSE

The ordinary least squares estimates (OLSE) of β_0 and β_1 are denoted as $\overset{\hat{\beta}}{b_0}$ and b_1, respectively.

$$\overset{\hat{\beta}_0}{\boxed{b_0}} = \bar{y} - \overset{\hat{\beta}}{b_1}\bar{x} \longrightarrow \text{OLSE of } \beta_0$$

$$b_1 = \frac{s_{xy}}{s_{xx}} \longrightarrow \text{OLSE of } \beta_1$$

where $s_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$, $s_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2$,

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$.

19

So, means if you just try to do a 1 minute algebra and can if you try to solve it you will get the value of β0 as here like this, $\overline{y} - \beta_1 \overline{x}$ and actually that you can see from here also in this equation if you try to substitute this equation to be equal to 0, then you get here n $\overline{y}$ or n will get cancel out $\overline{y}$ - β0 - β1 times $\overline{x} = 0$.

So, this implies that β0 will be $\overline{y}$ - β1 $\overline{x}$. So, definitely if you want to know the value of β0 then you need to know the value of β1. So, when I try to solve this second equation which is here from here you will get the β1 $= \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$.

So, you can see here that these values are depending only on your observations xi and yi that you have obtained through the experiment. So, now, this β0 is going to be known to us only when β1 is known to us. So, what we propose? That, first we try to obtain the β1 using this expression and then we try to substitute it, this value over here in this expression.

But now, let me try to explain you the symbolic problem. You are trying to represent this β1 this quantity which you have obtained here by β1, but β1 is the population parameter that is unknown to us, but this is the value which you have obtained here this can be obtained on the basis of given sample.

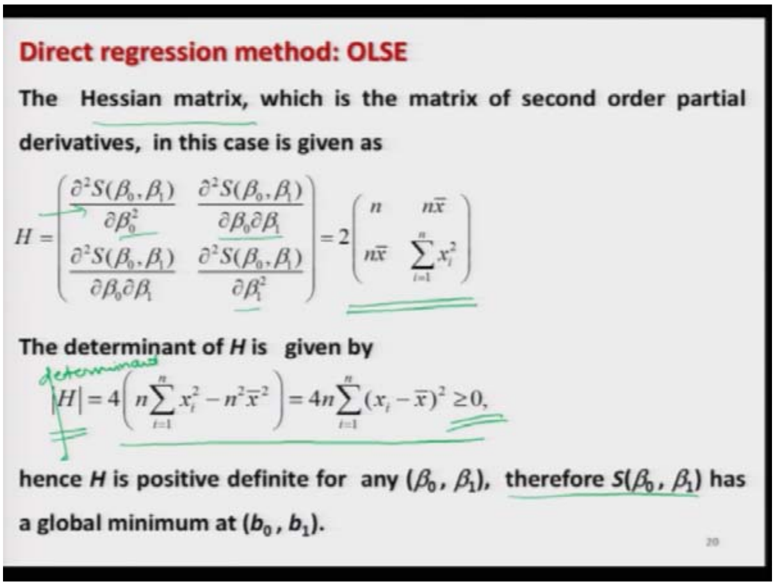
So, this is essentially an estimator of β1 and an estimator of a parameter in statistics is denoted by say $\hat{\beta}_1$ or $\tilde{\beta}_1$ you have to indicate it something. So, when you are trying to put the hat on this parameter that is indicating that the value with the hat can be estimated on the basis of given sample of data.

So, now you try to replace this $\hat{\beta}_1$ here and you will get here like this. So, we are denoting this $\hat{\beta}_1$ here as say here b1 and this will now become this will now become β0 hat and this β0 hat is indicated by b0. So, this is what exactly I have written here, right. So, this b0 is something like $\hat{\beta}_0$ and b1 is something like $\hat{\beta}_1$.

So, the value of $\hat{\beta}_1$ is obtained here $\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$, which I have symbolically denoted by sxy and sxx. So, $s_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}), \quad s_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2$. So, you can look that these two quantities are somewhat similar to the variance except that there is no divisor, right.

So, the value of β1 can be obtained by sxy upon sxx and once you obtain this value you can substitute in the value of β0 and you can obtain the estimator of β0, which is indicated here by b0, right. So, this is called as least square estimate. So, this is essentially the ordinary least square estimate of β1 and this b0 is the ordinary least square estimate of β0, right.

(Refer Slide Time: 40:59)



Direct regression method: OLSE

The Hessian matrix, which is the matrix of second order partial derivatives, in this case is given as

$$H=\begin{pmatrix}\frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0^2} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0\partial\beta_1}\\ \frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0\partial\beta_1} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_1^2}\end{pmatrix}=2\begin{pmatrix} n & n\overline{x}\\ n\overline{x} & \sum_{i=1}^{n}x_i^2\end{pmatrix}$$

The determinant of *H* is given by

determinant

$$|H|=4\left(n\sum_{i=1}^{n}x_i^2-n^2\overline{x}^2\right)=4n\sum_{i=1}^{n}(x_i-\overline{x})^2\geq 0,$$

hence *H* is positive definite for any (β_0,β_1), therefore $S(\beta_0,\beta_1)$ has a global minimum at (b_0,b_1).

20

So, what you have to; what you have obtained here, that you have obtained the values of β0 and β1, but now there is a doubt whether the value which you have obtained here for β0 and β1 are they really going to make this sum of square minimum or maximum that we do not know.

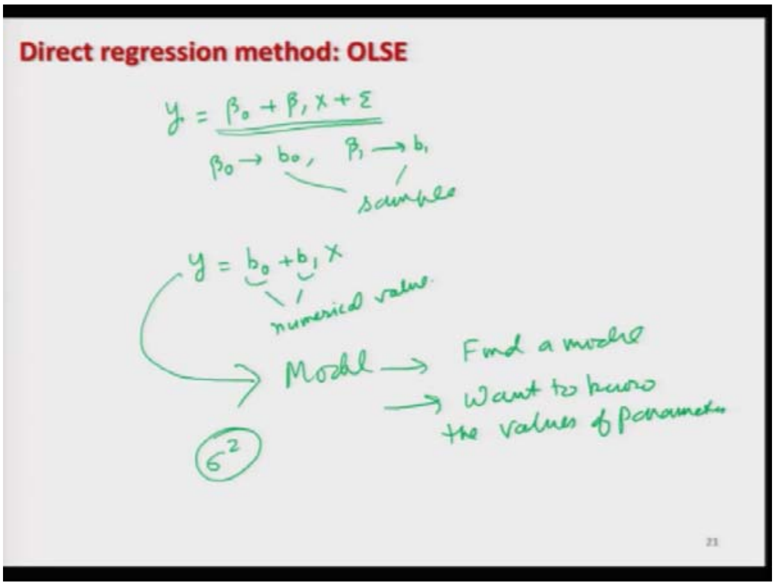
So, what we try to do? The principle of maxima and minima suggest that we have to take the second order derivative, and we have to check whether these values are really maximizing or minimizing the sum of squares. So, what we try to do?

We obtain here the Hessian matrix, so this is obtained here by the second order partial derivative of the as function with respect to β0, β1 and with respect to β0 and β1 that is a very simple algebra if you simply differentiate it you will get here this thing.

And then, in order to check whether this quantity is going to be maximized is a positive or negative we try to obtain its determinant and this determinant is obtained here, say here absolute said determinant value of H, remember this thing this is determinant not absolute, right. And this is here like this and this comes out to be $4n\sum_{i=1}^{n}(x_i-\overline{x})^2$ which is always greater than 0.

So, hence this matrix here capital H is a positive definite matrix and therefore, this function S(β0, β1) has a global minima at the values b0 and b1. So, now I can be confident that the value which you have obtained here b0 and b1 here they are really going to minimize the sum of squares, right.

(Refer Slide Time: 43:06)



Direct regression method: OLSE

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

$\beta_0 \to b_0, \quad \beta_1 \to b_1$

sample

$y = b_0 + b_1 X$

numerical value

Model → Find a model

→ Want to know the values of parameters

σ^2

21

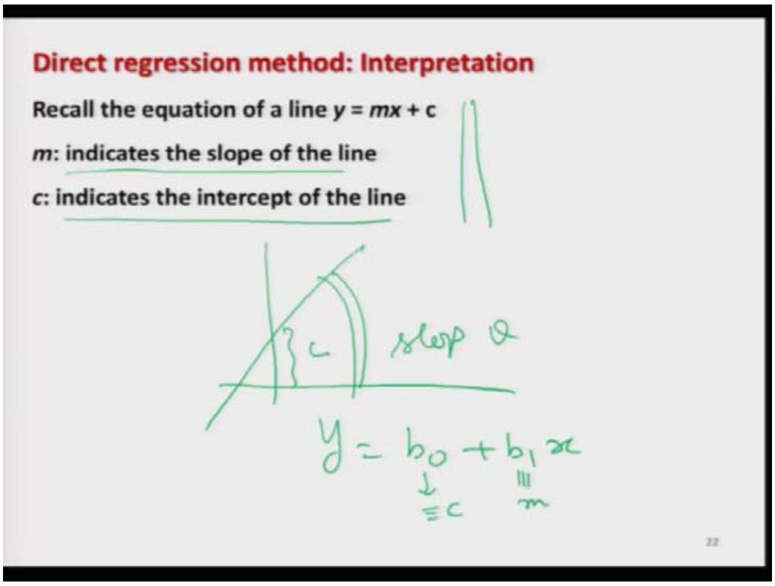
So, now if you observe what you have obtained, you started with the model y = β0 + β1 X + ε, where β0 and β1 are the population value so, they are unknown to us. So, what you have done, you have estimated the value of β_0 by b_0 and the value of β_1 by b_1. So, b_0 and b_1 are the values of β_0 and β_1 respectively, which can be obtained on the basis of given sample of data.

So, now what do I; what I can do? Now, I have obtained the unknown values of β_0 and β_1 on the basis of given a sample of data, and I can rewrite my model as $b_0 + b_1$ X, b_0 and b_1 X they are going to be some numerical value. It can be something like y = 2 + 3x. So, now you have obtained this equation and in simple way I can say you have obtained a model.

Now, you have a mathematical equation, which is going to represent the entire process and that will give you each and every information. So, you can see here when we say that we are going to find a model this is equivalent to saying that we want to know the values of parameters.

Although, one parameter here it is σ^2 which is still unknown to us and we will consider how to estimate it, right. So, this is what you mean by finding out a models, right. So, modelling is equivalent to the estimation of parameters in the assumed form of the model and you also have assumed of that the form of the model is a linear equation, ok.

(Refer Slide Time: 45:14)



Direct regression method: Interpretation

Recall the equation of a line *y* = *mx* + c

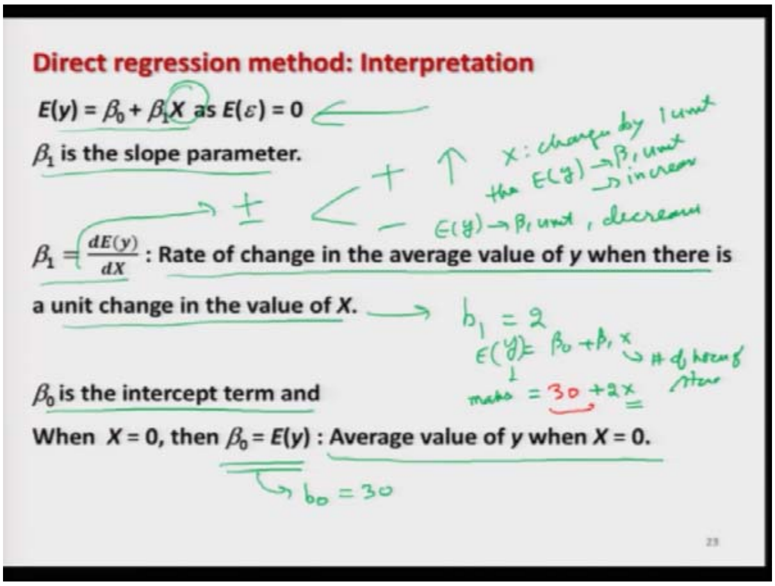
m: indicates the slope of the line

c: indicates the intercept of the line

Now, what are the interpretation of this b_0 and b_1? So, for that you try to recall the equation y = mx + c, what was you there m? m was indicating the slope of the line and c was indicating the intercept term of the line something like this is your here c and this here is your slope say θ then m = tan θ, right.

So, the same interpretation can be extended to this line also because, you have obtained here a line something like b0 + b1 x. So, this is comparable with mx equal to c where b0 is equivalent to here c and b1 is equivalent to here m.

(Refer Slide Time: 45:59)



Direct regression method: Interpretation

$E(y) = \beta_0 + \beta_1 X$ as $E(\varepsilon) = 0$

β_1 is the slope parameter.

$\beta_1 = \frac{dE(y)}{dX}$: Rate of change in the average value of *y* when there is a unit change in the value of *X*.

β_0 is the intercept term and

When $X = 0$, then $\beta_0 = E(y)$: Average value of *y* when $X = 0$.

23

So, now what are the interpretation? So, since you have written the model as expected value of y = β0 + β1 X something like y = mx + c. So, β1 is the slope parameter, and if you try to see β1 is the differential coefficient of expected value of i y with respect to x; d of expected value of y with respect to dx. What is this? This is simply the rate of change in the average value of y when there is a unit change in the value of X.

So, the value of β1 is indicating the rate of change in the average value of the response variable when the input variable has been changed by one unit. So, same is the interpretation for b1 also. So, if you get here for example, say b1 equal to 2 that means, if you try to change the value of x by 1 unit then there will be change of 2 units in the average value of y.

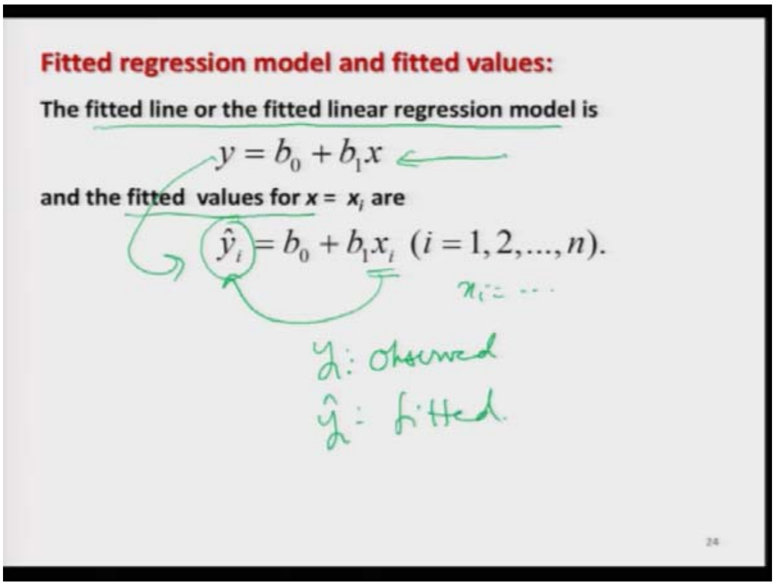
So, you had a model here something like y = say here β0 + β1 X that was expected and your X was number of hours of study, and y was marks obtained in the examination. So, now if you obtain here the model here plus some 2 times X so, that is indicating that if a student studies for 1 hour more, then the marks are going to be increased by 2 units.

And definitely this also will have a sign say plus minus. So, plus is going to indicate that the rate of change is positive that means, increasing this is increasing that means the if X changes by 1 unit then, expected value of y will change by this β_1 unit and this will increase; whereas, if this coefficient is negative that means, if X changes by 1 unit then expected value of y changes by β 1 unit, but it decreases, right.

And similarly β_0 is your intercept term that means, if you try to substitute X equal to 0 in this model. Then, you get here β_0 is equal to expected value of y. So, this is what the intercept term is going to indicate the average value of y when explanatory variable is 0. So, suppose if and the same interpretation goes for the estimated value b_0 also.

Suppose, if you get b_0 = 3 or say 30 then your model becomes here 30 plus twice of X. So, you are trying to say that if a student does not study at all. Even then the student will get 30 marks and it is possible because, the student is attending the classes also and possibly he has learnt in the class also, right.

(Refer Slide Time: 49:27)



Fitted regression model and fitted values:

The fitted line or the fitted linear regression model is

$$y = b_0 + b_1 x$$

and the fitted values for $x = x_i$ are

$$\hat{y}_i = b_0 + b_1 x_i \ (i = 1, 2, \ldots, n).$$

x_i = ...

y: observed

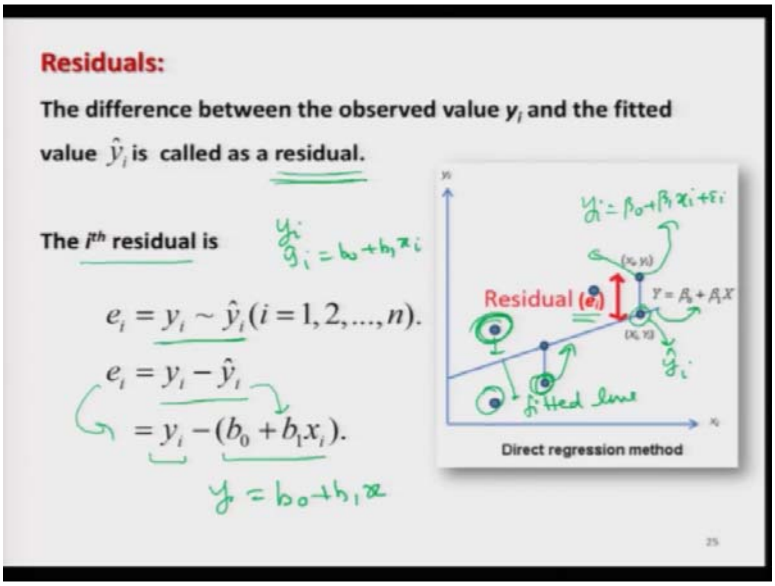
$\hat{y}$: fitted

24

So, the model which we have obtained after substituting the estimated value of β_0 and β_1 this is called as fitted model or this is called as fitted linear regression model and that is our objective that we want to find, this is our real model actually. So, when we say find a model we are trying to say that find the fitted model.

And in this model if you try to substitute the values of x_i for example, you have obtained in your example the values of x_i that how many number of hours a student is studying, if you try to substitute these values in the fitted model then the corresponding values which you will obtain for y they are called the fitted values and they are indicated by $\hat{y}_i$. So, y_i is the observed value and $\hat{y}_i$ is the fitted value, right.

(Refer Slide Time: 50:31)



Residuals:

The difference between the observed value y_i and the fitted value $\hat{y}_i$ is called as a residual.

The i^{th} residual is

y_i

$\hat{y}_i = b_0 + b_1 x_i$

$$e_i = y_i \sim \hat{y}_i (i = 1, 2, ..., n).$$

$$e_i = y_i - \hat{y}_i$$

$$= y_i - (b_0 + b_1 x_i).$$

$y = b_0 + b_1 x$

25

And if you try to plot it on your this scatter diagram that we considered earlier. So, if you remember this was your here the true model, but these observations they are going to follow something like $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. So, you can see here that now based on the fitted model you also have obtained the value of $\hat{y}_i$.

So, this $\hat{y}_i$ is going to lie here because, y_i hat is the value, which is now lying on the fitted line, right. And the difference between the value which you had observed and the value which you have obtained on the basis of fitted model for y this is called as residual, right. So, for a given observation so, this i-th observation, what I can do? I have observed value y_i and then I have fitted the model here $y_i = b_0 + b_1$ x.

So, I substitute here the value of x equal to x_i and I try to obtain here the fitted value and whatever is the difference between y_i and $\hat{y}_i$ this is the i-th residual. So obviously, you can see that the model is going to be good if these e i's are as small as possible and ideally if this is, if they are 0. So, that all the points which you have observed here they must lie on the line, right.

So, this is the concept of residuals. So, usually this is the difference between y_i and $\hat{y}_i$, but in practice we try to consider y_i minus $\hat{y}_i$ at least in this lecture, I am going to consider this definition the difference between y_i and $\hat{y}_i$ is indicated by y_i minus $\hat{y}_i$ I am not subtracting like y_i hat minus y. So, the value of the e_i can be obtained by y_i minus the $\hat{y}_i$ which is obtained by the fitted model $b_0 + b_1 x_i$, right.

So, now you can see here now, you have obtained here the model y equal to $b_0 + b_1x$ you have obtained the values of $\hat{y}$, if the model would have been following the model which you have obtained on the basis of given sample of data, right. So, now you have done you are done you have obtained the model and then you also have obtained the amount of error a sort of an idea about the amount of error, which is occurring in your data set because, you assume the model now you have obtained the fitted model.

So, there is some difference between the observed and the true values of y. And based on that we can do many things we can see whether my model is good fitted or bad fitted if those e $_i$i's are 0, then my model is the model which is best fitted on the basis of given sample of data but, in practice it will never happen; there will be some values and our objective is that as we try to make it as minimum as possible.

So, you can see here now in this lecture I have taken very small thing, very simple thing and I have tried to give you the idea that what is the interpretation of slope parameter how it is estimated on the basis of sample, how what is the interpretation of β_1, how it is estimated on the basis of given of sample, and then how is the difference between the observed and true values are obtained.

Now, this concept will definitely help you when we try to consider the multiple linear regression model and there we will introduce more notations more concepts also, but my idea in this lecture was to give you a basics of this regression modeling. So, you are familiar with those terminology so, it will become very easy for me to explain you the concept of multiple linear regression model.

But, the question here is how can we estimate all these things on the basis of R software? So, in the next time, I will try to take the same example which I have considered here and then I will try to show you that how these values can be obtained in the R software, but then the difference will be once you see the output you know what are you getting and how these values have been obtained.

Earlier you are simply clicking, clicking, clicking on the software and you will getting some values you were trying to possibly do some hit and trials for the interpretations and that is how you are possibly working means, I expect that you are not doing this thing, but this is really happening with many people and that is my objective in this course that I want to make people learn that what is the correct way. So, you try to revise this concept, try to think about it and I will see you in the next lecture, till then good bye.