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 - 41
Simple Linear Regression Analysis
Fitting Linear Model with R Software

Hello friends, welcome to the course Essentials of Data Science with R Software 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 continue with our chapter on Simple Linear Regression Analysis and in the last lecture we had considered the simple linear regression model where we had considered only one independent variable, and we had obtained the values of the parameters.

We also looked that how you can obtain the fitted model, fitted values and the difference between the observed values and fitted values. But, before I go further, first I would try to show you that how these things can be obtained in the R software. One point which you have to keep in mind that in order to obtain it I am going to use the package lm which means linear model.

And, this linear model is also used in the case of multiple linear regression model, and once I come to multiple linear regression model, there will be some more concepts which will be coming automatically into picture. So, at this moment I am going to use this package lm only for the simple linear regression model so, when you try to do it there will be some more outputs which you may not understand.

But that is not a very big deal because, after some time when we cover the multiple linear regression model then we will able understand each and every line in the software outcome, right.

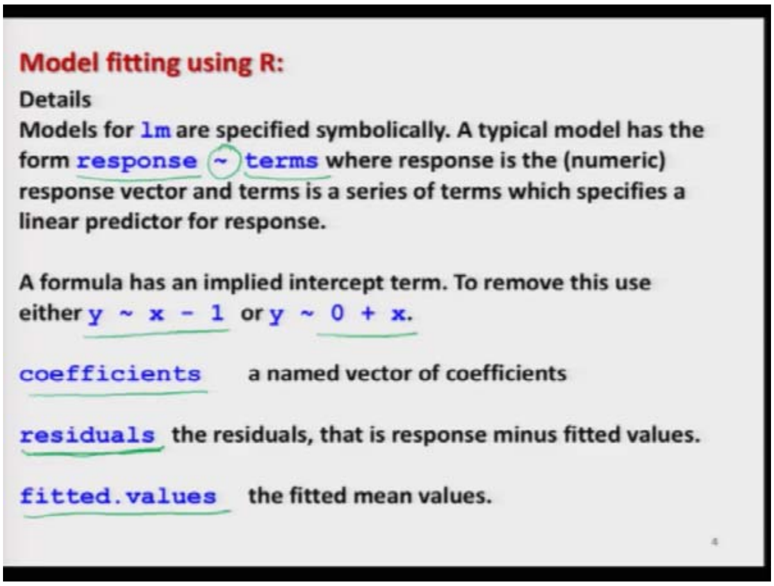
So, today my objective is this first I try to introduce you with this package then I will try to take the same example which I took in my first lecture, and we will obtain different values and I will try to show you that how they will help us, and after that I will come to R software also and on the R console I will try to show you, ok.

So, let us begin, ok. So, in order and the RPM of the fan depends on several factor like voltage, current etc., which is controlled by the regulator of the fan or the switch of the fan. In case, if you try to control the regulator from point number 1, 2, 3, 4, 5 you see that the RPM of the fan is increases or in simple language the speed of the fan increases.

But, in case if you simply switch off the switch that means current and voltage is 0. So, when voltage and current are 0 what will be the RPM of the fan? This will not start. So, in this case when the value of x is 0 this y is going to be 0. So, this is how you have to decide whether you want to have a model with intercept term or without intercept term.

Both the models have their own consequences different types of interpretation, different types of formulation that you will try to see gradually when we try to learn more in the regression analysis, ok.

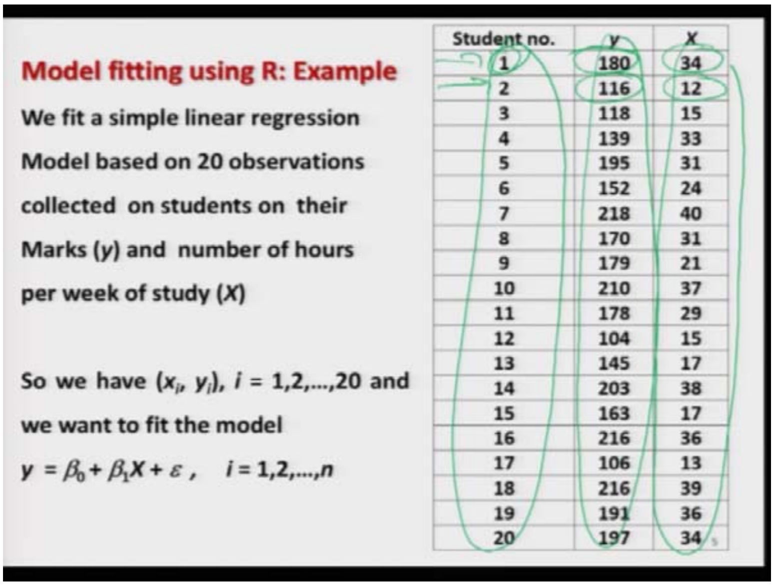
(Refer Slide Time: 03:53)



So, now once you have fitted a model there are different types of options which can be exercised to extract the particular value. Suppose, if you want to extract only the regression coefficient, so then you have an you have a option here to use coefficients if you want to have; if you want to extract the values of the residuals then you can use this command residuals with in a particular way.

And this will give you the value of residual that is the difference between the observed and fitted values. And if you want to find out extract only the fitted values then also that can be used that can be extracted by the command fitted dot values.

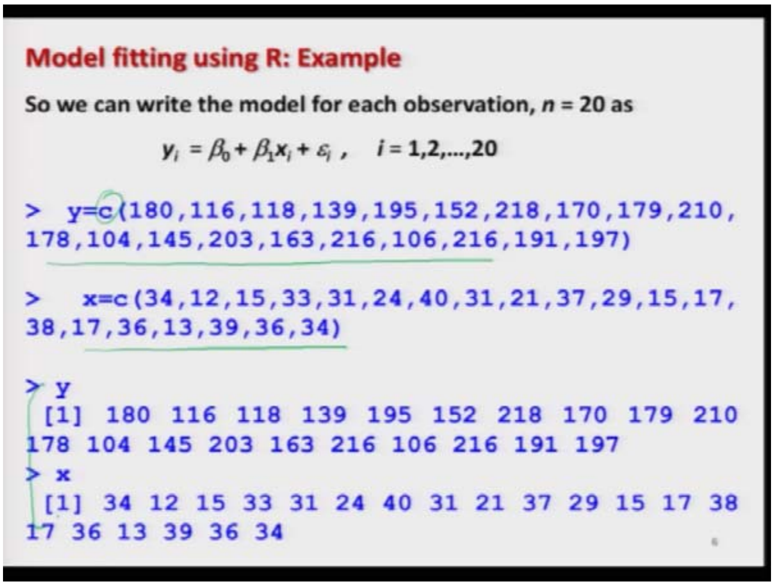
(Refer Slide Time: 04:37)



So, now let us take the example and we try to understand all these executions and their interpretation. So, this is the same example where I have considered 20 students and we have obtained their marks out of 250, and we also have recorded that how many hours a student has studied.

For example, student number 1 has got 180 marks and he has studied he or she has studied 34 hours in a week similarly the student number 2 has got 116 marks and the student has studied say 12 hours in a week.

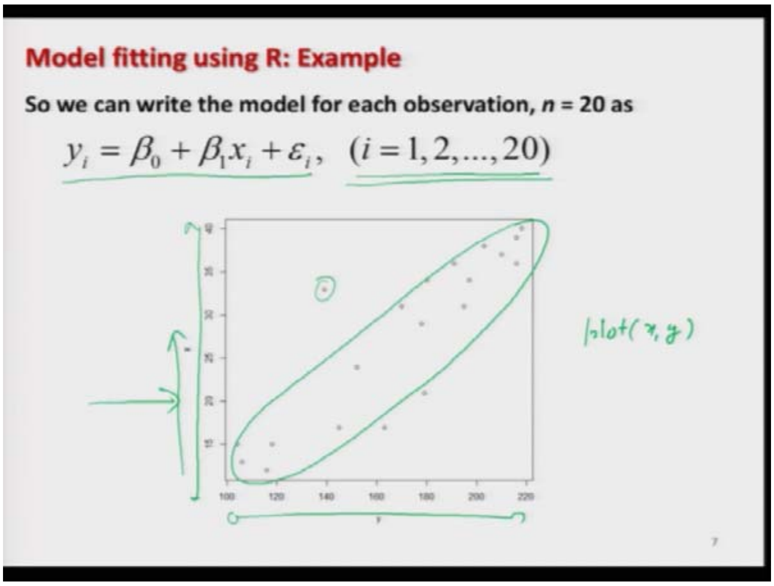
(Refer Slide Time: 05:15)



So, this is now 20 values of period observations on x_i and y_i. So, first I try to create the data vectors of this input variable, I have two options either I try to use the framework of data frame or I simply try to use two data vectors.

So, here I am trying to make the life simple. So, I have simply in type these values in the form of a simple data vector using the command c. So, the so these are the values on y which is marks and these are the value on here x on the number of hours of a study. So, you can see here this is the data, right.

(Refer Slide Time: 05:54)



Now the first question comes, that whenever you have got these 20 observations on say x_i and y_i you would like to fit a model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. So, first question comes this model can be fitted to this given set of data only when there is a linear relationship between x and y.

How to know this thing? Although, I have said that the linearity of the model depends on the parameters but, there is a possible way out. We can look at the scatter diagram, you can see here I have plotted here the scatter diagram of x and y simply the using the command plot x and y, I will try to show you on the R console.

But, you can see here that these are the location of the different pairs of observation. You also have to keep in mind that the scale here on the x and y axis because, sometime the scatteredness means you can see here one point is here, one point is here this scatteredness becomes very very large and it actually looks very large, but it is not.

Because, sometimes you are trying to make this scale quite wide or sometime the opposite also happen, but if the scale is too narrow then even the difference of 1000 will look very small. So, these are the minor thing which you have to keep in mind, but anyway if you try to look at this data set you can see here that more or less there is a sort of linear trend, right.

It cannot be 100 percent linear that is practically impossible in real life. So, this will this picture will this graph will give us a confidence yes we can use here the setup of linear model and we can find out the model using the simple linear regression analysis.

(Refer Slide Time: 08:01)

So, what I try to do here, that suppose I prepare here 4 lines, 4 fitted lines because, first you have to decide that how the line will look because you are going to use the principle of least square to fit these values. So, I have just created say artificially 4 lines on this on the same graph.

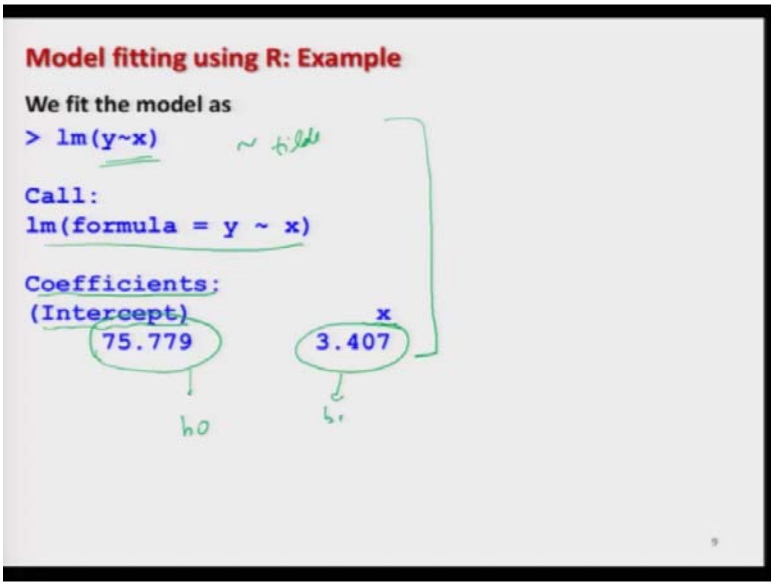
One line is going so, you can see that this is here the trend I mean this is here the data set and I have plotted here 4 lines in red color, this is line like this, this line is going more vertical, this line is something like this and this is here like this. So, what do you think? Which is the most representative line which will possibly fit to the data?

You can see here that this line is going here and most of the data is here and many many points are away from this line. So, this cannot be a very good line. Similarly, you can see here the data is here and your line is going here. So, this difference here is very high so, this is not a good fitted line.

Same thing is happening here the data is here and the line is here so, this difference with the line this observation this random errors are very high. So, this is also not a good line. So, these 3 options are not very good, but if you try to look in the fourth option this line is passing through with the maximum number of points within the cluster of the observation, and this error looks to be reasonably smaller than all the 3 remaining graphics.

So, in my opinion this looks like a good line. So, possibly we are looking to fit a line which will look like this one. How to do it? That we will try to see and at the end I will try to show you that the line which we have fitted and the line which we have which we are thinking that this is good, what is the difference between that two, ok.

(Refer Slide Time: 10:08)



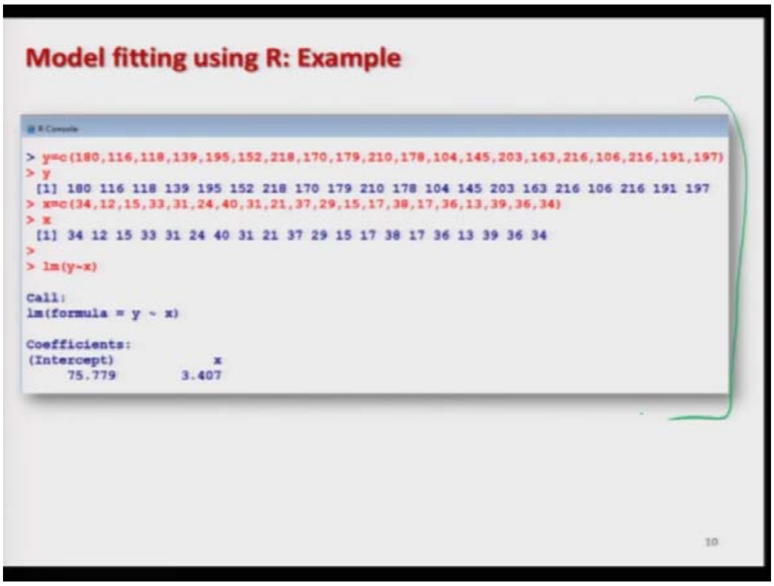
So, now here I use the lm command, I already have entered my data on y and x. So, I simply use here lm y equivalent sign x, right, this is indicated on your keyboard like this. So, this is a sort of tilde, right. So, if you try to execute this thing over here we would get an outcome like this one. So, now your objective is your job is to look into this outcome and try to see what are you getting.

So, first line will give you the formula which we have used then, we have a column line here coefficients and then it is giving you here intercept term and it is giving you here the value of x. So, your job is basically to know what are these values 75.779 and 3.407, right, this is your basic objective and this is your job.

Now, tell me one thing, if you have not done the earlier lecture on regression analysis, can you really know what are these thing? Somebody may tell you by some experience, ok this is something like the value of b_0 and this value is something like b_1, as long as the experiment is nice well you can work on it, but sometimes these values you will see that they are not really matching with the experiments outcome.

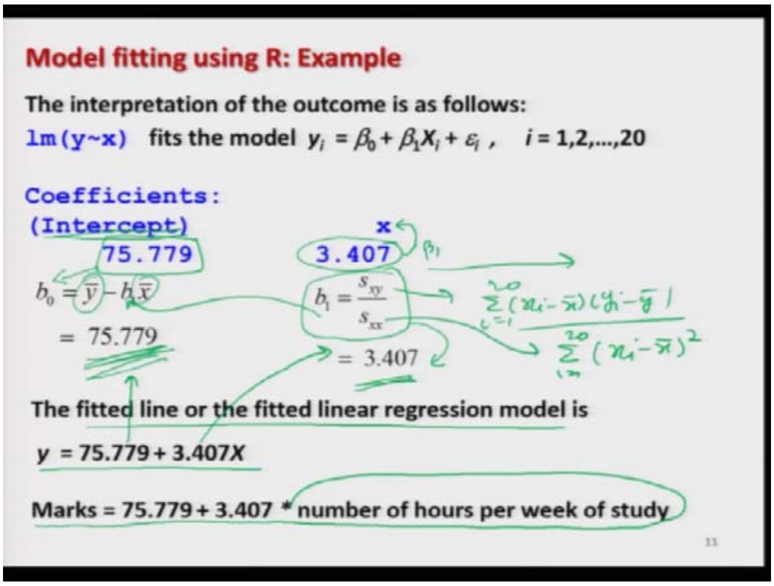
In that case you will get badly confused and then you have to see what is really happening and what are these values. So, at this moment my objective is very simple that I want to explain you the different outcomes in this R software. And I want to show you or I want to tell you that what are these things and how these values have been obtained, right.

(Refer Slide Time: 12:10)



So, if you try to see here that this is the screenshot of whatever I have done.

(Refer Slide Time: 12:16).



And now, let me try to take here one by one these outcomes and try to show you how you are going to interpret it. So, with this blue color you know that you have a you have the software outcome. So, now what is this here? Intercept term, this is 75.779 and this is the value of b_0 that you had obtained as $\bar{y}$ - $b_1 \bar{x}$.

So, this is the estimate of β_1 which is your intercept term. And what is your here this x 3.407 so, this is essentially indicating the value of β_1, which is the regression coefficient associated with x. If you have more variable this thing will be extended in this direction you will see, ok.

So, now the value of β_1 that you obtained after estimating by the estimator b_1 is equal to s_{xy} upon s_{xx} this comes out to be 3.407. So, if I give you an option that you try to compute the function $s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),\ \ s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Then, and then if you try to take the ratio of the two this value will be coming to be 3.407. So, whatever efforts and time you will spend in computing it manually the software is helping you and it is giving you this value 3.407 and if you try to substitute this value of b_1 here in this and try to obtain the value of $\bar{y}$ $\bar{x}$ manually.

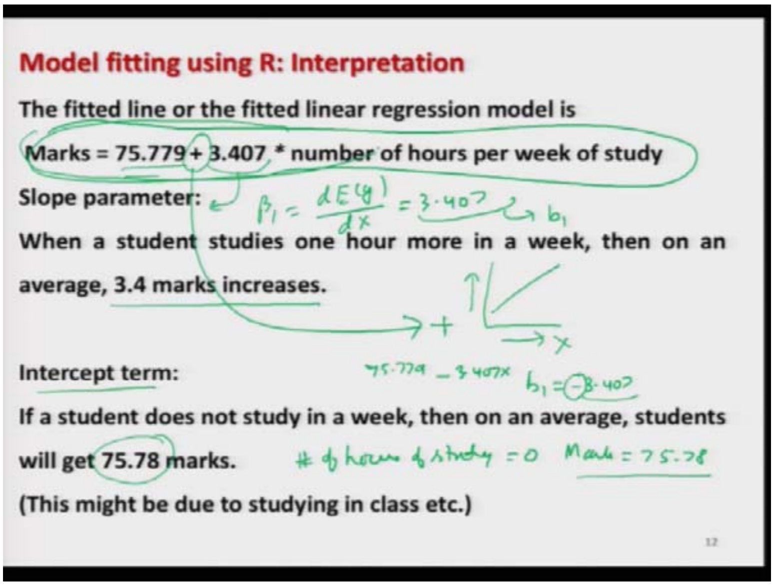
Then, you will get the same thing here 75.779. So, after obtaining the value of b_0 and b_1 I can obtain the fitted line or the fitted model or the fitted linear regression model as y is equal to $b_0 + b_1$ x. So, my model becomes here y is equal to 75.779 which is coming from here + 3.407 which is coming from here times x.

So, if you try to translate it in the in terms of the given variable now, you have here a model marks are equal to 75.779 plus 3.407 times the number of hours per week of study. So, now you can see here that here in this slide you had only the observations. There was no mathematical relationship which was known to us between x and y.

But, in this slide now, you have obtained the relationship between marks and number of hours and now you can see when I say you have to find a model or you have to find a simple linear regression model your job is only to find out the value of b_0 and b_1 that is all, right. Now, there can be different types of estimation approaches which can be implied to estimate the values of β_0 and β_1 there can be method of movement, there can be method of absolute deviation or the method of maximum likelihood and so on.

And it is possible that, different estimation techniques will give you different values also but, my personal belief is that if your data is good representative nice and if it is satisfying all the assumptions whatever method you choose they will give you the value of β_0 and β_1, which are not differing much and all the values should be nearly the same or say acceptable to us, right.

(Refer Slide Time: 16:13)



So, this is our model. Now, what is the interpretation of this thing? What is the use of this thing? Once you have obtained this fitted model now, this is here the value of slope parameter 3.047. So, as we had discussed that β_1 is the partial derivative or the derivative of expected value of y with respect to x. So, this is here 3.407.

So, what are you trying to say? β_1 is indicating actually this is the value of b_1 that is the value of β_1 which is obtained on the basis of sample. So, what are you trying to say? You are trying to say that if a student studies for one hour more then on an average 3.4 marks will increase, right.

This is the meaning. And you can see here one thing more here there is a positive sign; positive sign means the relationship is increasing something like this. So, you are trying to say if the number of hours on x are increasing then the number of marks will also increase. Suppose for a while, suppose this relationship would had been something like 75.779 - 3.407x.

This means what? The interpretation of the b_1 will remain the same that if a student studies for one hour more then, there will be a change in the marks by 3.407 units, but you are trying to say since there is a negative sign here so, the marks will decrease. So, you are trying to say somehow that if a student studies more than the marks will decrease which is really not possible if a student is studying honestly, right.

And, if you try to see this is the same thing which is indicated by this model also. So, whatever is happening in a real life that is also indicated by the model and that is what I meant when I said the model has to be representative, the model has to explain us the phenomena whatever is happening in a real life the same thing should be conveyed by the model also, right.

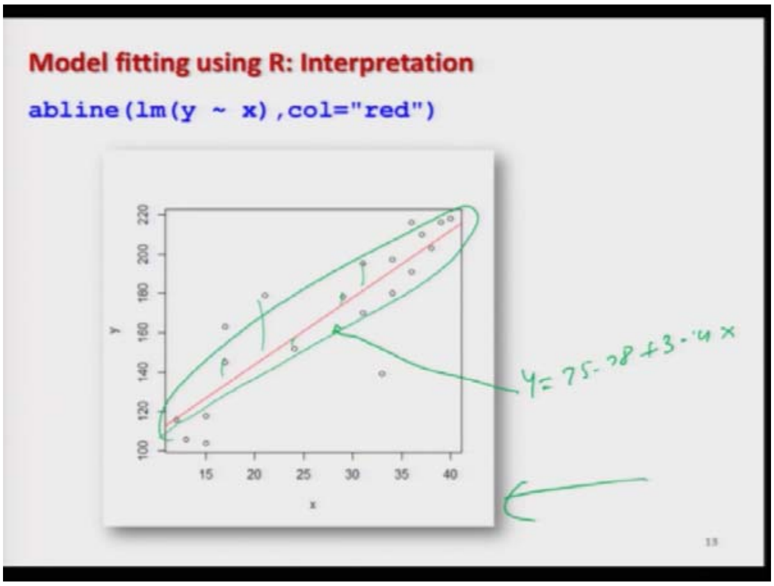
So, now we also have understood what is the interpretation of plus and minus sign of slope parameter, right, ok. Now, you come to the next value; intercept term, intercept term here is 75.779 or say nearly 75 say 70.78. So, you are trying to say in this model, if the numbers of hours are of study they are equal to 0 that is equal to 0.

Then, a still the marks are going to be say 75.78. So, you are trying to say if a that if a student that does not study in a week at all then, also on an average students will get 75.78 marks. Well, this is possible if a student is attending the classes very well, the student is studying in the library, the student is trying to read several books on the topic; that means, he is gathering that knowledge the only thing is this you had asked that how many hours you study in a week.

So, possibly he said no I am not studying at all, but you have not asked that are you going to the library, are you trying to read eBooks etc. And then, if a student is very intelligent also then just by attending the class he can get some marks.

So, that is what is being indicated by this model at least in my opinion that is my opinion you can have different things also and you can argue a lot that how the student is getting a mark without studying, well this is a hypothetical example to explain you, right, ok.

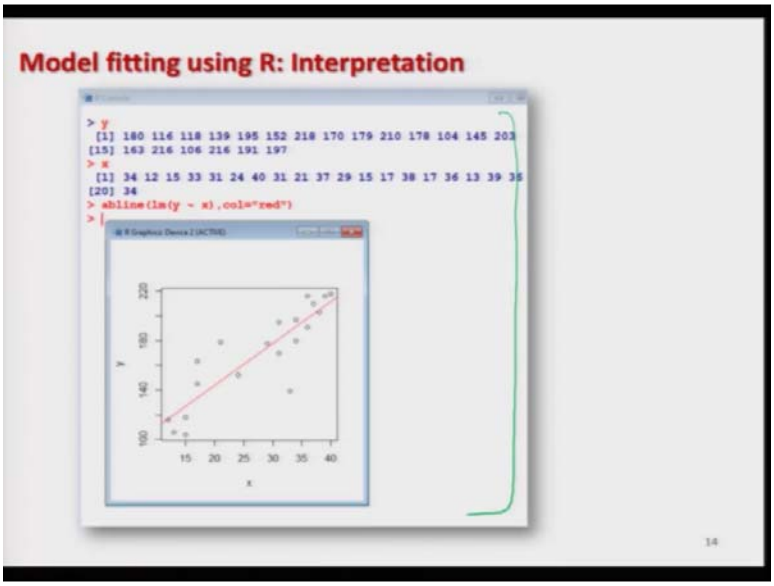
(Refer Slide Time: 20:58)



So, this is how we try to interpret the model and this is how we can do it. Now, I try to plot the data and I try to fit the line here what line which you have obtained here, y is equal to 7 this line y is equal to say 75.78 + 3.4 x I tried to plot this line here. So, on the same graphic I tried to plot this fitted line in red color; you can see here this line is passing through with the maximum number of points and in which this error is as small as possible.

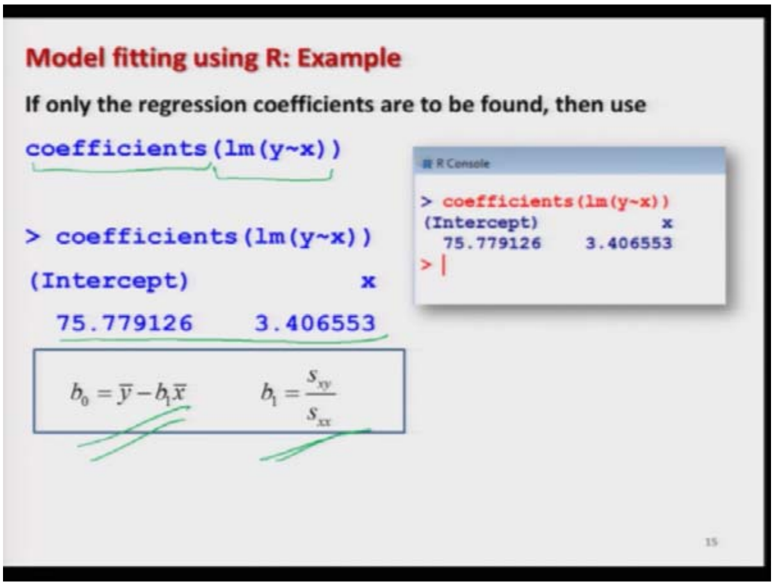
And, if you try to compare this graphic with this graph which we studied which we consider initially this graph here the fourth one they are more or less similar, and believe me I had created this graph only by manual means, I have just drawn this line on my laptop using the paint, ok. But, you can see here the line which I could guess and the line which I have obtained after doing a mathematical treatment of the data they are nearly the same, ok.

(Refer Slide Time: 22:20)



So, this is how we go further, ok. And this is the screenshot of the same graphic you can see here that with the data is the same and you are getting the same line well, you can also obtain it.

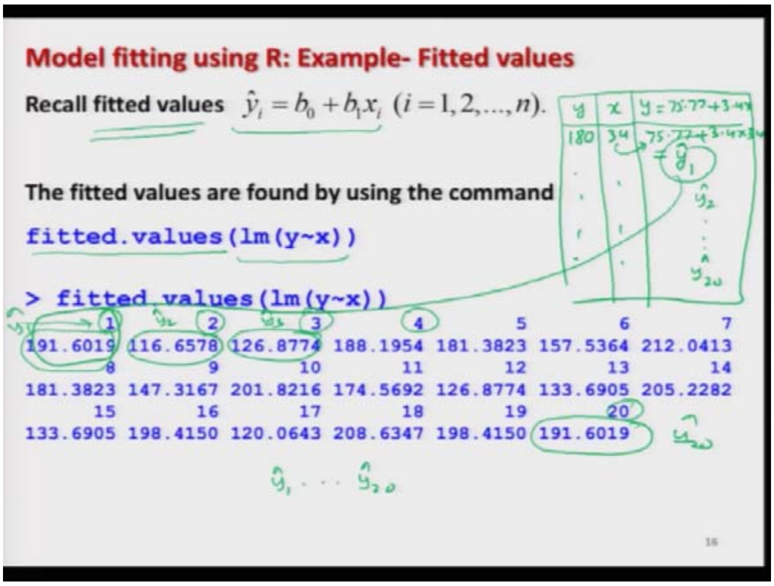
(Refer Slide Time: 22:28)



Now, suppose I am not interested in all this thing and I simply want to obtain the coefficients that is all. So, if you want to extract the coefficient in this analysis then, you can use the command here coefficient c o e f f i c i e n t s and inside the parenthesis you try to write down the function that is lm, and inside the parenthesis y tilde or equivalent sign x, right.

And if you try to execute it you will get the same thing here. So, it has given you only the values of β_0 and β_1 which are estimated by b_0 and b_1 here like this. So, this is how you can simply obtain the values of the coefficients, right.

(Refer Slide Time: 23:20)



Similarly, so now, suppose your interest is only in finding out the fitted values. What do you really mean by fitted values? These values are obtained by this expression $\hat{y}_i$ is equal to $b_0 + b_1 x_i$. What this actually mean? So, if you try to see what I do here I try to explain you with a very simple thing you have obtained the data on x and y.

So, y was suppose you can see here in this table where you had taken this data you can see here for this student number 1 it is 180 and 34. So, suppose this data is obtained here y here as a 180 and x here as a 34 hours in a week. Now, you have obtained the model y is equal to say 75.77 + 3.4 x.

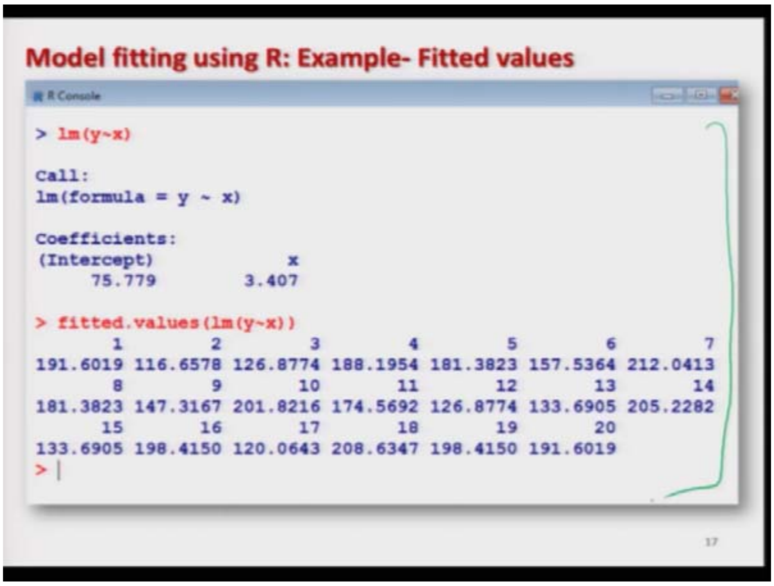
So, now what I am asking you use this x here and try to obtain the value 75.77 + 3.4 34. And then whatever is this value, this will be the value of y_1 which you have estimated from the sample. So, this is your first fitted value and this first fitted value is given here, and similarly if you try to take another set of data and if you try to obtain here

$\hat{y}_1, \hat{y}_2, ..., \hat{y}_{20}$. So, these are the values of y which are obtained on the basis of fitted model, but using the same input data, right.

So, this can be obtained by the command fitted dot values and inside the parenthesis you have to write lm and y equivalent x or y tilde x, ok. So, this is the you can see here once again you have here say 1, 2, 3, 4 up to here 20. So, what are these things? These are the 20 observations on what?

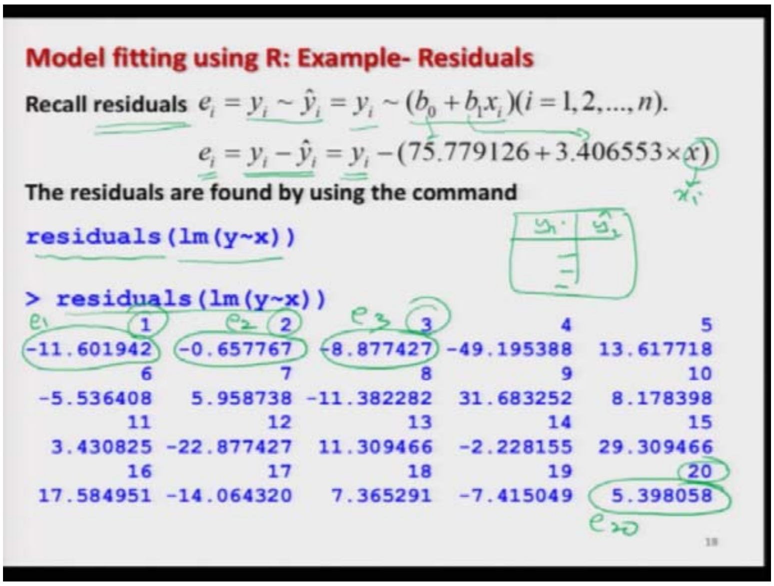
So, this 1st value this 1st value is here $\hat{y}_1$ and similarly here this 2nd value this is here $\hat{y}_2$, similarly here 3rd value this is here $\hat{y}_3$ and similarly here in the last this is $\hat{y}_{20}$, right. So, these are your fitted values. So, from this command you can obtain all the values $\hat{y}_1, \hat{y}_2, ..., \hat{y}_{20}$, right.

(Refer Slide Time: 26:15)



And now, if you want to find out the difference between the two that also can be obtained and this is here the screenshot of the same operation which I shown you.

(Refer Slide Time: 26:22)

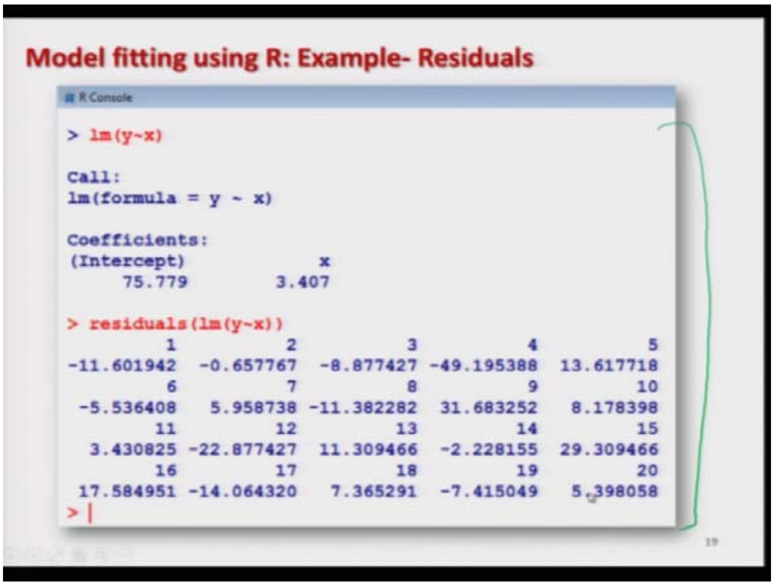


And suppose you want to find out the difference between the estimated and fitted value which we have called as residuals, ok. So, e_i was defined as y_i difference $\hat{y}_i$. So, this will be y_i and $\hat{y}_i$ has been obtained by $b_0 + b_1 x_i$. So, for example, so these are essentially use of e_i is equal to y_i - $\hat{y}_i$ y_i is given in the data then I try to obtain the $b_0 + b_1$ here x. So, x will be changing with i x_i.

So, what we have done? So, we already have obtained the $\hat{y}_i$ also if you try to now means in the earlier slide we had obtained y_i we had obtained $\hat{y}_i$. So, if you try to do it manually you simply have to take the difference of all $\hat{y}_i$ and $\hat{y}_i$. So, which has been given by the R software using the command residuals and inside the parenthesis lm y equivalent or tilde x, and you can obtain here these residuals.

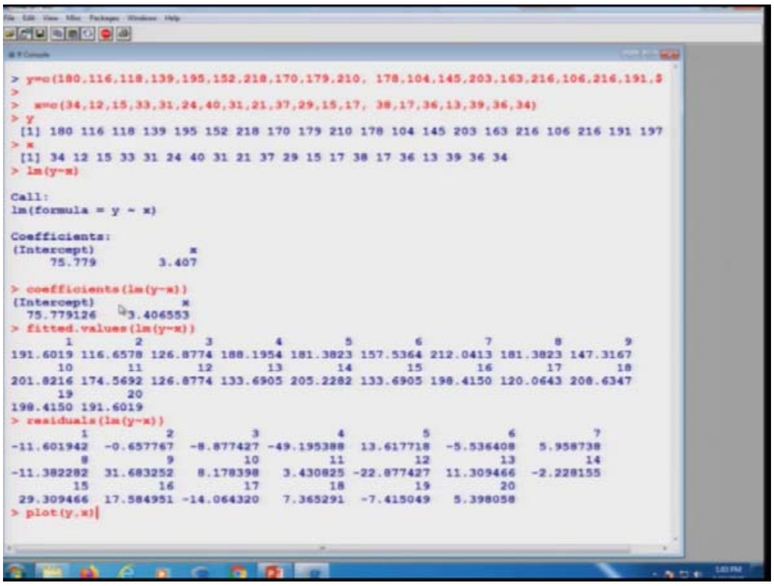
So, here again once again you can see these 1, 2, 3, 4 up to here number these are the 20 observations on the residuals. And the 1st residual that means, y_1 - $\hat{y}_1$ this is your here e_1 this is here like this - 11.60 similarly the 2nd value here is e_2 3rd value here is e_3 and similarly here the last value here is e_{20}. So, this is here e_{20}, right. So, we have now obtained e_1, e_2, e_{20} all the residuals and this is here the screenshot of the same thing, ok.

(Refer Slide Time: 28:16)



So, now let me come to R console and try to show you all these outcomes directly on the R console, right. So, what I have done here?

(Refer Slide Time: 28:45)

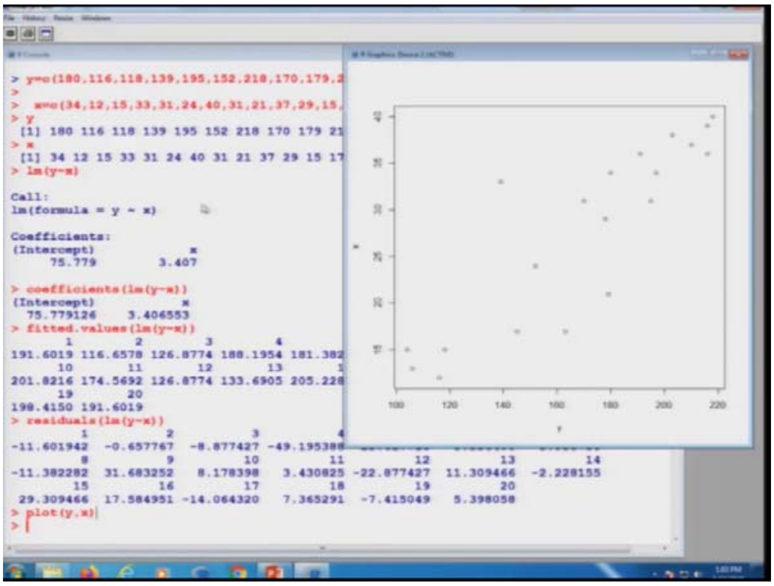


I already have copied this data say this here y and here x this I have already copied here you can see. So, you can see this is the data on y and this is data on here x. Now, I try to fit here a model lm y ~ x and you can see here this is the same thing which we had obtained earlier. Now, there can be some confusion that what is the command to obtain the coefficient so, if so the simple idea is just because try to use the same spelling which has been given here coefficients.

So, if I simply want to obtain here the coefficients I can write down here coefficients and say here lm y ~ x and you can see here you are getting here the values of the coefficients. And similarly, if you want to get the fitted values the command here is fitted dot values and you can see here these are the fitted values.

And if you want to obtain the residuals then the command is residuals and you can see here this is what you have obtained, right. And means if you want to plot this x and y so, you can see here plot y comma x this will give you a graphic here like this one. So, this is the same graph which I shown you here actually you should try to see this is the same graphic which I had obtained, right.

(Refer Slide Time: 30:21)



So, now let me stop here. So, I have shown you that how you can compute whatever quantities we have considered in the lecture in the R software and it is not difficult also. From the same command I will try to show you that we can get different type of information, but for that we need to learn something more and then I will be able to show you more outputs from the same command, right.

So, I will say try to take a small data set yourself and try to practice it try to obtain these different values try to plot them and try to see, whether they are really confirming the same thing which you expect from the data set. I will say take an artificial data set where you know what is happening try to take a data set where x and y are increasing try to take one more data set where you take x and y are decreasing means as the value of x increases y is decreasing.

Then, you try to fit a model then you will see that the value of the $\hat{\beta}_1$ this is b_1 will come out to be negative. So, try to practice and the success to become a good statistician who can produce a very good model is the practice and experience. So, you have practice it, gain experience and I will see you in the next lecture, till then good bye.