

**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 - 47**  
**Multiple Linear Regression Analysis**  
**Model Fitting With R Software**

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 continue on the topic of Multiple Linear Regression Analysis with R Software.

So, you can recall that in the earlier lecture we had considered the estimation of vector of regression coefficient  $\beta$  by the ordinary least square estimation, right. And, after that we also stated the estimator of  $\sigma^2$ , but how to obtain that we will try to do in the further lectures. And after that we had seen how we can obtain the fitted model and how we can obtain the residuals as well as fitted values.

So, before I go further with the estimation of  $\sigma^2$  and other and maximum likelihood estimation etc., let me try to show you whatever we have done that how those things can be executed in the R software. So, what I will try to do? I will try to take a take an example and this is the same example that I discussed in the 1st lecture on multiple linear regression model and I will try to fit a model.

And in order to do it, I am going to use the same command that we discussed in the case of simple linear regression model. So, it is very important that before starting this lecture you must have a quick review of the lecture in the case of simple linear regression model, where we had implemented the command `lm` in R software, ok. So, believing that you have the you have you had revised that lecture, let me begin this lecture now here, ok.

(Refer Slide Time: 02:11)

**Example**

Observations on 20 students are collected

Let

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

$X_1$  : Number of hours per week of study,

$X_2$  : Number of assignments submitted per month,

$X_3$  : Number of hours of play per week

Student no.	$y$	$X_1$	$X_2$	$X_3$
1	180	34	3	15
2	116	12	1	13
3	118	15	3	11
4	139	33	1	10
5	195	31	5	17
6	152	24	1	15
7	218	40	5	18
8	170	31	5	13
9	179	21	2	20
10	210	37	3	19
11	178	29	4	16
12	104	15	1	10
13	145	17	1	16
14	203	38	5	16
15	163	17	1	19
16	216	36	3	20
17	106	13	1	11
18	216	39	5	18
19	191	36	5	15
20	197	34	1	19

So now, this is the same data set where we have collected the information on 20 students about their marks denoted as  $y$ . The number of hours of study in a week, these are denoted by here  $X_1$  and this data is given here see here  $X_1$  and then number of assignment submitted per week this variable is denoted by  $X_2$  and the numbers of hours of play per week which is denoted by  $X_3$ .

And we have obtained here this data. So this data goes like this; that for the student number 1, the student has got 180 marks out of 250 and the student had studied 34 hours in a week. The student has submitted 3 assignments in a month and the student has played 15 hours in a week, and the same information for student 2, student 3 up to student 20, ok.

(Refer Slide Time: 03:13)

```
Model fitting using R:  
Fitting Linear Models  
lm is used to fit linear models.  
Usage  
lm(formula, data, subset, weights, na.action,  
method = "qr", model = TRUE, x = FALSE, y =  
FALSE, qr = TRUE, singular.ok = TRUE, contrasts  
= NULL, offset, ...)
```

So, now first I try to explain you the command. Command is the same, but now you will see here there are some more option, but definitely we are not going to use that option, because this command here lm which is used to fit the linear regression model that is a very general command.

And as I said, that the that there are many more topics in case of linear regression modelling, but we are not going to cover here all the topic, so I will restrict myself only to those option which I am going to use on the basis of the topics which we have covered in this course, right.

So, here once again just like the simple linear regression model, we use the command here lm which is here like this, and then after this there are different options which is here, one is here the formula, and another here is the data and there are many other things, but I am not going to discuss those things.

(Refer Slide Time: 04:13)

**Model fitting using R:**

**Arguments**

**formula**  
an object of class "**formula**" (or one that can be coerced to that class): a symbolic description of the model to be fitted.

$y \sim x_1 + x_2 + \dots$   
 $y \sim x$

**data**  
an optional data frame, list or environment (or object coercible by **as.data.frame** to a data frame) containing the variables in the model. If not found in data, the variables are taken from **environment(formula)**, typically the environment from which **lm** is called.

But, let me explain you the formula and data which are exactly the same what we did what we had done in the case of simple linear regression model. So, this formula will give you will be helpful in providing the model specification to the R software. So, this formula is an object of the class formula. And this has to be denoted using the  $y \sim$  say  $X_1 + X_2 + \dots$  and so on.

You remember that earlier in the case of simple linear regression model you had used the notation  $y \sim X$ , right. Simply, the second option is data, which is an optional data frame where you can input the data and if you do not want to use the data frame then you can input the data from the external means. You can specify the data vectors in your R console and then you try to use the `lm` command, ok.

(Refer Slide Time: 05:12)

**Model fitting using R:**

Details  
Models for `lm` are specified symbolically. A typical model has the form `response ~ terms` where response is the (numeric) response vector and terms is a series of terms which specifies a linear predictor for response.

A formula has an implied intercept term. To remove this use either `y ~ x - 1` or `y ~ 0 + x`.

coefficients  $\rightarrow \hat{\beta} = b$  a named vector of coefficients

residuals  $\rightarrow e$  the residuals, that is response minus fitted values.

fitted.values  $\rightarrow \hat{f}$  the fitted mean values.

anova the analysis of variance

So, now; so, and once you obtain the outcome of an `lm` command then you can exercise different types of option to retrieve information on different aspects. So, as I said, formula is going to be like this, a response variable, whatever is the name of the variable and then this symbol which is on your keyboard this is equivalent sign or  $\sim$  sign and then you have the terms, terms are your independent variable.

So, you have to give it in the terms of say  $X_1 + X_2 + \dots$  and so on, whatever are your variables. You have to just express all the variable with a  $+$  sign, ok. Now, in case if you want to have a model without intercept term then one option is this, you try to write down here  $y$  is equal to  $x - 1$ . So you have to insert the factor  $- 1$  in your formula, or alternative is that you can write down  $0 + x$  also that.

So that will indicate to the R, that in this case the model which you are going to going to find is a model without intercept term, right. And, one thing I can share with you that if there is no intercept term in the model then obviously, there is no intercept term so no  $\beta_0$  and so no  $b_0$ , but the form of the estimator for the slope parameter changes. Well, I have not considered those things here, but those things are there.

Well, in case if you want to retrieve the information on coefficient then you have to use the option `coefficients` and if you want to obtain the residuals you have to use the option here the commands `residual` with `lm` command and if you want to have fitted values, you

can use the command fitted dot values. And this will give you the value of here y hat, this residual will give you the value of e and coefficients will give you the value of  $\hat{\beta}$  which was your here b, right.

After this, there is another option here, anova which is analysis of variance. So at this moment in this lecture I am not going to talk about it. This is related to the test of hypothesis so which I will be doing later on when I start a discussion on the construction of test for test of hypothesis. But here, I am trying to inform you that this command anova, it is used to obtain the or extract the results on the analysis of variants of the given set of data, right.

(Refer Slide Time: 07:59)

**Model fitting using R: Example- data entry**

So we can write the model for each observation,  $n = 20$  as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1, 2, \dots, 20$$

$X1 = c(34, 12, 15, 33, 31, 24, 40, 31, 21, 37, 29, 15, 17, 38, 17, 36, 13, 39, 36, 34)$

$X2 = c(3, 1, 3, 1, 5, 1, 5, 5, 2, 3, 4, 1, 1, 5, 1, 3, 1, 5, 5, 1)$

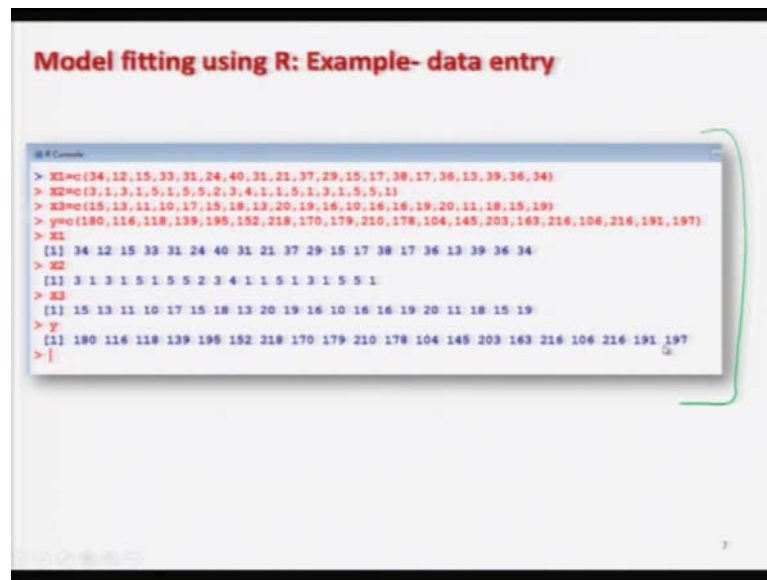
$X3 = c(15, 13, 11, 10, 17, 15, 18, 13, 20, 19, 16, 10, 16, 16, 19, 20, 11, 18, 15, 19)$

$y = c(180, 116, 118, 139, 195, 152, 218, 170, 179, 210, 178, 104, 145, 203, 163, 216, 106, 216, 191, 197)$

So now, let me take our this model. So, the model which we have written, I am considering here the model with intercept term. So I can write down the model here this  $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$  and whatever are the observations which are given for  $X_1$ ,  $X_2$ ,  $X_3$ , they are going to be satisfied for the by the multiple linear regression model. And we have here 20 observations, ok.

So now, I am storing this the values on  $X_1$ ,  $X_2$ ,  $X_3$  in three data vectors,  $X_1$ ,  $X_2$ ,  $X_3$  and again here y you can see here, these are very simple data vectors. And if I am not using the simple data vectors then I can club all this y,  $X_1$ ,  $X_2$ ,  $X_3$  in the framework of a data frame also. So, now you are comfortable in using the concept of data frame.

(Refer Slide Time: 08:53)

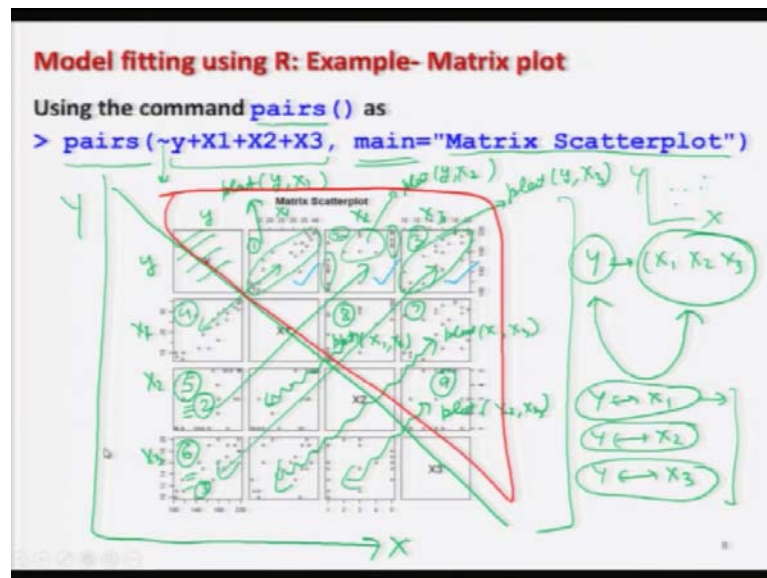


**Model fitting using R: Example- data entry**

```
> X1=c(34,12,15,33,31,24,40,31,21,37,29,15,17,38,17,36,13,39,36,34)
> X2=c(3,1,3,1,5,1,5,5,2,3,4,1,1,5,1,3,1,5,5,1)
> X3=c(15,13,11,10,17,15,18,13,20,19,16,10,16,16,19,20,11,18,15,19)
> y=c(180,116,118,139,195,152,218,170,179,210,178,104,145,203,163,216,106,216,191,197)
> X1
[1] 34 12 15 33 31 24 40 31 21 37 29 15 17 38 17 36 13 39 36 34
> X2
[1] 3 1 3 1 5 1 5 5 2 3 4 1 1 5 1 3 1 5 5 1
> X3
[1] 15 13 11 10 17 15 18 13 20 19 16 10 16 16 19 20 11 18 15 19
> y
[1] 180 116 118 139 195 152 218 170 179 210 178 104 145 203 163 216 106 216 191 197
> |
```

So you can see here, that this is how here I had entered the data in my R console. Well, I will try to show you also here.

(Refer Slide Time: 09:06)



But, before going into the linear regression modelling, the first assumption we have to verify that, whether the data has got a linear relationship or not. Now you see, the model is said to be linear when it is linear with respect to parameter, but our trouble is this I want to judge it in a real life.

And, our problem here is this earlier in the case of simple linear regression model we have only one  $X$  and one  $y$  so I could make here a plot, but now we have a situation here we have here  $y$  and a group of variable  $X_1, X_2, X_3$  and they are just inside a box, so you cannot just see them what is happening, you simply have here the values of  $y$ .

So, but what you want to check? You want to check that whether the relationship between  $y$  and all these variable  $X_1, X_2, X_3$ , is linear or not. So, one possible way out is this, I can check whether the relationship between  $y$  and  $X_1$  is linear or not, the relationship between  $y$  and  $X_2$  is linear or not, and  $y$  and  $X_3$  is linear or not, right. Definitely, I am not trying to check here whether there is a joint relationship, it is a difficult task to do it graphically.

And well, if you can come up with some idea that will be nice, but at this moment what we are assuming that whatever tools we have we can draw only here a graph between two variables. So, I am taking here  $X_1$  and  $y$ , another option is this we can also make some three dimensional graph, but again that will be restricted only to three variables not beyond that.

So, what we try to come up as the solution that, if we try to test or check and if we find that the relationship between  $y$  and with respect to each of the variable  $X_1$ , and  $X_2$ , and  $X_3$ , this is linear then we can conclude that the joint relationship between  $y$  and  $X_1, X_2, X_3$  is also linear although, there can be some interaction effects also, but I do not know how to take the observations on interactions.

So, I have no option here except to believe on or to rely on this idea and try to do as much as I can, right. So that is the reason. So, now what we try to do? That we will try to make a scatter plot between  $y$  and each of these variables, so basically, we will have here three scatter plots.

So, instead of going through the individual scatter plot, I can create a matrix of scatter plots, right. So, the advantage will be that you can view, you can visualize all the scatter diagram in a single shot, in a single diagram, which is more informative, which is more convenient.

So in order to create a matrix scatter plot in R we have a command here, `pairs; p a i r s` and this command has to be used like this, `pairs` and inside the parenthesis you have to



first use the equivalent sign or  $\sim$  sign and you have to give all the variables corresponding to which you want to create the matrix plot.

So, I am trying to choose here,  $y, X_1, X_2, X_3$ . There are four variable, so those four variables are expressed like this  $y + X_1 + X_2 + X_3$ . And, this option here is main this is used only to give the title to the main graphic. So, I am giving here a name Matrix Scatterplot. And, if you try to execute this on the R console, you will get here a picture like this one.

So now, first we understand how to interpret. So you can see here, here there is here a  $y$  and then here there is a  $X_1$  then  $X_2$  then  $X_3$  and the same thing is also here,  $y, X_1, X_2$ , and here  $X_3$ . So you can consider as if here is some X axis and here this is Y axis, it is something like this, right. So now, you come to here first diagram. So, this is the diagram between  $y$  and  $y$ , so which is of not interest. Now, if you come to this one, please try to watch the movement of my pen.

So, if you come to here, let me put it here 1, if you come to block number 1, this is a plot between this is equivalent to plot between  $y$  and  $X_1$ , right. So, you can see here possibly you can see here well there is a good linear trend, right, and it is not very difficult. Well, now then you come to plot block number 2 here, this is actually the plot between  $y$  and  $X_2$ . You can see here, it looks little bit haphazard, means some points are here, some points are here, some points are here, right.

So, well I can assume linear relationship but I am not really confident, ok. So, let us move forward. Now, you come to the block number here 3. So, this is the plot between  $y$  and  $X_3$ . You can see here, well there is a reasonable linear plot, it is not that bad, right. So now, before I try to conclude let me try to give you the interpretation of all other blocks also.

Now, if you try to make here a diagonal then this matrix plot is a symmetric. This means, this here if you look at the block number 1 and block number here 4, so this is the same thing. The only difference is that the axis  $y$  and  $X$  are interchanged, that is all, right. So, that is the reason, and then whatever you are looking here in the block number 5, this is same as block number 2 plot.

You can see here only the direction is changed, means instead of using the command plot  $X y$ , I am using the command plot  $y X$ . And similarly, if you try to see here in the block number 6, this is the same as block number 3, because they are symmetric. This is symmetric, this is symmetric, right, ok.

Now, if you try to come here block number here see here 7, so what is this thing? This is a plot between see here,  $X_1$  and  $X_3$ . And, if you come to here plot block number 8 this is the plot between  $X_1$  and  $X_2$ , right. And, if you come to here block number here 3, so this is the plot between see here  $X_2$  and  $X_3$ .

So, although it is not our interest in looking the relationship among  $X_1$ ,  $X_2$ ,  $X_3$  because at this moment, we are primarily interested in the relationship of  $y$  with respect to  $X_1$ ,  $X_2$ ,  $X_3$ , right. But still, these three plots in block number 7, 8, 9, they will try to give you an information about the independence of the explanatory variable, right.

So, well, I am not discussing here, but looking at this graphic that will give you a firsthand information about the multi collinearity problem in the data. And after that if you try to see here, this means these two graphics they are means again the same only the exchange of axis over here.

So, you can see here that this is the symmetric plot. So essentially, in a matrix scatter plot you have to look only into the upper part of the diagram which is this one, that is all. Either upper one or lower one also, means you have to choose only one thing and the and all other information can be followed exactly.

Now, if you try to see if you try to look into the blocks number here, 1, 2 and 3, so what we can see here that the linear relationship between  $y$  and  $X_1$  and  $y$  and  $X_3$  is 100 percent clear, between  $y$  and  $X_2$  we are not 100 percent sure, but we have no other options so, because there can be an interaction effect also there can be other problems also so, but for a while we assume that, ok this is reasonably linear and we move forward.

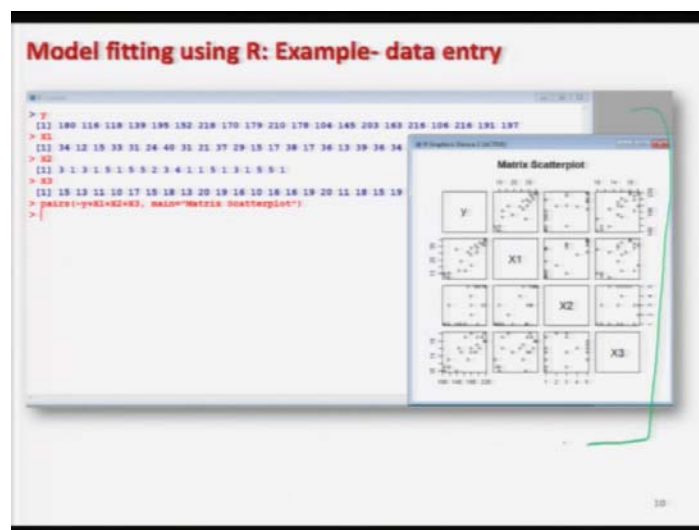
(Refer Slide Time: 18:22)



Then we will try to see whether we are getting a good model or not. If there is any problem then we have to come back and then we have to modify our variable, we have to use some transformation of variable and then we can move forward, ok. So, you can see here, this is the I mean the same scatter plot which we have obtained, but now in a much bigger way.

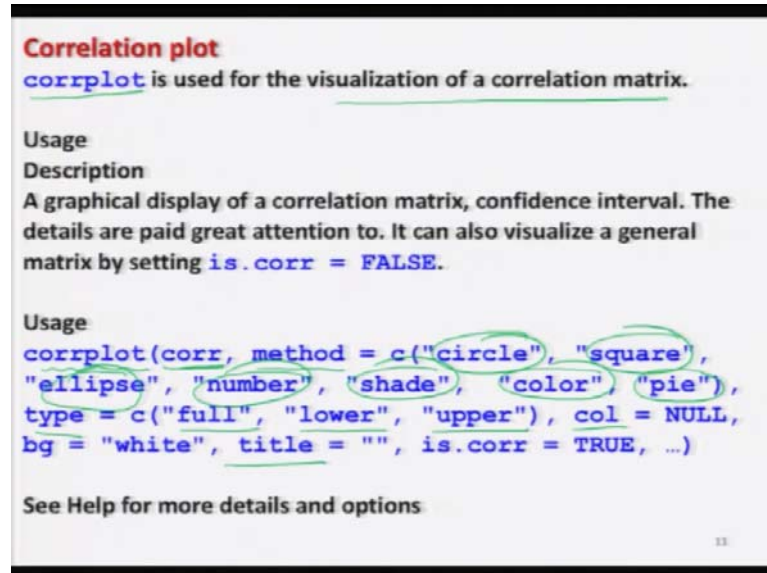
So you can now have a look, I have given it because in the earlier one I had made lots of this marks, so I have I am reproducing it here again once again for you. So that now you can look at this graphic and you can see what you can really infer for different type of relationship, ok.

(Refer Slide Time: 18:53)



So, now this is the scatter diagram so that you can be confident that I have reproduced the same thing which you will obtain, ok.

(Refer Slide Time: 19:02)



**Correlation plot**  
`corrplot` is used for the visualization of a correlation matrix.

**Usage**  
**Description**  
A graphical display of a correlation matrix, confidence interval. The details are paid great attention to. It can also visualize a general matrix by setting `is.corr = FALSE`.

**Usage**  
`corrplot(corr, method = c("circle", "square", "ellipse", "number", "shade", "color", "pie"), type = c("full", "lower", "upper"), col = NULL, bg = "white", title = "", is.corr = TRUE, ...)`

See Help for more details and options

11

Now, there is another option to check this relationship. And if you remember, in the beginning of the lecture we had used the option correlation plot and I had explained you in detail even I introduced some basic concept about the R software, that how to construct the correlation plot.

So, I will not go into that detail, but I will, but if you have forgotten it, I will say that go back to your earlier lectures and try to see how you had obtained, how you how you had interpret the correlation plot. And now I will try to make this plot in the context of our data. So, just for a quick review; so this correlation plot is obtained by the command corr plot which is used for the visualization of the correlation matrix, right.

So, the main thing is this; main commands which we will which we had used earlier and we will use here again, these are the main command is corrplot and then there are different option which is here corr; corr is the matrix or the correlation matrix that you have to give, which you want to plot.

Then you have to give the method means, method means the correlation structure and the degree of correlation coefficient is going to be indicated by a circle, by a square, by an ellipse or by an number or by some shade or colour or pie diagram. And similarly, you

have one more option here type, type is to show whether you want a full matrix or a lower diagonal matrix or an upper diagonal matrix using the command, full, lower and upper respectively.

Similarly, if you want to change the colour, you have the command col and so on if you want to give the title etcetera. So, there are many many commands, I will request you to go to help menu and see all the things. But here, my objective is to create a plot for the given set of data.

(Refer Slide Time: 20:50)

**Correlation plot**  
 Creating a correlation plot  
 corrplot is used to fit linear models.  
 First install the package corrplot and load it.  
`install.packages("corrplot")`  
`library(corrplot)`  
`X123y=data.frame(y,X1,X2,X3) #Data frame creation`  
 Creates a correlation matrix  
`X123y_cor = cor(X123y) # Creates correlation matrix`  
 Creates a correlation plot  
`corrplot(X123y_cor, method = "number")`

*Handwritten notes:*  
 Correlation matrix,  $(X_1, X_2)$   
 $\rho = \text{Corr}(X_1, X_2) = \lambda_{X_1, X_2} \in [-1, 1]$   
 Correlation matrix,  $(X_1, X_2, X_3)$   
 $\begin{matrix} X_1 & X_2 & X_3 \\ \lambda_{X_1, X_1} = 1 & \lambda_{X_1, X_2} & \lambda_{X_1, X_3} \\ X_2 & \lambda_{X_2, X_1} & \lambda_{X_2, X_2} = 1 \\ X_3 & \lambda_{X_3, X_1} & \lambda_{X_3, X_2} & \lambda_{X_3, X_3} = 1 \end{matrix}$

So, as we discussed earlier in the lecture, in order to construct this correlation plot you need to install a package corrplot. So, we first we install the package corrplot using this command; install dot packages and then we load it using the command, library. Now, in this correlation plot the data has to be entered into the framework of a data frame.

So, what I try to do, I have got here a get a set of four variables, y, X1, X2, X3 so I try to create a data frame here and I give the name as X 1 2 3 y so that you can remember it is X1, X2, X3 and y, right. And then, this data frame has to be converted into a correlation matrix. So, before I go further, let me explain you what is the correlation matrix, right.

Suppose, if I say I have got here two variables say, X and say here X1 and X2 suppose, then the correlation coefficient between X1, and X2 let me denote it by here  $r_{X_1, X_2}$ , right. So, now, suppose I have got here three variables, X1, X2, and X3, and suppose I want to create a correlation matrix. So the correlation matrix of this X1, X2, X3 will be a matrix

like this. The 1st element will be the correlation between  $X_1$  and  $X_1$ . So you can write down actually here that is more easy to understand,  $X_1, X_2, X_3, X_1, X_2, X_3$ .

So, the first entry will be the correlation coefficient between  $X_1$  and  $X_1$ , which is equal to here 1. So, and similarly, the second diagonal element will be  $r_{X_2, X_2}$  that is the correlation coefficient between  $X_2$  and  $X_2$  which is 1. And the third diagonal will be the correlation coefficient between  $X_3$  and  $X_3$  which will again be equal to 1, right.

Now, if you come to this other elements suppose here, if you try to see, so this is the correlation coefficient between  $X_1$ , and  $X_2$ , and the next element will be correlation coefficient between  $X_1$  and  $X_3$ . Similarly, here is the column which is going to be a symmetric matrix, because the correlation coefficient between  $X_1$  and  $X_2$  is the same as the correlation coefficient between  $X_2$  and  $X_1$ .

So, I can write down here correlation between  $X_2$  and  $X_1$  and similarly, the 3rd row will be 3rd row 1st element will be correlation coefficient  $X_3$  and  $X_1$ . So,  $X_3$  is coming from here and  $X_1$  is coming from here, you can see. Similarly, here the last element in the 2nd row will be the correlation coefficient between  $X_2$  and  $X_3$  and similarly here in the 3rd row and 3rd column this will become here  $r_{X_3, X_2}$ , right.

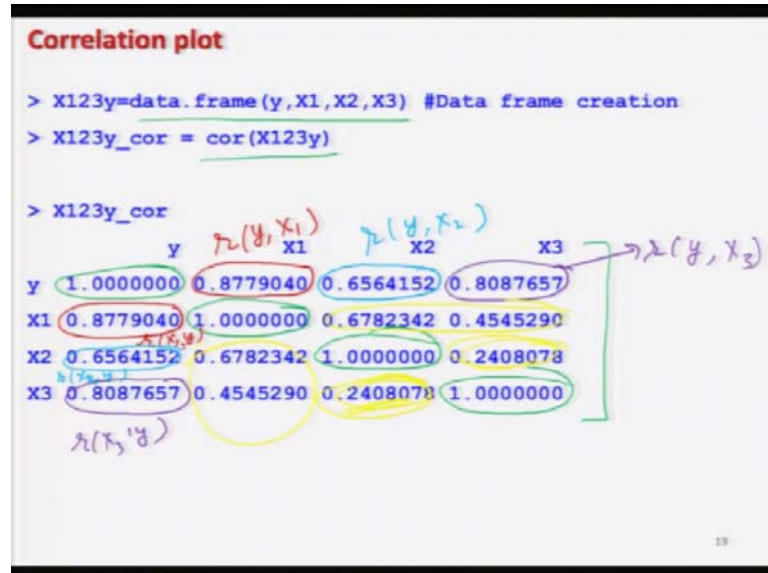
So you can see here now, this correlation matrix is going to be a matrix where all the correlation coefficient are presented together and all the values of correlation coefficient will be lying between 0 and 1. Now we are trying to create a matrix plot of this correlation coefficient values from this matrix. And we try to make it more beautiful more informative for that we use different types of option, ok.

So, first I try to create here my correlation matrix. The correlation matrix can be created using the same command `cor`. So, if you try to give here a scalar it will give you scalar value, but if you try to give here a say this data frame then it will give you a matrix value, ok. So, I try to first I try to create this correlation matrix and I store it under `X_1_2_3_y_underscore_cor`. So, means you can imagine that you can remember that this is the matrix of  $X_1, X_2, X_3$  and  $y$  for correlation coefficient, ok.

Now, the correlation plot is created by the command `corrplot`; `corrplot`, first you have to give the name of the correlation matrix remember, not the data frame. So, you try to give

it a correlation matrix and then you have to give here the method number or whatever you want to choose, I will try to take different options.

(Refer Slide Time: 25:25)



Now, if you try to see, this is how you are going to obtain. First, I try to create the data frame of y, X<sub>1</sub>, X<sub>2</sub>, X<sub>3</sub> and then I try to create the correlation matrix and this correlation matrix will look like this. So, you can see here this is your here the correlation coefficient between y and y this is the diagonal elements correlation coefficient between X<sub>1</sub> and X<sub>1</sub> and similarly, for X<sub>2</sub> and X<sub>3</sub> here.

Now, if you try to look on the non diagonal elements, so obviously, the diagonal elements are always going to be 1 and we are not interested in those things because these are obvious values. The more important values are on the off diagonal parts. And, if you try to look in the first value I am trying to use here different colours of pen.

So, this is here 0.8779040, this is the value of the correlation coefficient between y and X<sub>1</sub>. And this is the same value if you try to see here, in the 1st column 2nd row, this is the same value this is the same value, but this is the correlation coefficient between X<sub>1</sub> and y, but that does not make any difference. They are the same value.

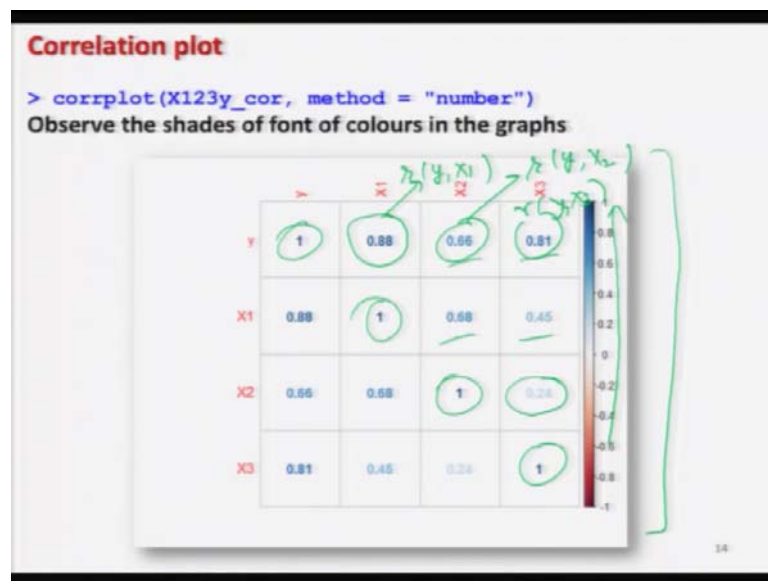
Similarly, if you try to go for other values for example, if you try to look at this value this is the correlation coefficient between y and X<sub>2</sub>. And similarly here, in the 1st column, this is the value here of correlation coefficient between X<sub>2</sub> and y, right. So, they

are again the same values. And similarly, if you try to go for the third value here, this value is the correlation coefficient between  $y$  and  $X_3$ . And this is the same value here correlation coefficient between, but  $X_3$  and  $y$ , right.

So, these are the values of interest, and similarly you can see here these values here, which are here like this they are the here like this. They are the correlation coefficient ah, but they are the correlation coefficient between  $X_1$ ,  $X_2$ ,  $X_3$  and so on. But since they are not our interest so we are not going into the those details.

But you can see here, that ah here in this case possibly the values are quite small, but in other cases there is some correlation, but that is going to happen in most of the real life situation. Particularly, when you are trying to deal in the data sciences where there are millions and billions of observations. So, that is what we have to learn that how you are going to interpret and look into the data.

(Refer Slide Time: 27:56)



Now, the outcome of this matrix plot will look like this, you can see here. So, on the diagonal elements we have here the values. And, you can see here, now try to observe on two things; one is the value and another is the shade of the colour. You can see here, here the value is 0.24, but the colour is quite low, right. Colour is not dark as compared to this value here 0.88; 0.88 you see the values also high and the shade of the colour is also darker.

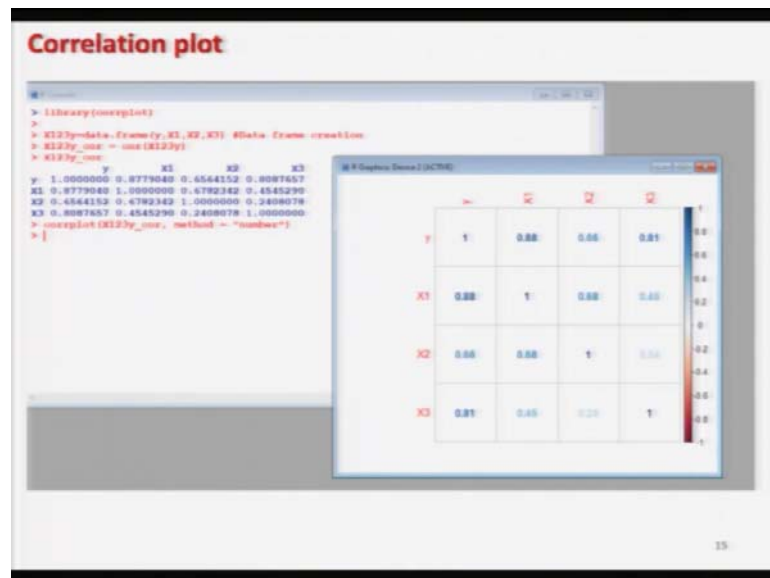


So now, what is the meaning? The shade of the colour is going to indicate the degree of the correlation. So, you can see here you can see here, these values and you can see here they have got the different shades. It is not like that in the recording you are ah looking at this shades and they are looking to be faded, no. Degree of shade is proportional to the magnitude of the number. So, lower the number, lower the shade, higher the number, darker the shade.

And this scale is also given here you can see, right. So, this is your correlation matrix plot for example, this 0.888, it is trying to give you the correlation coefficient between y and X<sub>1</sub>. This is trying to give you the correlation coefficient between y and X<sub>2</sub> and 0.81 is the correlation coefficient between y and X<sub>3</sub> and so on, right.

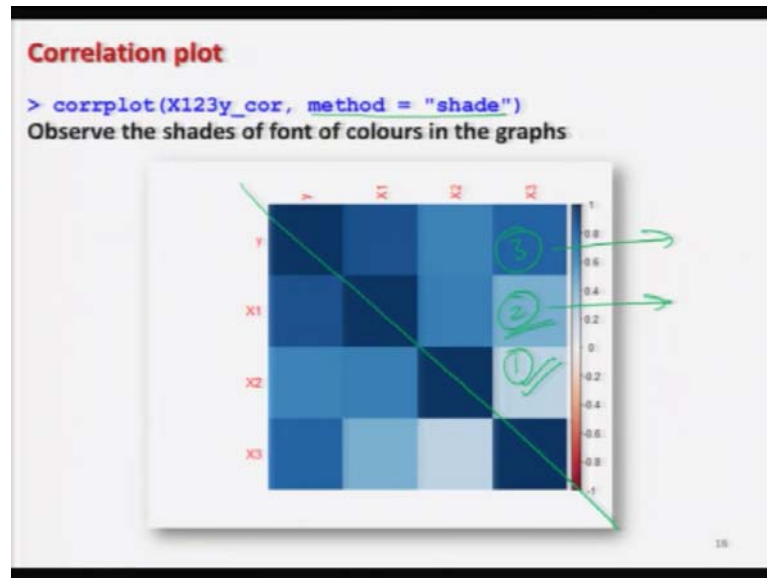
So, this is how you can by looking at this picture at a glance, you can very easily see what is the nature of degree of correlation coefficient and the degree of linear relationship. And this is going to give you much better information in terms of shade of the colour as well as magnitude both are embedded together.

(Refer Slide Time: 29:49)



, right. And this is the screenshot of the same outcome, ok.

(Refer Slide Time: 29:54)

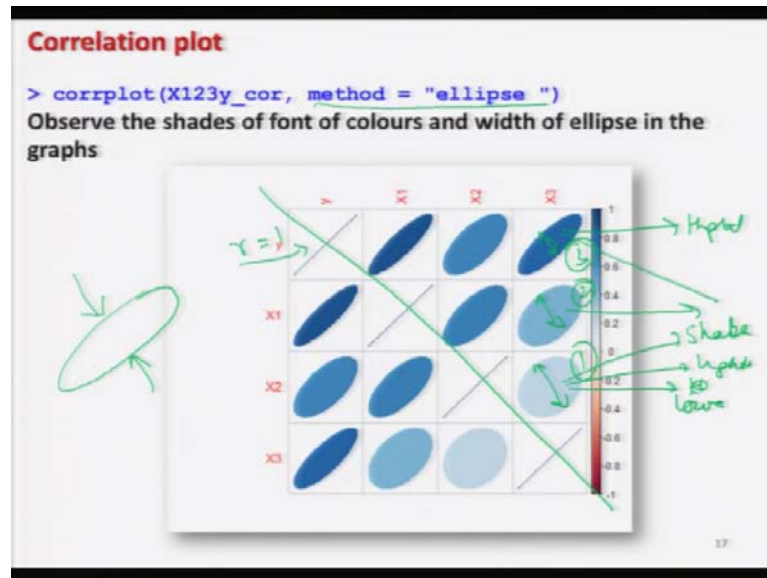


Now, means other option is this, instead of using the option method equal to number, I can use here method equal to shade. So, in this case you can see here that the degree of shade is proportional to the degree of correlation, but here you will not get the magnitude of the correlation coefficient.

So, I can make here a very simple rule that the degree of correlation coefficient is proportional to the degree of shade of the colour, right. For example, you can see here on the diagonal element it is the darkest, because the correlation coefficient is always 1. So that is the maximum value.

But, you can see if you try to look over here, compared to compare block number here 1, block number here 2 and block number here 3, you can see here that the degree of shade is changing, so the correlation coefficient in the block 1 is the lowest. Then followed by the correlation coefficient in the block number 2 and then the block number 3 has the say highest correlation coefficient among this block number 1 2 and 3, right.

(Refer Slide Time: 30:59)



Similarly, if you try to take here say, this another option here method equal to ellipse. So, in this case what is happening? Those numbers are replaced by or those squares are replaced by the form of an ellipse. So, if you want to understand it, you can first look into the diagonal elements.

So, the diagonal elements you have got here a simple line. What does this mean? If you try to take here an ellipse and try to press it from the upper and lower side by your hand then it is something like this. You take a ball; ball in your hand something like this and try to push from both the sides like this. So, what will happen? Suppose, I have here sort of here ball and I am trying to put my hands here and I am trying to press it.

So what will happen? This ball will become flatter. It will look like an ellipse and if you try to push more force or if you try to push the ball with more force it will become practically flat. So that is the maximum what you can achieve by pressing the ball. This is the same thing which is indicated in this slide also, ok. So, this is what is indicating here, that the corr and this width of the slide is indicating the correlation coefficient.

So, on the diagonal element the correlation coefficient is 1, so the width is going to be nearly very small. So, the rule is this, this width of the ellipse that is inversely proportional to the correlation coefficient. Higher the value of correlation coefficient; that means, the width is going to be lower.

So, in this case if you try to look into figure number 1 2 and 3 here, you can see here, 1 has the maximum width, so the correlation coefficient is lowest and then the this 2nd one, block number 2, this has got some width and this block number 3 has got this width. So, you can see here block number 3 has got the lowest width among this 1, 2 and 3.

So, the correlation coefficient of this one is going to be the lowest and the correlation coefficient of this one is going to be the highest among block number 1 2 and 3. And, the shade also you can see here the shade is the, if we try to see a shade; shade here shade here. So, you can see here, here the shade is lighter then block number 2 and block number 3.

So, if your, so if your magnitude is less than your shade is also lighter. So by looking at the width of the ellipse and the shade of the ellipse, colour of the ellipse, you can have a fair idea that how the correlations are present in the variable  $X_1, X_2, X_3$ .

(Refer Slide Time: 33:53)

**Model fitting using R: Example- Finding OLSE**

So we can write the model for each observation,  $n = 20$  as:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \quad i = 1, 2, \dots, 20$$

> `lm(y ~ X1 + X2 + X3)`  $lm(y \sim X)$

Call:  
`lm(formula = y ~ X1 + X2 + X3)`  $y$  is to be regressed on  $X_1, X_2, X_3$

Coefficients:

(Intercept)	X1	X2	X3
8.769	1.997	3.918	6.106
$\beta_0 = b_0$	$\beta_1 = b_1$	$\beta_2 = b_2$	$\beta_3 = b_3$

The fitted model is

$$y = 8.769 + 1.997X_1 + 3.918X_2 + 6.106X_3$$

$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix}$

So now, we are confident that, ok we take a call that, ok our relationship between  $y$  and  $X_1, X_2, X_3$  is approximately linear and we will try to go with the linear model. So in order to fit this model, we  $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$  based on 20 observation, we are going to use the same command that we use in the case of simple linear regression model, but I am trying to extend it.

So the command is the same here. Earlier you had written `lm y ~ X`, now I am writing it here  $y$  and now my independent variables are  $X_1, X_2, X_3$ . So I am trying to write all the

variables which are joined by the + sign, right. So, I am writing  $X_1 + X_2 + X_3$  and this equivalent or  $\sim$  sign. So, this is going to indicate that  $y$  is to be regressed on  $X_1, X_2$  and  $X_3$ , right.

So now, if you try to execute it you get here this type of outcome. So, you can see here, what are you getting here? The first value here is intercept term, so it is trying to give you the value of  $\hat{\beta}_0$ , which is you are indicating by  $b_0$ . Thus the value corresponding to variable  $X_1$  it is 1.997, this is the value of  $\hat{\beta}_1$  which you are indicating by  $b_1$ . So, and then the value corresponding to  $X_2$  is 3.918 which is the value of here  $\hat{\beta}_2$ , which is here indicated by  $b_2$ .

And,  $X_3$  has a value 6.106 which is the value of  $\hat{\beta}_3$ , which you are trying to indicate by  $b_3$ . So, what is happening here? You had the coefficient vector  $\beta$ , as  $\beta_0, \beta_1, \beta_2$  and  $\beta_3$  and these values have been obtained by the ordinary least square estimate  $b$ , whose values are given by  $b_0, b_1, b_2$  and  $b_3$ , right.

So now based on that, your fitted model will become  $y$  is equal to, ok. So now the fitted model is  $y$  is equal to this  $b_0$  coming from here +  $b_1$  coming from here  $X_1$  +  $b_2$  coming from here  $X_2$  +  $b_3 X_3$ . So, this is our fitted model which we can obtain from the this outcome, right.

(Refer Slide Time: 36:39)

**Model fitting using R: Example- Finding OLSE**  
 To find the OLSE, use command `coefficients`

```
> coefficients(lm(y~X1+X2+X3))
(Intercept)      X1      X2      X3
  8.769453    1.996746    3.918402    6.106034
```

The OLSE is

$$b = (X'X)^{-1}X'y = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 8.769 \\ 1.997 \\ 3.918 \\ 6.106 \end{pmatrix}$$

So, this is exactly what I have rewritten here, that your ordinary least square estimator is given by like this. So now, these are the values of  $b_0, b_1, b_2, b_3$  which we have obtained here like this. So now, at least I have finished my first part that looking at this outcome you know that how these values have been obtained. They are obtained using this expression. And, what are these values? These are these values which are the estimated values for  $b_0, b_1, b_2$  and  $b_3$ .

So, now you understand that is what is happening inside this R software.

(Refer Slide Time: 37:17)

```

Model fitting using R: Example- Finding OLSE

# R Console
> lm(y~X1+X2+X3)

Call:
lm(formula = y ~ X1 + X2 + X3)

Coefficients:
(Intercept)      X1      X2      X3
  8.769      1.997      3.918      6.106

> coefficients(lm(y~X1+X2+X3))
(Intercept)      X1      X2      X3
 8.769453      1.996746      3.918402      6.106034
  
```

And this is the screenshot what I did.

(Refer Slide Time: 37:22)

Model fitting using R: Example- Fitted values

Recall fitted values  $\hat{y} = Xb$ .

The fitted values are found by using the command `fitted.values(lm(y~X1+X2+X3))` *lm(y~X)*

```

> fitted.values(lm(y~X1+X2+X3))
  1      2      3      4      5
180.0045 116.0273 117.6422 139.6408 194.0632
  6      7      8      9     10
152.2003 218.1399 169.6390 180.6586 210.4189
 11     12     13     14     15
180.0452 103.6994 144.3291 201.9344 162.6472
 16     17     18     19     20
214.5282 105.8119 216.1432 191.8348 196.5919
  
```

And, after this exactly going on the same line as we did in the case of simple linear regression model, we will try to find out the fitted values as well as residuals. So, the fitted values can be obtained by here the expression  $\hat{y} = Xb$ .

So, in case of multiple linear regression model you have to use the same command. That, fitted dot values and inside the parenthesis. Earlier, we had used `lm(y ~ X)`, now I am using here the new object `y ~ X1 + X2 + X3`.

And if you try to do it here, you have this outcome. So, you can see here the 1<sup>st</sup> value, it is the value of  $\hat{y}_1$  which is 180.0045, 2nd value here is the value of  $\hat{y}_2$ , and similarly here this 20<sup>th</sup> value is the value of  $\hat{y}_{20}$ , right. So, you can see here that it is very easy to obtain the fitted values even in the case of multiple linear regression model. And now, you can compare these fitted values from your observed values of  $y$ .

So, it is something like this, if you try to use the values of  $X_1, X_2, X_3$  for the first set of observation in the fitted model then you will obtain the value of  $y$  as here like this, which is  $\hat{y}_1$ . And similarly, if you try to use the second set of observation on  $X_1, X_2, X_3$  and if you try to substitute those values in the fitted model, you will get here the value 116.0273 which is  $\hat{y}_2$  and so on, right.

(Refer Slide Time: 39:01)

```

Model fitting using R: Example- Residual vector

R Console
> lm(y~X1+X2+X3)

Call:
lm(formula = y ~ X1 + X2 + X3)

Coefficients:
(Intercept)      X1      X2      X3
      8.769      1.997      3.918      6.106

> residuals(lm(y~X1+X2+X3))
      1      2      3      4      5
-0.004535796 -0.027252133  0.357774992 -0.640814986  0.936830147
      6      7      8      9     10
-0.200272488 -0.139917983  0.360967235 -1.658607628 -0.418910837
     11     12     13     14     15
-2.045241880  0.300612730  0.670915130  1.065642529  0.352812314
     16     17     18     19     20
 1.471800875  0.188070427 -0.143171999 -0.834831230  0.408130581
> |
  
```

And this is the screenshot of the same outcome.

(Refer Slide Time: 39:06)

**Model fitting using R: Example- Residual vector**

Recall residual vector  $e = y - \hat{y} = y - Xb$

$\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$  observed fitted

The residual vector is found by using the command

`residuals(lm(y~X1+X2+X3))`

$e_1 = y_1 - \hat{y}_1$   
 $e_2 = y_2 - \hat{y}_2$

```
> residuals(lm(y~X1+X2+X3))
  e1  1    e2  2    e3  3          4          5
-0.004535796 -0.027252133  0.357774992 -0.640814986  0.936830147
  6          7          8          9         10
-0.200272488 -0.139917983  0.360967235 -1.658607628 -0.418910837
 11         12         13         14         15
-2.045241880  0.300612730  0.670915130  1.065642529  0.352812314
 16         17         18         19         20
 1.471800875  0.188070427 -0.143171999 -0.834831230  0.408130581
```

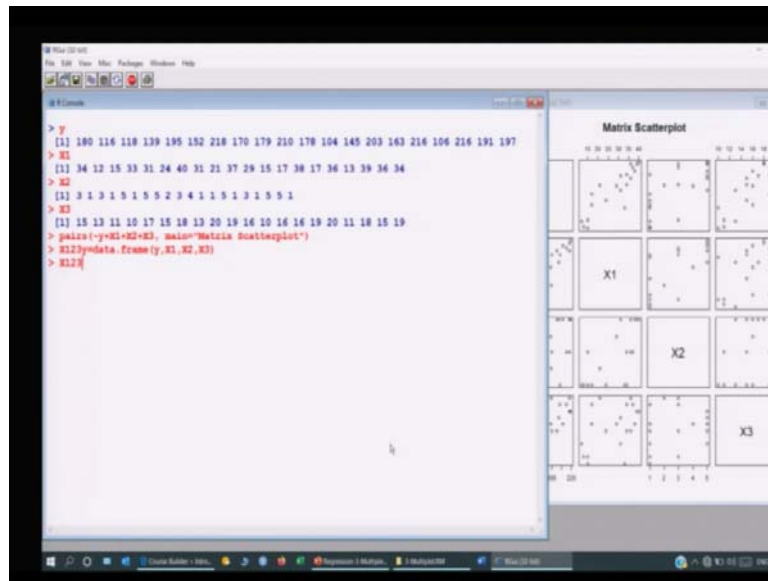
And now I try to find out the residual vector. So, you can see here, now in this case  $e$  is a vector  $e_1 e_2$  here  $e_{20}$ , because we have 20 observation. So, now here, this  $y$  which is here observed and  $\hat{y}$  here is fitted. And if you try to take the difference, you will get this value. So now, in order to obtain the residual vectors we have to use the command residuals.

And then inside the parenthesis you have to same you have to use the same object  $\text{lm}(y \sim X1 + X2 + X3)$  and you will get here this type of outcome. So you can see here, we have the similar thing. The first value here is the value of  $e_1$ , that is the first residual, second value is the value is the second residual, third value is the third residual and the last value 20 here is the last residuals. So, the first residual is  $y_1 - \hat{y}_1$  and  $e_2$  here is  $y_2 - \hat{y}_2$  and so on, right.

So, you can see here it is now difficult to obtain this residuals, right. So, let me first try to show you all these things on the R console also. And this is here the screenshot.



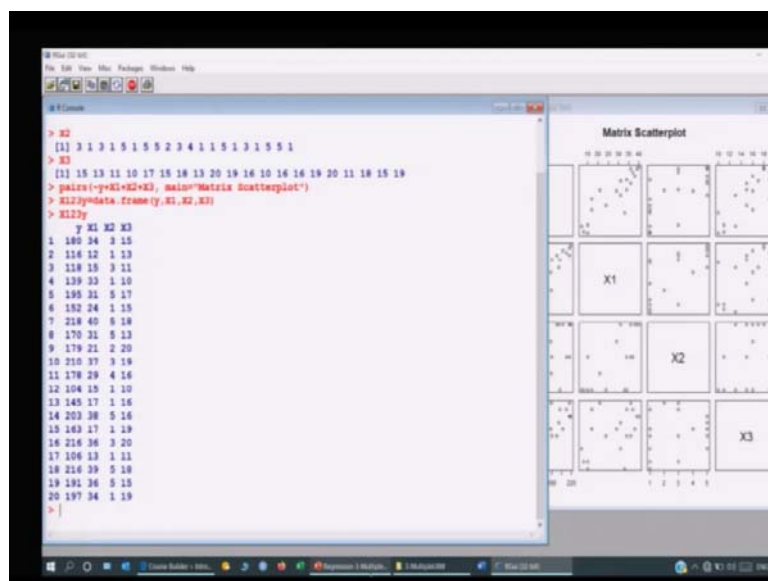
(Refer Slide Time: 40:30)



So, you can see here, I already have entered the value here y of y, X<sub>1</sub>, X<sub>2</sub>, and X<sub>3</sub>. So, first I try to create here the matrix scatter plot. You can see here this comes out to be like this. So this is the same screen plot which you have obtained, right. Now, after this if you want to create this correlation plot, so I will use the first command here, that first you have to create the data frame.

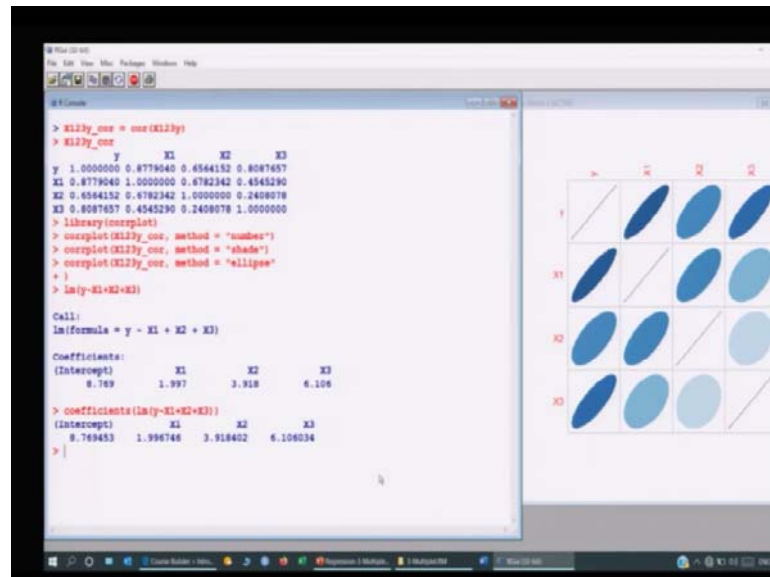
So, I try to create here the data frame you can see. So you can see here, this is your data frame, right same data set which you had obtained.

(Refer Slide Time: 41:15)



Now, you try to find out the correlation matrix of this data set. So I clear the screen.

(Refer Slide Time: 41:28)



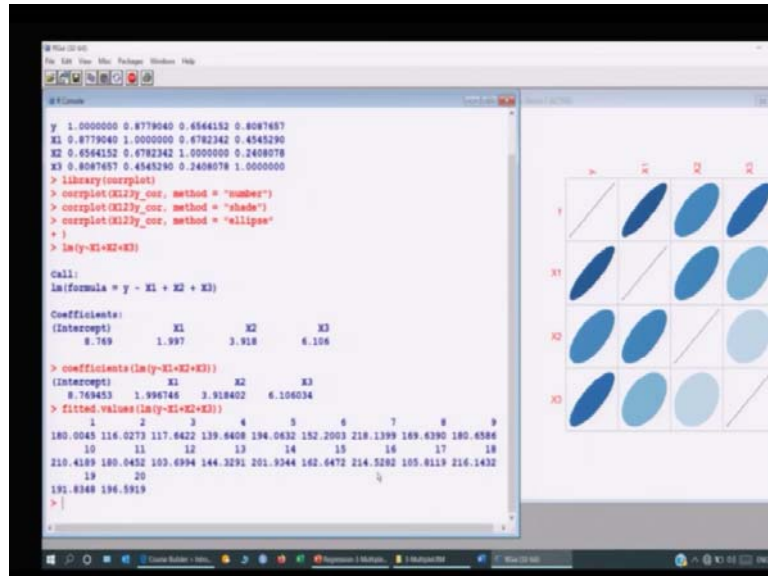
It will be here like this. So, this the outcome will here look like this, you can see here. Then, after this I would like to create the correlation plot. So, I already have installed this package corrplot on my computer. So, you can see here this is loaded, and now if you try to use here the command for this here corrplot for the number, you can see here you get here this outcome.

You can see here now very clearly that how the shades are changing. And if you want to use here the method equal to here shade, so you have to simply use this command on the R console, you can see here you are getting here the shade, right. And similarly, if you want to have here ellipse, so what you have to do? You simply have to give here command here ellipse, right.

So, this is the same outcome which I shown you on the slides and after that if you want to fit here a model, so you can see here this fitting can be using the command lm. You can see here this is the same outcome which you have obtained here, right. Just by writing the command  $lm(y \sim X1 + X2 + X3)$ , and if you want to retrieve the coefficients from this outcome you simply have to use the command coefficients and you will get here only the value of the coefficient.

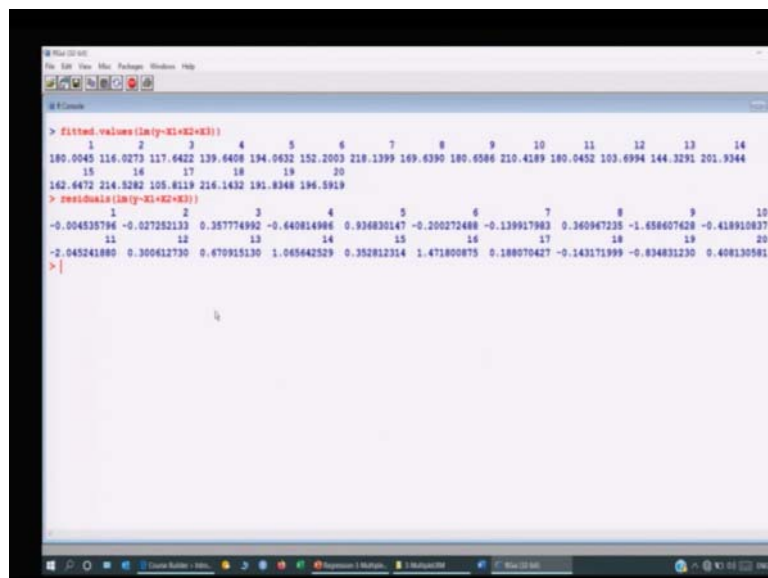
And similarly, if you want to obtain the fitted values of this output then you have to simply use the command, fitted dot values.

(Refer Slide Time: 43:23)



And, inside the parenthesis the object l m, you can see here you have got these many values. Well, if I try to make my this slide bigger and if I try to just show you once again.

(Refer Slide Time: 43:37)



You can see here this is like this. So, there are 20 values, which are trying to give me the values of  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{20}$ . And now, if you want to find out the residuals, you can see

here you can use the command residuals and inside the parenthesis you have to write the l m object or the formula which you have used for fitting the model. And you will see here you can get here the.

So, you can see here it is not difficult at all, right. So now you have seen that it is not difficult at all to find out those values which you have obtained in theory on the basis of a data set inside the R software. I am sure that this will remove the fear from your heart that sometime because of matrix theory and other things people may think that the things are very complicated but they are not.

Using them is very simple, straight forward, but the main challenge is how to interpret the value. This is something like this, when a doctor looks in the X-ray or an MRI or an ECG, based on that the person has to determine whether his patient is healthy or not or if there is any problem or not, right.

Now, if somebody knows only to look into the ECG or this X-ray report, but if the doctor does not understand what is happening correspondingly inside the body, do you think that will you go to that doctor once again, who do not know by looking at the X-ray that what is what has happened inside the body? I am sure you will not go to that doctor one means anymore.

Same thing is with the statistician also or as a data scientist also. If you just know how to interpret the data, but you do not understand what is happening what is happening behind the data, what is happening inside the data, you cannot be a data scientist. You will simply be a compounder.

So, decide what you want to be now. If you practice, if you learn the software outcome, if you learn the theoretical construction, I am not asking you to have the proofs always, but you have to understand the construction, how the tools have been constructed and whatever is the software outcome, what are the values which are being computed in the software outcome. That is my experience. Your experience may be different.

But in my opinion it is not possible to become a good statistician or a good data scientist without having a one to one correspondence between the outcome of the software and the theoretical construct or the basic concept behind that value. So, you try to take some simple examples, just try to take a data set of 5 values, 10 values.

Try to create it yourself and try to see whatever issue you are trying to create inside the data is this getting reflected in your model. You can take a data by just creating yourself where the curve is non-linear and then try to fit a linear regression model and see what happens. Well, in the next lectures I will be going more into the details and then all other aspects will come into picture.

So, till then you practice and I will see you in the next week, goodbye.