

## **Lecture – 33**

### **Fitting of Linear Models: Least Squares Method – One Variable**

Welcome to that lecture on the course descriptive statistics with R software, in this lecture we are going to start with a new topic  
Refer Slide Time: (00:22)

# **Descriptive Statistics With R Software**

## **Fitting of Linear Models**

**::**

## **Least Squares Method – One Variable**

**Shalabh**

**Department of Mathematics and Statistics**

**Indian Institute of Technology Kanpur**

fitting of linear models. So, you see what happens, whenever we have a sample of data, then the first information is obtained from the graphical tool and suppose we have data on two variables and both the variables are associated, they are not independent, then definitely this observation will have some correlation structure. So, first step will be to create a plot like as a graphic plot, a smooth a scatter plot and so on, those plots are going to give us an idea, there is an association present in the data. Now, after this we try to use different types of tools for example, correlation coefficient to quantify the degree of Association or the degree of linear relationship. Now, the final question which is remaining is, can we find out the mathematical relationship between the two variables or can we find a statistical model between the two variables and if yes, how to do it? So, that is the question which we are going to entertain in this topic fitting of linear models. Now, we are assuming that there are two variables or there can be more than two variables and there exists some relationship between those variables Right! and whenever there is a relationship, there are going to be two types of variable one, output variable, that mean the the variable on which, we obtain the values of the output and second aspect will be input variable. So, whatever are the values which are given to the so called input variable based on that we will have an output, for example, suppose if I say that whenever we are doing some agriculture, then the weight of a crop that depends on the quantity of fertilizer, that in case if I try to increase or decrease the quantity of fertilizer in the field, then up to a certain extent the craft will increase or decrease, Yeah! obviously if you try to increase it more than the crop will get burnt up. So, this is a case where we can see that relationship between the yield of the crop, the quantity and the quantity of fertilizer. Similarly, if I try to extend this relationship, then we know from our experience that the outcome of a variable in this case for example, the field of the crop, it does not depend only on one variable the quantity of fertilizer but it depends on

other variables also, quantity of fertilizer, temperature, rainfall, irrigation, moisture and so on. So, now I have given you two situations where the outcome is going to be dependent on one variable and on more than one variables. So, now how to handle this situation this is what we are going to understand in this lecture,

Refer Slide Time: (04:03)

**Relationship Between Variables**

**Relationship exists between two variables.**

**Output of a variable is affected by one or more than one variables.**

**Example:**

- **Yield of crop increases with an increase in quantity of fertilizer.**
- **Speed of electric fan (rotations per minute) increases as voltage increases.**
- **People drink more water as weather temperature increases.**
- **Yield of a crop depends upon other variables like quantity of fertilizer, rainfall, weather temperature, irrigation etc.**

we assume that a relationship exists between two variables or this can be generalized, that the relationship exists between R variable and more than one variables and so on, that we will see later on and in this type of relationship, we assume that the outcome of variable is affected by one or more than one variables for example, the example which we have just considered that the yield of a crop increases with an increase in the quantity of fertilizer, Right! Similarly, if you try to see, incase if you try to observe the phenomena of an electric fan. So, whenever we try to increase the speed, how it is done that is controlled by a switch and we try to control the switch from position number one to position number two, position number two to position number three and in this process, what is happening, inside the switch, the quantity of current that is increased. So, I can say that when I am trying to increase the position of the switches from one to two, to three and so on, then automatically inside the switch the amount of current flowing in the circuit increases and finally the outcome of the fan which is the RPM, that is rotation per minutes increases or in simple words the speed of the fan, say increases. Similarly, in another example, we have seen that as the temperature of the weather increases, then the quantity of water consumed is more you can see that will consume more water during summer than winter, Right! So, in this type of example, where I am saying

that the speed of an electric fan which is measured by the rotations per minute, because increases as the voltage or current increases, people drink more water as the weather temperature increases, and similarly the same example that the yield of a crop depends on other variables also, like as quantity of fertilizer, rainfall, weather, temperature, irrigation etc., Right!

Refer Slide Time: (06:26)

**Relationship Between Variables**

Relationships are expressed through models.

**Model:**  
Relationship among the variables depicting the phenomenon.

Relationship is characterized by variables and parameters.

What type of relationships ?

Relationship can be linear or nonlinear.

So, now we are assuming that such type of relationships, can be expressed through models and model is a very fancy word, in nowadays everybody wants to find out a model among the variables and so how to do it, but first question comes what is a model, this model is only a, so accord of mathematical or statistical relationship among the variables and this relationship is in such a way such that it is representing or depicting the phenomena that whatever is happening, this is indicated by the mathematical functioning of the mathematical relationship or the statistical relationship. So, when we talk of this modeling, then modeling and relationship among the variables they are the same thing and in such a case, the relationship is characterized by two things one is variables and say another is parameters. So, the first question comes that what is the difference between variables and parameters. So, this one will try to address in a very simple language through a very simple example, but before going further back we clarify here that what type of relationships can exist, then there can be different type of relationship and broadly I can classify them into two parts linear and nonlinear. Now, in this course we are going to entertain only the linear relationships.

Refer Slide Time: (08:17)

## Input and Output Variables

Usually any phenomenon has two types of variables

- input variables and
- output variables.

• Marks depend upon number of hours a student studies ✓

or

~~Number of hours of study depends upon the marks obtained by student.~~ ✗

• Yield of a crop depends upon the rainfall and weather temperature ✓

or

~~Rainfall and weather temperature depends upon yield of crop.~~ ✗

Now, I come back to my means earlier issued, that how to start. So, whenever there is a phenomena or whenever there is something is happening, then in that phenomena is you have to observe that usually there will be two types of variables or the variables can be divided into two categories, one category is input variables and another category is output variables. Now, the first question comes that how to pick a call or how to decide that which of the variable is an input variable and which of the variable is an output variable. So, this I can explain you with the two simple example for example, we from our experience that whenever a student is studying more usually, he or she will get more marks. So, now I have here two variables, one is the marks in the examination and second is the number of hours a student studies. So, now in this situation I will try to ask two question that there are two possibilities, one possibility is that I can assume that marks depends on the number of hour to studies and second option will be that the number of hours studied depends on the marks obtained. Now, in such situation there is no mathematical rule or statistical rule which can explain you, but this is only your experience with the data, the information about the phenomena that is going to help you in taking a call or in taking a decision that which is affecting what. So, in this case, we have two options as I have given in the slides that the marks depend upon the number of hours a student studies or the second option is the number of hours of study depends upon the marks obtained by the student, Right! Now, each one is correct this is correct and this statement is wrong and similarly, if I try to take another example, that the end of a crop depends on the quantity of fertilizer and temperature of weather or the reverse happening, that the quantity of fertilizer or the temperature of that depends on the yield. So, in this case second option is not possible and we know

only from our experience that weather, temperature and the quantity of fertilizer, both are going to affect the amount of yield up to a certain extent .So, in this case the quantity of fertilizer and the temperature of weather, they become the input variables and the yield of the crop becomes an output variable. So, in this case if you try to see we have two options which are written on the slide that the yield of a crop depends upon the rainfall and weather, temperature or the second option is that the rainfall and this is either temperature depends upon the field of the crop. So, obviously the four sentence is correct and second sentences wrong, this is wrong and similarly in the first case this was wrong, Right! So, this is how we try to decide in a phenomenon that what is going to be an input variable and what is going to be an output variable in any given situation. Now the next question is that whenever we are trying to write down a model, model is essentially going to be a mathematical equation, Yes, the statistical concept will be used to obtain that mathematical equation but in that equation, there will be two components, one is variable and another this parameter. So, I will try to take here a very simple example to explain you, that what is the difference between the two and what is the role of variable and parameters in a model. Now, I am going to take an example of the equation of a simple linear line which you have possibly studied in class 10, 11 or 12,  $y$  equal to  $mx$  plus  $c$ , you see  $y$  equal to  $mx$  plus  $c$  has two types of component, one category I can define as  $x$  and  $y$ , and another category is  $m$  and  $c$ , out of these two categories, one set of quantities is a variable and another set of quantities is a parameter. Now, how to take a call, how to decide what is what. So, let me try to explain you here.

Refer Slide Time: (13:30)

**Variables and Parameters**

**Example**

**Equation of a straight line**

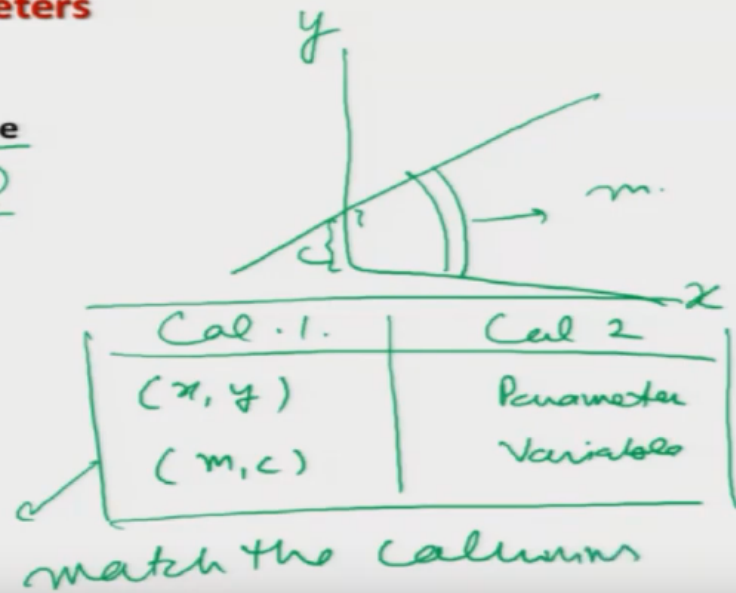
$$y = mx + c$$

**$c$  : Intercept term**

**$m$  : Slope of line**

**$x$  : Values on  $x$  - axis**

**$y$  : Values on  $y$  - axis**



Col. 1.	Col. 2
$(x, y)$	Parameter
$(m, c)$	Variables

match the columns

So, we know that the equation of a straight line is given by  $y$  equal to  $mx$  plus  $c$ , where your  $c$  is interceptive, and this line suppose it looks like this. So, here  $x$  is the  $x$ -axis and this  $x$  is going to indicate the values on the  $x$  axis, and this is my here  $y$ -axis and here this  $y$  is going to denote here the values on  $y$ -axis, and this quantity here is see here,  $c$ . So,  $c$  is going to be an intercept term, Right! and this angle that is going to be represented by  $m$  in terms of intern of the angle the trigonometric function. So, this is how we denote this equation  $y$  equal to  $mx$  plus  $c$ . Now, in this equation, if you see there are two options which I can explain you here, one is here  $x$  and  $y$  and second here is  $m$  and  $c$  and now out of this one of them is parameter, one of the set is representing parameter and another set is representing variable. Now, means I can give you a simple theory, that please match these two columns. So, column one and column two that is the simple type of question that you have solved that please match, match the columns. So, now let us try to understand through an example,

Refer Slide Time: (15:04)

**Variables and Parameters**

**Example**

**Option 1:** Knowing the values of  $(x, y)$ , say  $x = 4, y = 2$ , can we know all the information about the line?

For example,  $2 = 4m + c$

$y = mx + c$   
 $2 = 4m + c$

**Option 2:** Knowing the values of  $(m, c)$ , say  $m = 5, c = 6$ , can we know all the information about the line?

For example,  $y = 5x + 6$

**Option 1** : **Incorrect**

**Option 2** : **Correct**

that how you can do it, suppose I give you first option as that you know the value of  $x$  and  $y$ , suppose I know the value of  $x$  and  $y$  to be say  $x$  equal to 4 and  $y$  equal to 2, then my question is can we know the entire information or all the information about the line. So, in this case your  $y$  equal to  $mx$  plus  $c$  will become, say here  $y$  equal to 2,  $x$  equal to 4 and then  $m$  plus  $c$ . Now, looking at this line, do you think that you have the entire information about the line, certainly not. I just know that there is a point  $x$  equal to 4 and  $y$  equal to 2. Now, the second option is this that instead of having the values of  $x$  and  $y$ , I know the value of  $m$  and  $c$ , and suppose  $m$  is equal to 5 and  $c$  is equal to 6, in this case, the equation becomes here  $y$  is equal to  $5x$  plus 6. Now, my question is do we know the entire information about this line or can we



really know by looking at this equation all the information contained in this equation about the line, the answer is yes, if you wish even I can plot this line here see here 1, 2 up to, up to here. So, here is 6. So, this line is somewhere here like this where this is going to present the intercept  $mc$  and similarly this angle is given by this quantity here 5, 10 of  $\theta$  is equal to 5 or 10 of angle is equal to 5. So, now one can see here by looking at the values of  $m$  and  $c$ , I can have the entire information about this line. So, this option one was incorrect and option two is correct.

Refer Slide Time: (17:08)

**Variables and Parameters**

$(m, c)$  : parameters

$(x, y)$  : variables

Knowing the parameters is equivalent to knowing the line

$y = mx + c$

Knowing  $(m, c)$

If we know  $m$  &  $c$ , we know the entire line

variable  $\rightarrow$  collect two data on variables

marks =  $m \cdot x$  +  $c$

# of hours of study

parameters

So, now if you try to have here in more detail that what is happening in this equation, you have to keep in mind that your ultimate objective is to know the equation,  $y$  equal to  $mx$  plus  $c$ . Now, when I say that I want to know the equation  $y$  equal to  $mx$  plus  $c$ , then it is equivalent to knowing  $m$  and  $c$ , if you tell me the values of  $m$  and  $c$ , then I know the entire line, if we know  $m$  and  $c$ , we know the entire line, whereas just by knowing  $x$  and  $y$ , I do not know the entire line. So, in this situation such quantities  $m$  and  $c$ , they are called as parameters, another values on which we try to collect the values they are called as variables. So, what will happen that if I try to take the earlier example of marks obtained in the examination, they depends upon see here some parameter here  $m$  in to numbers of hours of study plus  $c$ . So, we see here that here this marks and here number of hours of study, they are my variables, and  $m$  and here  $c$ , they are the parameters. So, what we try to do, we try to conduct an experiment and we collect the data on variables, collect the data on variables, Right! So, now I can solve your this question and I can say here that  $x$  and  $y$  are the variables and  $m$  and  $c$  are the parameters. So, if you try to see what is the advantage or how you



are trying to make a decision, the parameters are those values which we give you the entire information about the model. So, whenever we call that we want to find out a model in simple word, I'm trying to say I want to know the values of the parameter. So, whenever we hear the sentence like that we want to construct a model that is equivalent to saying that I want to estimate or I want to know the values of the parameters on the basis of given sample of data, Right! So, now how this data is collected and how to indicate in our statistical scale language, how to make symbols and notation for the data? So, we try to take an example here and and we try to understand,

Refer Slide Time: (20:29)

**What We Have?**

Suppose  $X$  denotes the quantity of fertilizer (in Kg.) and  $Y$  denotes the yield of a crop (in Kg.)

We want to find the relationship between  $X$  and  $Y$ .

*plot → fixed size*

Conduct an experiment and collect observations

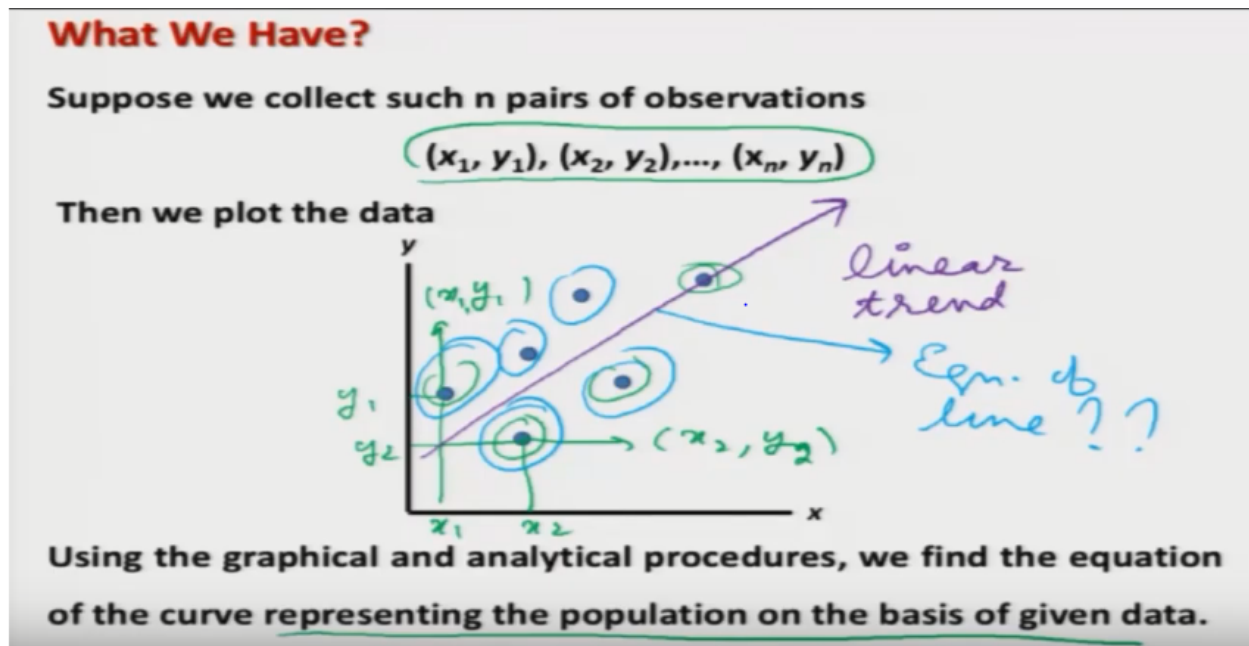
<u><math>x_1 = 1</math> Kg of fertilizer,</u>	<u><math>y_1 = 6</math> kg of yield is obtained</u>
<u><math>x_2 = 2</math> Kg of fertilizer,</u>	<u><math>y_2 = 7</math> kg of yield is obtained</u>
<u><math>x_3 = 3</math> Kg of fertilizer,</u>	<u><math>y_3 = 6</math> kg of yield is obtained</u>

and so on. *paired obs.  $(x_1, y_1)$   $(x_2, y_2)$   $(x_3, y_3)$  ...  $(x_n, y_n)$   $n$  no. of obs*

so now we take an example here, this is the same example what we can consider about the quantity of fertilizer and yield, suppose capital  $X$  is a variable, which is denoting the quantity of fertilizer in kilogram and  $Y$  is the variable which is denoting the field of a crop in kilogram. So, what is being done here, we are trying to conduct an experiment and we are trying to collect the data, so that we know what we have in our hand and our objective is that we want to find out the relationship between  $x$  and  $y$ , that is the quantity of fertilizer and yield of the crop. So, now we conduct the experiment and collect the observation in the following way, suppose I take a plot of some fixed size, Yeah! don't change the size of the plot, fixed size and then we put 1 kilogram of fertilizer in the field, and after some time we get here say 6 kilogram of yield. So, this is going to give us the value of  $x_1$ , and this 6 kg of yield is going to give me the value of  $y_1$ . Similarly, I try to repeat the experiment and then on a plot of the same size I try to put 2 kilogram of fertilizer. So, this quantity will be denoted as say  $x_2$ , which is the second observation on  $X$  then after some time, we are suppose we observe that 7 kg of yield is obtained. So, in this case the value of  $Y$  for the

second observation is converted as  $y_2$ , and similarly I can keep on repeating for example, I can take 3 kilogram of fertilizer and its value is going to be denoted by  $x_3$  and then we obtain 6 kg of yield, suppose and this value is here  $y_3$ . So, you can see here we are going to obtain here, the paired observations like, when I give the value of  $x_1$  then I get the value of  $y_1$ , when I give the value of  $x_2$  then I get the value of  $y_2$  and when I give the value of  $x_3$  then only I get the value of  $y_3$  and so on, and suppose we say that we have obtained I say,  $n$  number of observations. So, all these paired observations are going to be denoted by  $x_1, y_1, x_2, y_2$  up to here  $x_n, y_n$ , Right!

Refer Slide Time: (23:12)



Now, once we have obtained such paired observations  $x_1, y_1, x_2, y_2, x_n, y_n$ , then the first information is given by the graphical plots. So, we try to plot this data on a scatter diagram. So, for example I have just made it here these are the point which are indicating the data points, suppose this is here  $x_1$ , this is here  $y_1$ . So, this data point is denoting the value of  $x_1, y_1$ . Similarly, suppose this is here  $x_2$ , this is indicating here  $y_2$ . So, this data, is point is the location of the point  $x_2, y_2$  and so on, Right! So, now you can see here that looking at this graph, you can decide whether there is going to be a linear trend or not, Right! So, you can see here that the things are going in this direction, so that there is a presence of linear trend in the data or they can be a nonlinear trend also, but my objective is here that by looking at the values of these observations, how to know the equation of this line, how to know this thing or how to know the equation of the curve and this equation is going to be found in such a way, such that it is representing the population. What is the meaning of this contents you see, whenever we are trying to make a model, the model is given in on the or say for the entire population and the problem is that we do not know the entire

population so we have to work on the basis of given sample of data, for example have you ever heard a statement like, this medicine controls the body temperature of Americans for seven hours and the same medicine controls the body temperature of Indians, say for 10 hours and the same medicine controls the body temperature of say, say German people only for five hours, it doesn't happen, medicine is a medicine, and the effect of medicine under the similar type of persons, we will also be the same and that will be valid for the entire population all over the world, Right! But, when we are trying to know the duration of the temperature control, we try to conduct an experiment by giving the doses of the medicine to some people, we try to obtain the data and then we try to find out the equation of the curve or the line, and based on that we make a conclusion and this conclusion is valid for the entire population, this is the entire process of modeling but here in this course, we are going to find only the equation, Right! the remaining part, there is a course on say linear regression analysis and the tools are for linear regression analysis gives you all the information that how to construct a linear model, Right! But, here we are just going to concentrate only on one aspect, Okay? So, that is our now objective.

Refer Slide Time: (27:10)

**What We Want?**  
**Example**  
**Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:**  
**We know from experience that marks obtained by students increase as the number of hours increase.**

Marks →	337	316	327	340	374	330	352	353	370	380
Number of hours per week →	23	25	26	27	30	26	29	32	33	34
	①	②	③							

Marks →	384	398	413	428	430	438	439	479	460	450
Number of hours per week →	35	38	39	42	43	44	45	46	44	41
										④

28:27 / 54:03

So, now that we try to take here the same example, what we had considered in the earlier lectures and we try to see how to get this equation of the curve. So, you may recall that we had considered an example, we had recorded the marks in the examination obtained by twenty students out of five hundred marks and the number of hours they studied in a week. So, this data is given here for example, in the first row here the marks are obtained and in the second row the number of hours studied by that corresponding students are

given. So, say this is to number one, he studied he or she studied for 23 hours and he or she has got 37 marks out of 500 and so on. So, this data is here for twenty students. So, this is to number one, this is to number two, this is to number three up to here is student number twenty, Right! So, now we want to know what is the relationship between the marks obtained and the number of hours studied in a week. Although, we know from our experience that marks obtained by students increase as the number of hours increase, but we would like to see whether this statement is correct or not, Right!

Refer Slide Time: (28:31)

**What We Want?**

**Example**

**marks =**  
 c(337, 316, 327, 340, 374, 330, 352, 353, 370, 380, 384, 398, 413, 428, 430, 438, 439, 479, 460, 450)

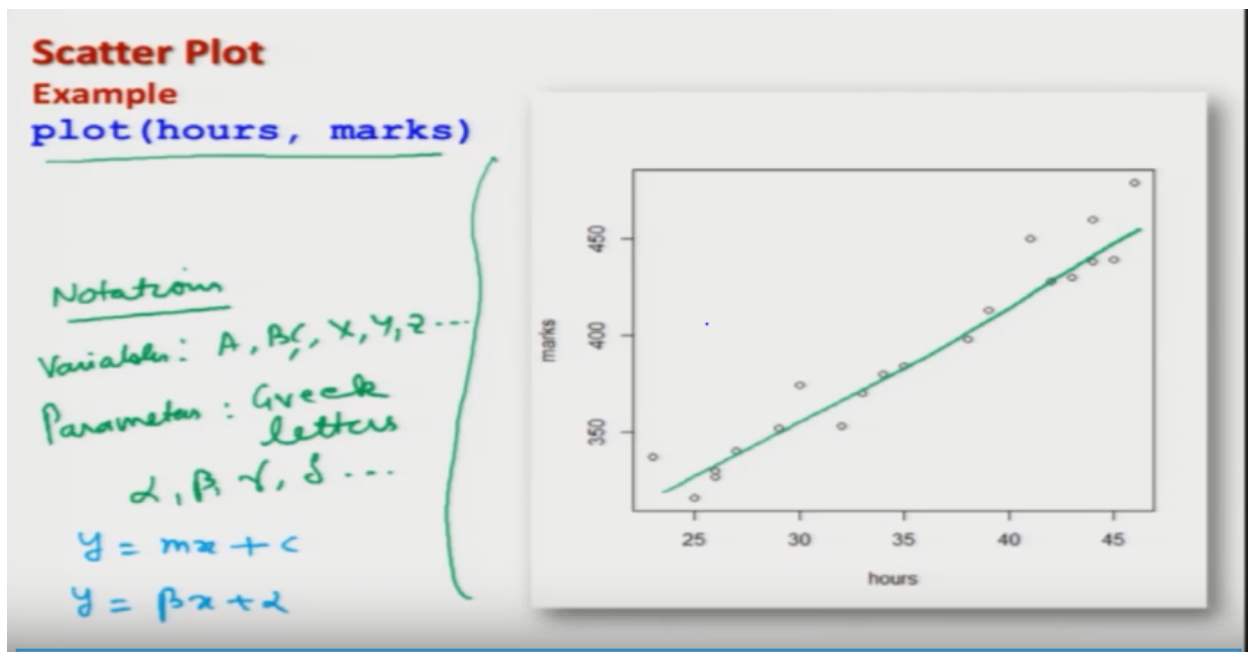
**hours =**  
 c(23, 25, 26, 27, 30, 26, 29, 32, 33, 34, 35, 38, 39, 42, 43, 44, 45, 46, 44, 41)

**Representation**  
 hours = c(23, 25, 26, ...) marks = c(337, 316, 327, ...)

$x_1 = 23$  hours,  $y_1 = 337$  marks  
 $x_2 = 25$  hours,  $y_2 = 316$  marks  
 $x_3 = 26$  hours,  $y_3 = 327$  marks and so on.

So, suppose I have a stored this data on marks inside a variable named marks, like this, in the R software and similarly the data on number of hours is stored inside a variable named hours, like this, inside the R software and how this data is presenting that is what you have to see. So, we have collected this data and this is how we are going to represent the data that  $x_1$  is equal to 23 then  $y_1$  is equal to 337 score that data inside the hours and marks, a data vector that is given in the same order you can see here, this is 23, 337. So, this is 337 in the first position and 23 in the first position. Similarly, we have the second value 25 of  $x_2$ , and 316 for  $y_2$ . So, this is again in the same order, say 316 here and 25 here. So, the data for the paired observation is given in two different vectors but the order of the observation remain the same in both the vectors, this is very important and you should keep in mind, so you can see here this 23 occurring is here, 25 occurring is here, 26 occurring is here, 337 is occurring here, 316 is occurring here, 327 is occurring here, and then after this, these are the peer observation 23 and 337, 25 and 316 and similarly here 26 and 327, Right!

Refer Slide Time: (30:12)



this is what you have to keep in mind. Now, what is your first step, first of all I would try to create here a plot that we already have actually done by including the command plot inside the argument hours and marks, Right! So, you get here a plot like this and you can see here that there is a sort of linear trend and in case if you try to make plots like this scatter is smooth another thing they will also give a sort of estimated close line, but my objective here is to know that how this line is created. So, before going into the details of this, let me give you here a small information on the notations. In the language of statistics, the variables are denoted by Say English alphabets like as A, B, C, X, Y, Z and so on and whereas the parameters they are indicated by the Greek letters like Alpha, Beta, Gamma, Delta and so on. So that is our standard language. So, now the equation which I had just expressed as  $y$  is equal to  $mx$  plus  $c$ , we had understood that  $m$  and  $c$  are the parameters. So, I try to represent it in the statistical way, and we try to write down the variable here as say  $y$ , and instead of here  $m$  I use here beta  $X$ , plus here alpha. So, that is a standard notation, when we are trying to say that the model is linear.

Refer Slide Time: (32:09)

## Scatter Plots with Line

### Example

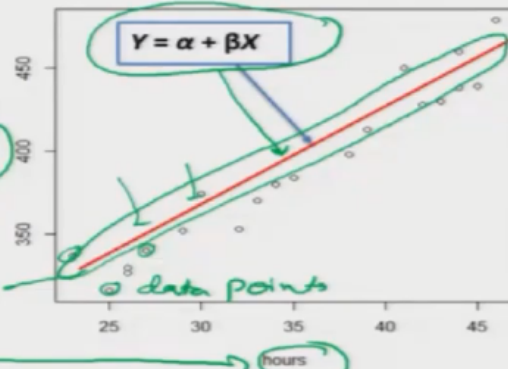
Next question?

What is the equation of this red line?

Let the equation of line be

$$Y = \alpha + \beta X$$

$X$  : Hours,  $Y$  : Marks



We want to find the relationship between  $X$  and  $Y$  in terms of

$$Y = \alpha + \beta X$$

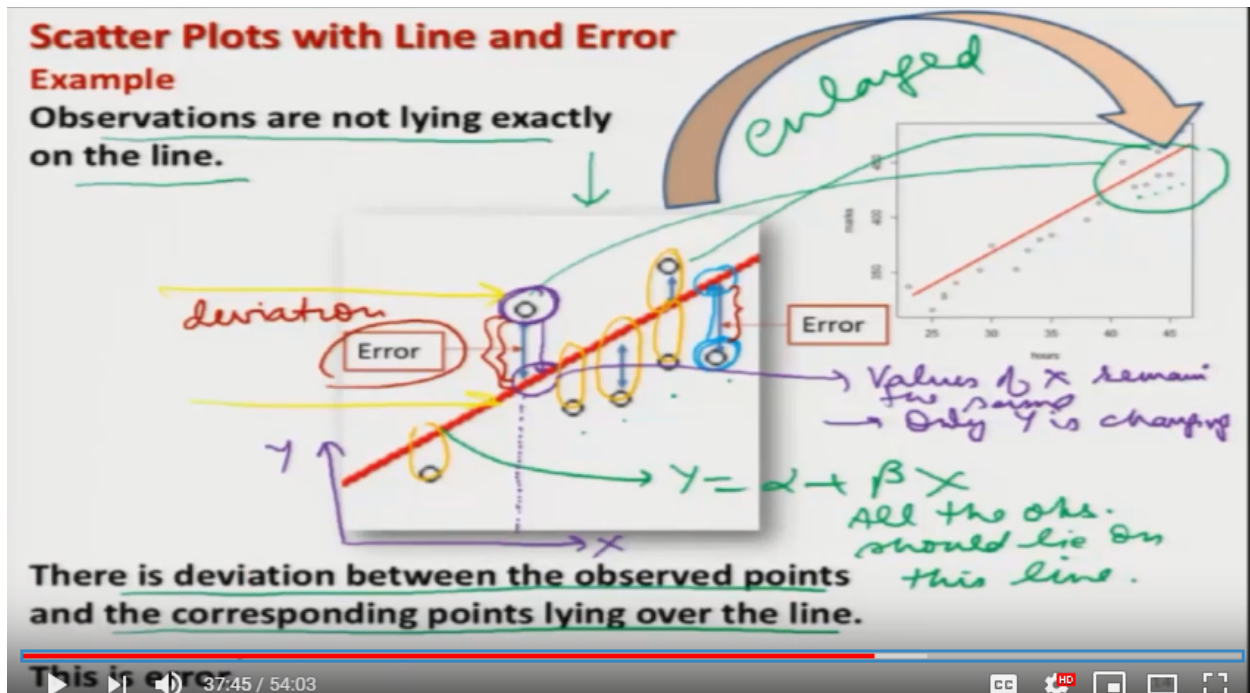
If we know  $\alpha$  and  $\beta$ , the equation will be known.

How to know  $\alpha$  and  $\beta$ ?

Now, we consider the same example and we move forward. So, you can see here in the same scatter plot, these points here this small circle, they are going to denote the data points and by my experience, I have drawn a line here, which is indicated in the red color and I have done it manually and I feel that this is a line which is representing the values or the relationship between the values of  $x$  and  $y$ , and my objective is to know the equation of this line which is in here in red color. So, suppose I try to denote the equation, the mathematical equation of this line by here,  $y$  is equal to  $\alpha$  plus  $\beta X$ , and now you know that  $Y$ , I am using here  $\alpha$  and  $\beta$  and not  $m$  and  $c$ . So, now in this equation  $y$  equal to  $\alpha$  plus  $\beta X$ , this  $X$  is going to denote the number of hours which is here and  $Y$  is going to indicate the marks obtained which is plotted on the  $y$ -axis here, and now my objective is very simple, I want to find out the relationship between  $x$  and  $y$  in terms of  $y$  equal to  $\alpha$  plus  $\beta X$ , and now you also understand and we have already discussed that this line is known to us only, when the parameters  $\alpha$  and  $\beta$  are known to us. So, I can say if we know  $\alpha$  and  $\beta$  then the entire equation will be known to us. So, now the fundamental question comes in front of us, how to know this  $\alpha$  and  $\beta$ .

Refer Slide Time: (34:08)



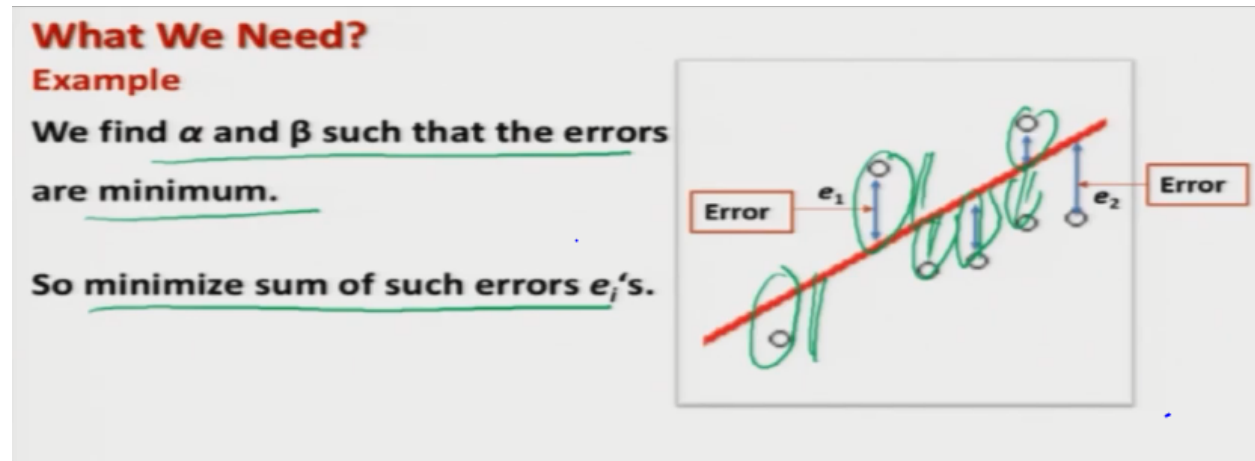


So, that is the objective what I am going to now explain here, but before going that try to observe, suppose I try to take in the same figure, I try to take this small section and I am trying to enlarge it. So, this is the figure here which is simply the enlarged part of this, say circle. So, you can see here that this point is here, this point is here and so on and there are one, two, three, four points here, one, two, three, four points here, Right! So, now if you try to see what are you trying to do, you are saying that this red line, this is  $y = \alpha + \beta X$ , and you assume that all the observation point should lie on this line, that is your idea that all the observations should lie on this line, so that you can say that this is the mathematical equation between  $X$  and  $Y$  and this is how the experiment is being controlled, but in practice this will not happen, all the points are not going to lie exactly on the same line. So, you can see here, if here is this point, then you want or you expect that this point should lie here, somewhere here and similarly if you try to take another point here, see here this point, you expect that this point should lie exactly on this line. So, you can observe in this graph that this is not really happening and there is some difference between these two points and that difference or the deviations or the deviation between say this point and this point indicating here is a sort of error which is happening in our approximation, Right! Similarly, you can see here that in all other cases also you as, there's an error here here, there is error is here, the error is here and it is here, error is here. So, you can now notice here, that there is a deviation between the absurd points and the corresponding points lying over the line, Right! and one thing what you have to observe here that in this case, there are two values  $X$  and here  $Y$ , one is here indicating here  $x$ -axis and here it is indicating  $y$ -axis. So, if you try to see here in this case, in the case where I am making this line, the values of  $X$  remains the same, in this case values of  $X$  remain the same, only  $Y$  is changing, why



Y is changing, you can see here one here is this Y, and another here is this Y. So, this is essentially it called as error.

Refer Slide Time: (37:47)



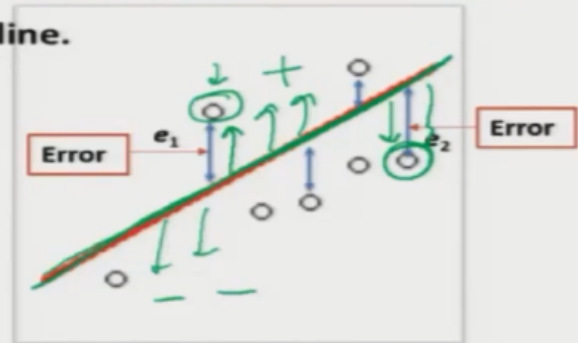
And now we would like to find out the values of alpha and beta, such that these errors are minimum and you can see here that these others are happening in each and every observation here, here, here, here, here and so on. So, now the question is how to compile all these errors, so that by minimizing that quantity I can find out the value of alpha and beta. So, one objective is to minimize the sum of such errors, I can simply measure this error and then I try to take the sum of all the errors and I try to minimize them,

Refer Slide Time: (38:29)

## What to Do?

### Example

Some errors/deviations are in positive direction and some errors are in negative direction with respect to line.



Hence the sum of errors/deviations may be close to zero indicating that there is no error or very small error.

to minimize the sum of squared errors.

but you can see here that when we want to measure these errors, then these errors are measured with respect to this line, this red line. So, this observation which I'm indicating here, this has got an error  $e_1$ , but this is above the line and this second observation here for with the error here is  $e_2$ , this is lying below the line. So, now we need to measure the direction of the points whether they are lying above the line or below the line. So, we assign that all the points which are lying above the line, I will indicate them by plus sign and all the observation which are lying below the line that I'm trying to indicate with negative sign. So, now when I am trying to act all these errors, then some errors are in the positive direction and some errors are in the negative direction with respect to the line and hence when I try to sum them, there some may become very close to zero or exactly zero, and this might be indicating as if there is no error in the data or if this sum is very small very close to zero, that will indicate that the amount of error in the data is very small which is not correct. So, now I need to devise a methodology by which I can change the sign or I can get rid of the sign. So, I have two options, either I take absolute value or I try to square these in errors. So, I opt here that we try to consider the sum of squares of this errors, that is a better option, why I am calling the better option, it's just due to mathematical simplicity, I can say at this stage and in this case I can find out the clear-cut expressions for the values of alpha and beta. in case if you try to take here absolute values, that is also possible, but that I am not discussing in this lecture.

Refer Slide Time: (40:53)

## How to Find the Line?

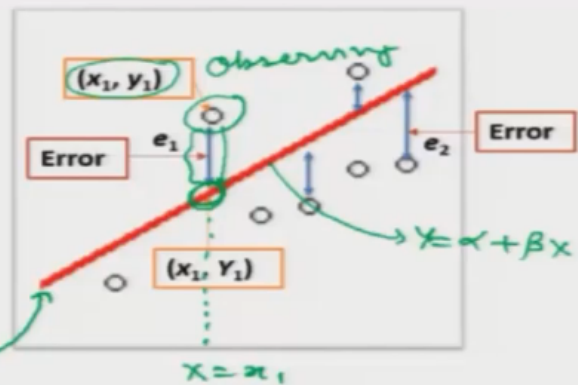
### Example

Suppose we collect such  $n$  pairs of observations

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

and every pair  $(x_i, y_i)$  satisfies

$$y_i = \alpha + \beta x_i + e_i, i = 1, 2, \dots, n$$



Find a line using the data set  $(x_i, y_i), i = 1, 2, \dots, n$  such that

- It passes through with maximum number of points
- The deviations of points with the fitted line are minimum.

So, now you can see here that when you are really trying to represent the values of this observation. So, as we have discussed here that there are two values corresponding to every observation, for example, this is the value here which I am observing inside the experiment and these values are supposed corresponding to here  $X$  equal to say here  $x_1$ . So, this quantity will have coordinate say here  $x_1$  and small  $y_1$ . Now, I assume that this point should lie on this line and this line is being denoted by  $Y$ , Capital  $Y$  is equal to alpha plus beta, capital  $X$ . So, this coordinate is going to be small  $x_1$  and capital  $Y_1$ , Right! and there is some error in this data and we call or read, Express this errors as  $e_1$ . So, now incase if I try to express this fact here, then I can express in general that every pair of the observations satisfies the equation like  $y_i$  is equal to alpha plus beta  $x_i$  plus  $e_i$ , where  $e_i$  is are the errors they can be in the positive direction or they can be in the negative direction, and now we are going to find out the values of alpha and beta, such that the sum of squares of this  $e_i$  is minimum. So, now we take a call that we will try to find out the equation of this line, which line, this line, red color line on the basis of the given data sector say it using all the small  $n$  paired observation on  $x_i$  and  $y_i$ , such that he line is passing through with maximum number of points and the deviations or points with the fitted line are minimum.

Refer Slide Time: (43:10)

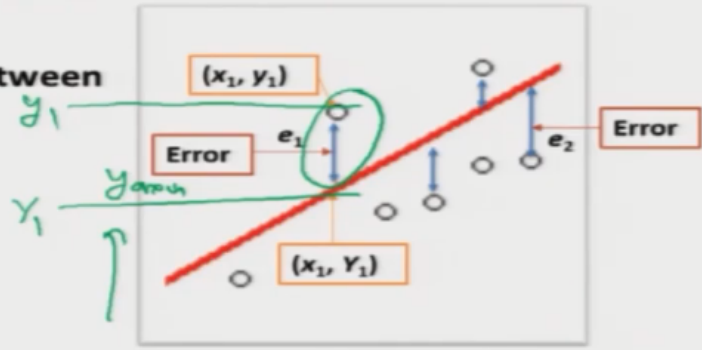
## What are Errors/Deviations?

### Example

Errors are the differences between  $y_i$  and  $Y_i$  as

$$e_i = y_i - Y_i, i = 1, 2, \dots, n$$

$$+ - < \begin{matrix} y_i - Y_i \\ Y_i - y_i \end{matrix}$$



We find  $\alpha$  and  $\beta$  such that the sum of square of errors/deviations  $e_i$ 's is minimum.

So, you can see in the same picture that this difference or this error or this deviation this is essentially the difference between this value, see here capital  $Y_1$ , and this value here is small  $y_1$  on the  $y$  axis. So, this difference is denoted as say  $y_i$  difference, Capital  $Y_i$  and this difference between  $y$  and capital  $Y$  can be positive or can be negative, but you have to use the same structure that either you try to measure them by  $y_i$  minus Capital  $y$  or  $y_i$  minus  $y_i$  Right! So, now we will try to find out the value of alpha and beta such that the sum of squares of these deviations  $e_i$  is minimum,

Refer Slide Time: (44:06)

## Method of Least Squares

Find the values of parameters such that the line passes through maximum number of given data points and the sum of squared errors/deviations from the line is minimum.

Use principle of maxima and minima to minimize

$$S = \sum_{i=1}^n e_i^2$$

how to obtain it, Right! So, for that I can use the principle of maxima and minima and we minimize the quantity, summation  $i$  goes from 1 to  $n$ ,  $e_i$  square which is the sum of square of all the deviations,  $e_i$  and this is denoted by capital  $S$ . So, now my objective is now defined, I want to find out the values of parameters alpha and beta, such that the line is passing through with the maximum number of given data points and the sum of squared deviation or errors from the line is minimum and this is called as method of least square or principle of least squares, and the principle of least square is simply saying that try to find out the the equation of the lines in such a way, such that the line is passing through with the maximum number of given data points and the sum of squared deviation from the line is as minimum as possible. So, now in order to find out the value of alpha and beta using the principle of least square,

Refer Slide Time: (45:16)

**Method of Least Squares**

$$S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

Find  $\alpha$  and  $\beta$  such that  $S$  is minimum.

$$\frac{\partial S}{\partial \alpha} = 0 \Rightarrow -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0$$

$$\sum_{i=1}^n y_i - \sum_{i=1}^n \alpha - \beta \sum_{i=1}^n x_i = 0$$

$$n\bar{y} - n\alpha - n\beta\bar{x} = 0$$

$$\bar{y} - \alpha - \beta\bar{x} = 0$$

$$\Rightarrow \alpha = \bar{y} - \beta\bar{x}$$

$\Rightarrow \alpha = \bar{y} - \beta\bar{x}$  provided  $\beta$  is known.  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ ,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

Handwritten notes:  $e_i = y_i - \alpha - \beta x_i$ , Find values of  $\bar{x}, \bar{y}$ ,  $(x_i, y_i) i=1, \dots, n$ , sample

we try to write down the sum of squares of errors, like here this and you know that  $e_i$  is given by  $y_i$  minus alpha minus beta  $x_i$ . So, I try to replace this  $e_i$  over hereby  $y_i$  minus alpha minus beta  $x_i$  whole square, and now I have to use the principle of maxima and minima. So, for that I need to find out the first order derivative of this  $S$  with respect to alpha and beta, I need to put it equal to zero and then, I need to check by finding out the second order derivative that whether the maxima or minima has been achieved. So, I try to find out the first order partial derivative of  $S$ , with respect to alpha and this give us this equation and following the principle of maximum, minima I try to put this first order condition equal to zero. So, now incase if you try to solve it you get here, that summation  $y_i$ ,  $i$  goes from 1 to  $n$ , minus summation  $i$  goes from 1 to  $n$  alpha minus beta times summation  $i$  goes from 1 to  $n$   $x_i$  is equal to 0. So, this quantity is

simply you're here,  $n$  times  $\bar{y}$  minus  $n$  times  $\alpha$ , minus  $n$  times  $\beta \bar{x}$ , Right! Because,  $\bar{x}$  and  $\bar{y}$  are defined as  $\frac{1}{n}$  upon  $n$  summation  $x_i$ , and  $\bar{y}$  is  $\frac{1}{n}$  upon  $n$  summation  $y_i$ , Right! So, now you can solve this equation, and this gives you here  $\bar{y}$  minus  $\alpha$  minus  $\beta \bar{x}$  is equal to 0, and this gives us that  $\alpha$  is equal to  $\bar{y}$  minus  $\beta \bar{x}$ , which is here. So, now on the basis of given set of data, I can find out the values of say here  $\bar{x}$  and  $\bar{y}$ , because we have observations  $x_i$  and  $y_i$ ,  $i$  goes from 1 to  $n$ . So, now this value of  $\alpha$  is going to be known to us, if  $\beta$  is known. So,  $\bar{y}$  is known from the sample data,  $\bar{x}$  is also known from the sample data and  $\beta$  is unknown. So, now my next objective is how to find out this  $\beta$ .

Refer Slide Time: (47:43)

**Method of Least Squares**

$$\frac{\partial S}{\partial \beta} = 0 \Rightarrow -2 \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0$$

$$\beta = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\beta}, \text{ (denoted as } \hat{\beta} \text{)}$$

$$\alpha = \bar{y} - \hat{\beta} \bar{x} = \hat{\alpha} \text{ (denoted as } \hat{\alpha} \text{)}$$

obtained on the basis of given data  $(x_i, y_i)$

So, we try to do the same process and we obtain the first order partial derivative of  $S$  with respect to  $\beta$  and I try to substitute it equal to 0, then I get this equation and I put it equal to 0 this is the first order equation, and now if you solve it, that's a pretty simple algebra, you will get the value of here  $\beta$  like this, which is summation  $x_i$  minus  $\bar{x}$   $y_i$  minus  $\bar{y}$ , summation goes from 1 to  $n$ , divided by summation  $i$  goes from 1 to  $n$ ,  $x_i$  minus  $\bar{x}$  whole square, and now this is the value of  $\beta$  that can be obtained on the basis of given set of data on say  $x_i$  and  $y_i$ , Right! So, this estimated value of  $\beta$  which is obtained on the basis of given sample, it is denoted here as a  $\beta$  hat, that just write  $\beta$  and put here a gap, hat, Right! So, now I can obtain the value of  $\beta$  as  $\beta$  hat from the given sample of data and I try to substitute the value of  $\beta$  equal to  $\beta$  hat in this equation of  $\alpha$ . So, once I try to substitute here  $\beta$  equal to the  $\beta$  hat, I get here the value of  $\alpha$  which now I can find out on the basis of given set of data. So, this is denoted as  $\alpha$  hat. So, now you can see here we have obtained two values of

parameters alpha hat and beta hat. So, alpha hat is the value of alpha and beta hat is the value of beta that can be obtained on the basis of given sample of data.

Refer Slide Time: (49:36)

**Method of Least Squares**

$$\left. \begin{aligned} \frac{\partial^2 S}{\partial \alpha^2} \Big|_{\alpha=\hat{\alpha}} &> 0, \\ \frac{\partial^2 S}{\partial \beta^2} \Big|_{\beta=\hat{\beta}} &> 0. \end{aligned} \right\}$$

$y = \alpha + \beta x + e$  (Model)  
 $y = \hat{\alpha} + \hat{\beta} x$  (fitted model)  $\sum e_i^2$

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

Least squares estimate of  $\beta$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Least squares estimate of  $\alpha$

Now, mathematically the next issue is how do I know, whether this value of alpha and beta are minimizing the sum of squared deviations or not. So, I try to find out here the second order partial derivatives of S with respect to alpha and beta and I try to substitute the value alpha equal to alpha hat and beta equal to beta hat and these values comes out to be zero, that you can verify yourself. So, now I have a value of alpha hat and beta hat like this. So, this beta hat is given by this expression, which is telling us the what is the value of beta on the basis of given sample of data and this beta hat is called as least squares estimate of beta, this is based on the principle of least squares, and similarly, the value of alpha we which is obtained here is alpha hat, y bar is minus beta hat, x bar and this can also be estimated on the given sample of data and this is called the least square estimate of alpha. So, now you can see here the equation y equal to alpha plus beta x plus e, that was our original model that we wanted to find, and now we have found the value of alpha to be alpha hat and beta to be beta hat x, and now it will become 0, because we have obtained this equation in such a way such that sum of  $e_i$  square is minimum and the value of  $e_i$  for which this sum is minimum is 0. So, that is why this is called as fitted model, Right! and this is simply called here as a model, and now we try to compute these two values, the values of beta hat and alpha hat on the basis of given sample of data,

Refer Slide Time: (51:28)



## Method of Least Squares

### Example

Solving it for the given data on **marks** and **hours**, we get the values of  $\alpha$  and  $\beta$  as follows:

$$\bar{y} = \frac{1}{20} \sum_{i=1}^{20} y_i = 389.9, \quad \bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = 35.1$$

$$\hat{\beta} = \frac{\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{20} (x_i - \bar{x})^2} = 6.3,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 168.65$$

Model:  $\text{marks} = 168.65 + 6.3 * \text{hours}$

*fitted model*

on the basis of example, that we were considering earlier. So, you can see here, this marks is here your Y and number of hours per week is your here X. So, this is the value of  $y_1$  and this is the value of  $x_1$  and this is the value of  $x_2$ , this is the value of  $y_2$  and so on, this is you here all that data set, now I try to compute the values of  $\bar{x}$  and  $\bar{y}$  which is here  $\bar{y}$ , if you try to compute  $\frac{1}{20} \sum_{i=1}^{20} y_i$ , this comes out to be 389.9 and  $\bar{x}$  says simply comes out to be a sample mean of all the values of  $x_i$ 's as 35.1, and similarly the expression of  $\hat{\beta}$ , if you try to substitute all the values, this will come out to be  $\hat{\beta}$  is equal to 6.3 and the expression for  $\hat{\alpha}$  will come out to be 168.65. So, now you can see here that your model becomes here marks is equal to 168.65 plus 6.3 in two hours. So, this is you're here, fitted model and you can see here that this model has been obtained on the basis of given sample of data only, Right! Okay. So, we stop here now in this long lecture and you have seen that how we have computed the values of parameters  $\alpha$  and  $\beta$  on the basis of given sample of data, but we have done this computation manually, now I will try to show you in the next lecture that how to get all these values from the R software directly, but it is important for you to understand that how the values inside the software are obtained and what are the computations and philosophy and the concept which has been used in the computation. So, you try to understand this concept and I will see you in the next lecture, till then goodbye.