

Descriptive Statistics with R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture- 29
Association of Variables - Correlation Coefficient using R Software

Welcome to the next lecture on the course Descriptive Statistics with R software. You may recall that in the earlier lecture we started a discussion on the concept of association between two continuous variable and we learned about the Correlation Coefficient. And we also have understood now that how to interpret the values of correlation coefficient with respect to the magnitude and direction of the association.

Now, in this lecture I am going to demonstrate that how you are going to compute the value of correlation coefficient using the R software and how you are going to implement it, how you are going to use it when you get a data set. So, first just a quick review what we had done earlier.

(Refer Slide Time: 01:10)

Covariance
X, Y : Two variables
n pairs of observations are available as $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
The covariance between the variables x and y is defined as

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

where

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

Similar definition is available for grouped data in frequency table.

2

So, you may recall that we had discussed the concept of covariance between the two variables X and Y and for which we had obtained the n pairs of observations as $x_1, y_1, x_2, y_2, \dots, x_n, y_n$ based on that we have computed the covariance between x and y as like this and this was for the ungroup data and similar definitions fall for the group data also.

(Refer Slide Time: 01:37)

Covariance

R command:

x, y : Two data vectors

$\text{cov}(x, y)$: covariance between x and y .

Command $\text{cov}(x, y)$ calculates the covariance with divisor $(n - 1)$

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

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

And we also had discussed that if I have two data vectors in R software which I denoted as x and y then the command cov that is covariance between x and y . So, you write cov and inside the argument write the data vectors this will give us the value of covariance. But remember one thing this command covariance between x and y in R software is going to give you the value of covariance in which this divisor is 1 upon n minus 1. And whereas, we had defined the covariance as 1 i goes from 1 to n x_i minus \bar{x} y_i minus \bar{y} which has the divisor n .

(Refer Slide Time: 02:18)

Coefficient of Correlation

Also called as **Karl Pearson Coefficient of Correlation**

$$r \equiv r(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}}$$

$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ > 0
 < 0

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

> 0 $< +$
 > 0

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

The sign of Cov. coeff. is determined by the Covariance.

So, after this we had define the correlation coefficient this was defined as here say r and finally, we had obtained the expressions here like this a simplified version of this correlation coefficient and after that we had understood that how to interpret the values and the sign.

(Refer Slide Time: 02:36)

Coefficient of Correlation

R Command

`cor(x, y)` computes the correlation between x and y

`cor(x, y, use = "everything", method = c("pearson", "kendall", "spearman"))`

Karl Pearson
Spearman's Rank Correlation Coefficient

x: a numeric vector, matrix or data frame.
y: a numeric vector, matrix or data frame with compatible dimensions to x.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Now, the next question is how to compute this correlation coefficient inside the R software? So, if I say that I have the same setup that I have two data vectors x and y, then the correlation coefficient between x and y is computed by the command cor and inside the argument we have to write the data vector. So, the cor inside argument x and y will compute the correlation coefficient between x and y that is the data vectors x and y.

And when you try to look into the help of the cor function, then there are several options and I am going to detail here some important things which we are going to use actually. You will see here the function here the cor and inside the argument this and this I am writing several options.

So, now we try to understand them one by one, this x and y that we know these are going to denote the data vectors here like this. Now, there is here another command here use and I have written here inside the double quotes as everything.

(Refer Slide Time: 03:57)

Coefficient of Correlation

use : an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings "everything", "all.obs", "complete.obs", "na.or.complete", or "pairwise.complete.obs".

method : a character string indicating which correlation coefficient (or covariance) is to be computed. One of "pearson" (default), "kendall", or "spearman" can be abbreviated.

Now, what does this everything means and what is the use of this syntax use use. Actually this use is an optional character in this cor function which is trying to give a method of computing the covariance in the presence of missing values. If you remember earlier we had used the command like na dot r m, so this command also have a similar utility.

So, then in this case we have several option to give here say everything when you try to use all the observations or say all observation complete observation na dot or dot complete or pairwise complete observations and so on because, there are different test situation in which one would like to compute this coefficient of correlation.

So, we are simply going to use here the option here everything where I want to compute the correlation based on all the data right remaining details you can look into the help menu. After this there is another option here which now I am trying to denote in blue color, so that you can see clearly this is here method and inside the c command inside the argument I am writing here three options; pearsons, kendall and spearman.

Actually there are several types of correlation coefficient, up to now what we have studied the r and if you remember I had told you that this r is also called as say this Karl Pearson coefficient of correlation and this is how this Pearson is coming here. Second or say another correlation coefficient is rank correlation coefficient, which we will discuss

in the next lecture and this rank correlation is also called as Spearman, Spearman's correlation coefficient or Spearman's rank correlation coefficient.

So, this option here which now I am highlighting in red color spearman inside the double quotes, this option is used when we want to compute the rank correlation coefficient. And similarly there is another form of the correlation coefficient which is defined by the kendall which we are not using at the moment.

So, essentially we are going to use here the option pearson the correlation coefficient that was defined by say here summation $x_i - \bar{x}$, $y_i - \bar{y}$ and square root of summation $n(x_i - \bar{x})^2$ and summation $(y_i - \bar{y})^2$ this is actually computed by this option pearson right.

And this is method is mentioned here in the next slide that method is a gives us a characterizing indicating that which correlation coefficient of the covariance is to be computed for say pearson, for kendall or say spearman this has to be abbreviated inside the method and the default will be pearson.

(Refer Slide Time: 07:13)

The image shows a screenshot of an R console with several lines of code and their outputs. Handwritten annotations in red and blue highlight specific parts of the code and output. A green arrow points to the first line of code. A red circle highlights the output of the second line. A blue circle highlights the output of the third line. A red circle highlights the output of the fourth line. A green arrow points to the fourth line of code. A red circle highlights the output of the fifth line. A red circle highlights the output of the sixth line.

```
Example
Covariance
> cov( c(1,2,3,4), c(1,2,3,4) )
[1] 1.666667

R Console
> cov( c(1,2,3,4), c(1,2,3,4) )
[1] 1.666667

> cov( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1.666667

R Console
> cov( c(1,2,3,4), c(-1,-2,-3,-4) )
[1] -1.666667
```

Now, I try to illustrate you first some example if you try to see here I have taken here two data vectors 1, 2, 3, 4 and 1, 2, 3, 4 same values and I am trying to compute the covariance between them. Once I try to do it this value will come out to be 1.66 and so on and here you can see the slide. And now what I do in the next example which I am

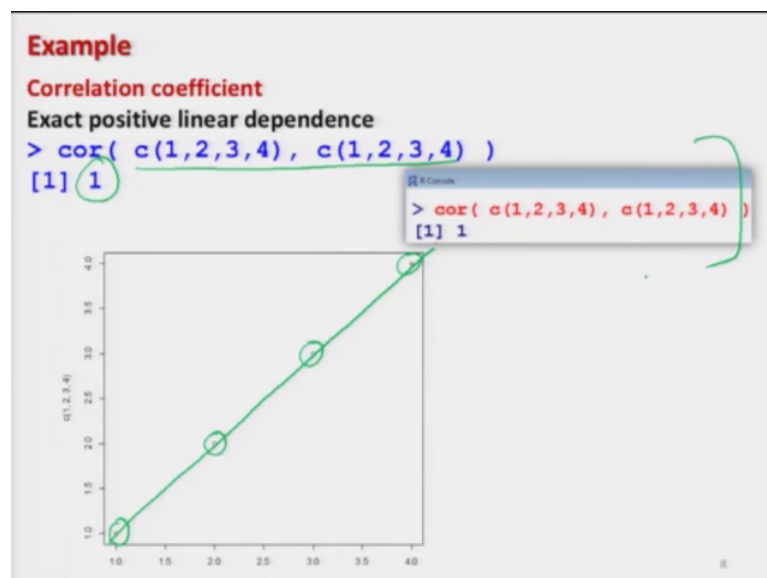
denoting now in say here purple color I try to take here the same data set c 1, 2, 3, 4, but in the next the data set I try to change the signs from the second data vector.

So, now if you try to see what happens here, the value of the covariance comes out to be minus 1.66 and so on. So, you can see here that these two values which I am now highlighting in red color this and this the magnitude of this values are the same, but only the sign is opposite and here is the screenshot. What is the meaning of this?

Now, if you try to look into this slide of the definition of correlation coefficient. We had understood that r can take positive value and r can take negative value, but if you try to see in the denominator this value will always be greater than 0, this value will always be greater than 0. So, this is the only value which is the covariance between x and y, this can be greater than 0 or this can be smaller than 0.

So, the direction of the correlation coefficient is determined by the covariance. So, I can say that the sign of correlation coefficient is determined by the covariance. And this is what I am trying to show you here that if the data points here in the first case here and in the second case here they have got a opposite sign, then this is given by this negative sign here right.

(Refer Slide Time: 09:50)

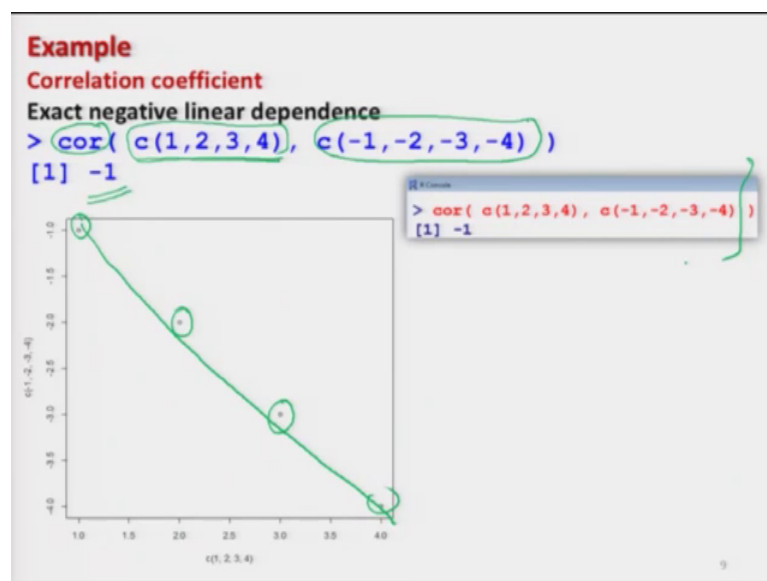


And if you try to plot these two data points how they will look like. For example, now I try to find out the correlation coefficient and by which I will also show you the direction.

So, when I try to find out the correlation coefficient between the two data vectors 1, 2, 3, 4; 1, 2, 3, 4 they are the same, then this correlation coefficient comes out to be here 1 that you can see.

And you see this is also indicated in the scatter plot which is given here you see these are the point 1, 2, 3, 4 and they are lying exactly on the straight line So, this is a case where we have exact positive linear dependence and this is the screenshot of the same operation, I will try to show you on the R software also.

(Refer Slide Time: 10:38)



And similarly if I try to take the data vectors with exactly opposite sign, say first get a vector with the all positive sign data and second data vector minus 1 minus, 2 minus, 3 minus, 4 having the opposite sign and if I try to find out the correlation coefficient between the two this comes out to be minus 1.

And you can see here this is indicated in the scatter diagram here these are the four points and this relationship is decreasing and this is the case of exact negative linear dependence between the two variable and here is the screenshot of the of both the things.

(Refer Slide Time: 11:17)

Coefficient of Correlation
Example

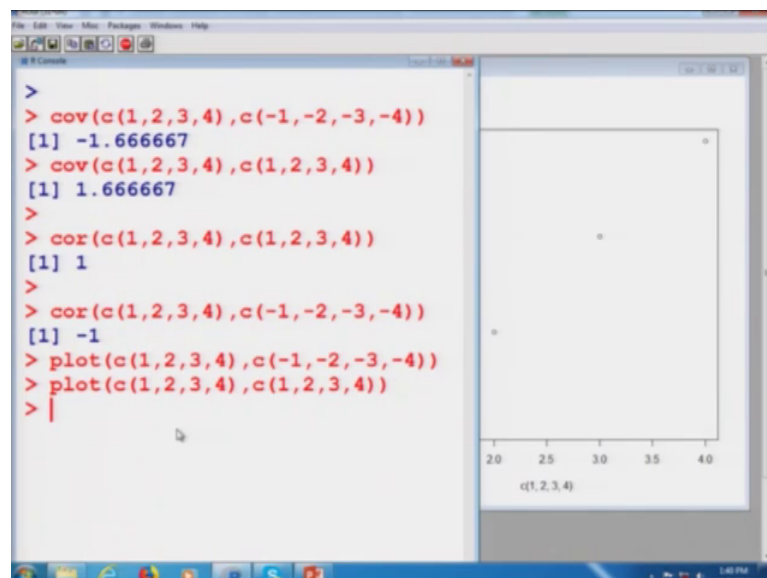
Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:

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

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

(Refer Slide Time: 11:21)



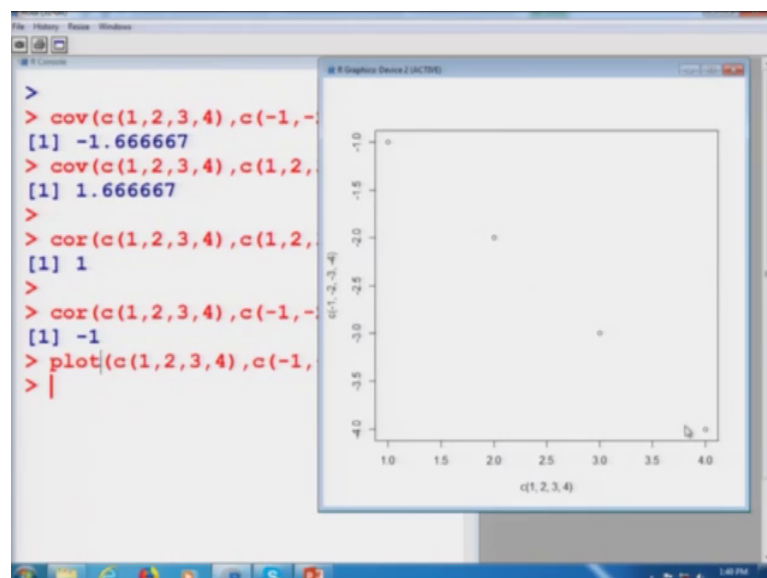
Now, before taking an example let me try to show you this things over the r console also. So, you try to see here I try to find out the covariance between 1, 2, 3, 4 and here minus 1, minus 2, minus 2 and minus 4, you can see here this comes out to be minus 1.6667 and if I try to remove the signs and suppose both the data vectors have got the same value, then in this case the magnitude remains the same, but the sign become positive.

So, in this case you can see here the covariance is positive. Now in case if I try to find out the correlation in the same case where the covariance is positive you can see here this

is obtained by the function `cor` and this comes out to be 1 for the sign of positive is maintained here. Similarly, if I try to take the data set with negative signs and if I try to find out the correlation coefficient between 1, 2, 3, 4 and minus 1, minus 2, minus 3, minus 4 you can see here this comes out to be here minus 1 right.

Now, if I try to make here the scatter plots I will try to show you the both the things here. So, you can see here say I can make here the scatter plot of 1, 2, 3, 4 and minus 1, 2, 3, 4 you can see here this comes out to be like this where the direction of my cursor is indicating the negative relationship.

(Refer Slide Time: 12:56)



And similarly if I try to make it here positive; that means, that two data vectors both are here positive 1, 2, 3, 4; 1, 2, 3, 4 in this case you can see here that this scatter diagram is trying to give a positive exact relationship. Now, I try to take an example to show you that how the interpretation of correlation coefficients comes into picture. Suppose I try to take here a data set where I have obtained the marks and the number of hours of study by the 20 students, this is the same example which I have considered in the earlier lectures while making different types of graphics or say bivariate plots.

So, we have recorded their marks in the first row in both the tables and the number of hours per week they studied in hours here right this is the same example which we have considered several times and this data has been stored inside two variables marks and here hours right.

(Refer Slide Time: 14:05)

Coefficient of Correlation
Example

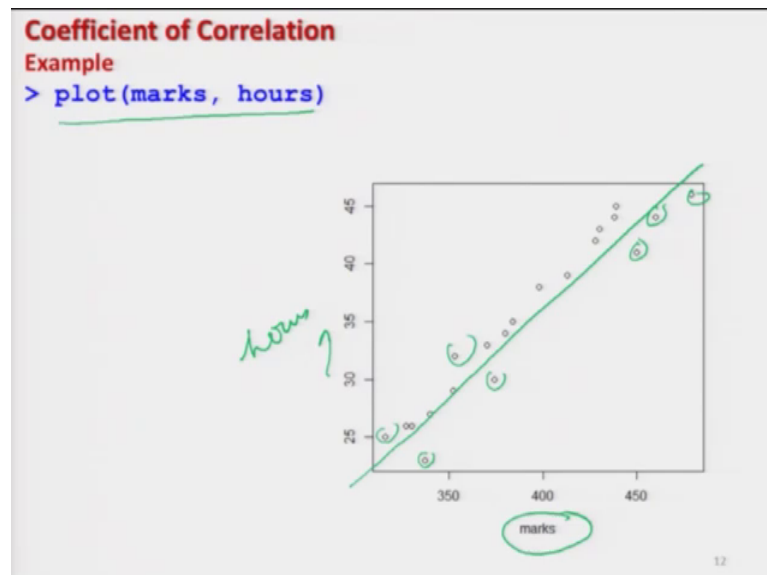
marks =
c(337, 316, 327, 340, 374, 330, 352, 353, 370, 380, 384, 398, 413, 428, 430, 438, 439, 479, 460, 450)

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

11

Now, the first thing what it comes to our mind that now we have got the data and before using the concept of correlation coefficient we would like to see what is the type of relationship whether it is linear or not. So, we simply use the option plot and we try to plot here the marks versus hours.

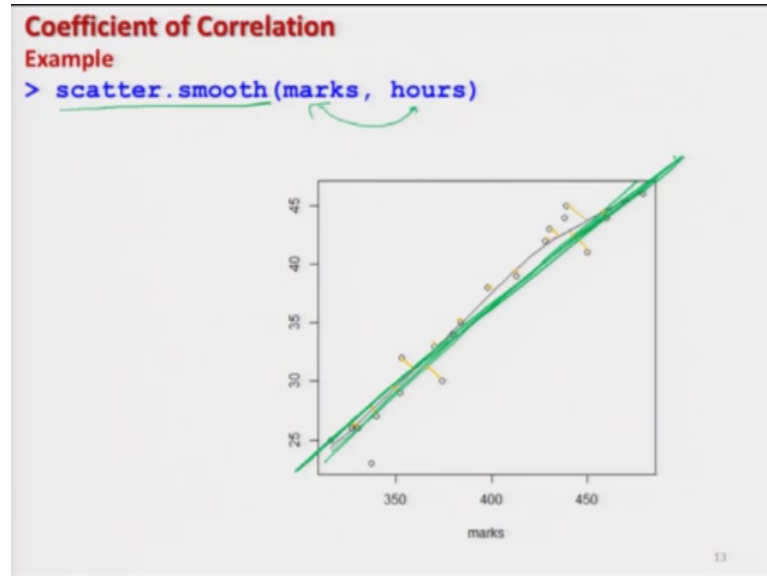
(Refer Slide Time: 14:23)



So, you can see here in the data here is lying like this all the data here are your marks and here is your actually hours right and you can see here this there is a linear trend here. So,

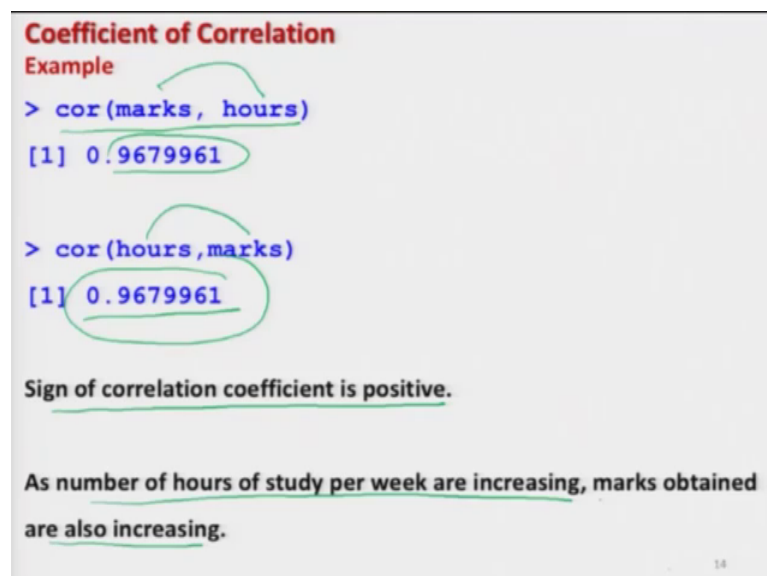
this gives us a sort of assurance that the relationship between marks and hours is well close to linear and now I am, but still I have made the trend line by hand.

(Refer Slide Time: 15:02)



And now we also have learned how to make this trend line, so for that I will use the same command that we discussed earlier scatter dot is smooth and I will try to make your line between the two data vectors marks and hours. And you can see here that this line is now also indicating that this is nearly linear and this gives us a confidence that in this case I can use the correlation coefficient.

(Refer Slide Time: 15:29)

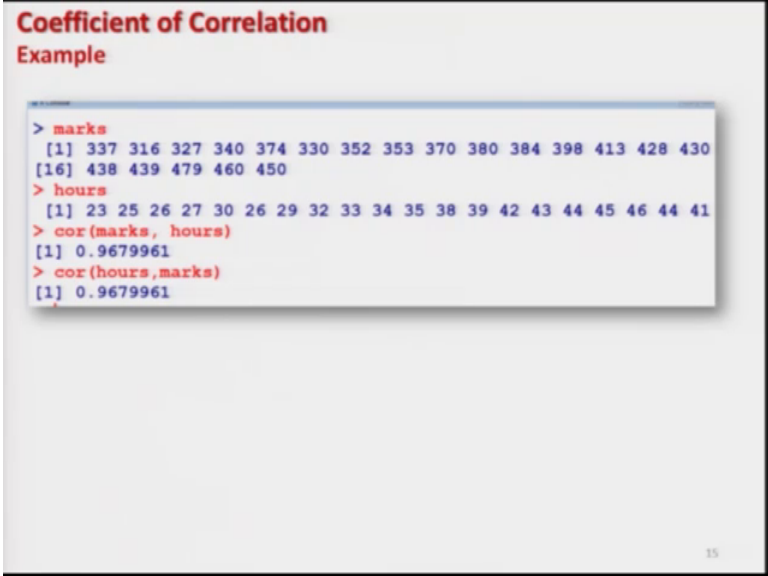


And then I try to find out the correlation coefficient between marks and hours you can see here this comes out to be 0.96. So, you can see here what is this 0.96 indicating. If you try to look into this curve you can see here that these deviations which I am plotting in orange color they are very close to lines and these deviations are very small and these points are lying very close to the trend line.

So, this is possibly indicating that on an average the degree of linear relationship between marks and hours is very strong and this is nearly 96.79 percent. Similarly, if you try to interchange those variables, so earlier I have taken say marks and hours and now I try to take hours and marks. So, you can see here this value of the correlation coefficient remain the same this is what we have discussed in the last lecture also right.

So, in this case you can see here the sign of the correlation coefficient is positive and this is indicating that the relationship between x and y is positive which is here you can see here. And we can conclude that as the number of hours of study per week are increasing, the marks obtained by the students are also increasing because there is a positive relationship and the value of correlation coefficient is pretty high. So, that is why I can say that whatever the data is telling that is correct and my conclusions based on the data they are correct and it is also showing a linear trend.

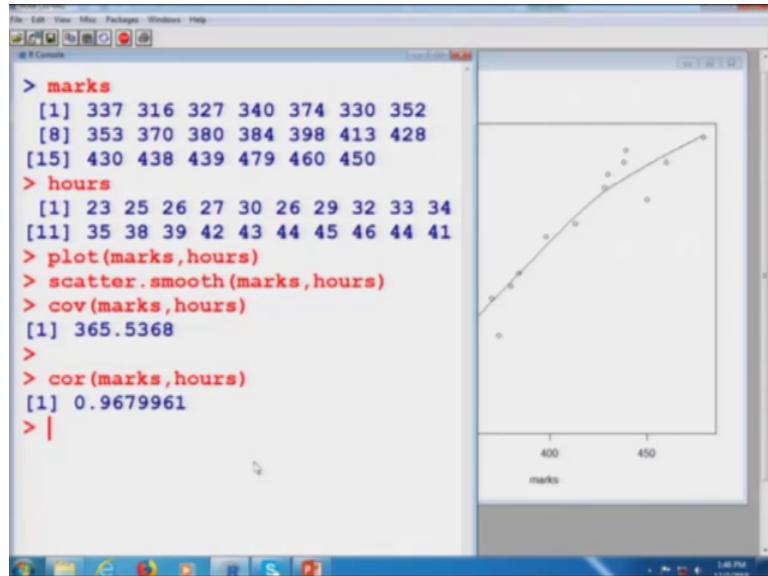
(Refer Slide Time: 17:08)



```
Coefficient of Correlation
Example

> marks
[1] 337 316 327 340 374 330 352 353 370 380 384 398 413 428 430
[16] 438 439 479 460 450
> hours
[1] 23 25 26 27 30 26 29 32 33 34 35 38 39 42 43 44 45 46 44 41
> cor(marks, hours)
[1] 0.9679961
> cor(hours, marks)
[1] 0.9679961
```

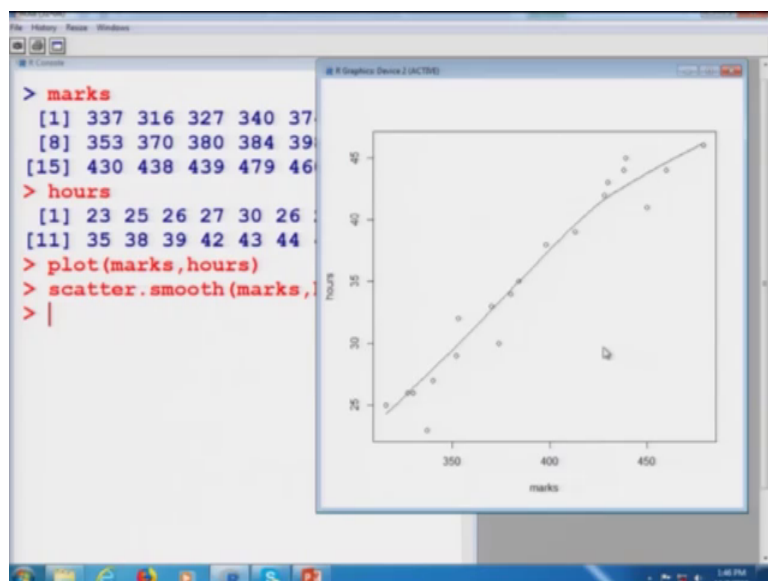
(Refer Slide Time: 17:19)



And this is here the screen shot I will try to show you it on the r console also before I take one more example of the negative one, so I will try to show you here. I already have stored the data, so you can see here this is the data for marks and this is the data of our say hours.

So, if I first try to plot marks versus hours or say hours versus mark whatever you want this gives you us you can see here nice scatter plot indicating the linear trend, if you try to make it here a trend line also, so try to give a scatter dot smooth and you can see here that a line is also plotted here right.

(Refer Slide Time: 17:36)



And now I try to find out the covariance between marks and hours that will indicate the direction. So, you can see here that the value of the covariance is 356.5368 and it is positive. So, it is indicating and it is matching with the work on every whether were a graphical conclusion that the relationship is positive.

Now, I would try to find out the correlation coefficient between marks and hours and you can see here this is 0.96 right. So, you can see here that this correlation coefficient is trying to take care of the variance as well as covariance between the two variables.

(Refer Slide Time: 18:41)

Coefficient of Correlation
Example
 A medicine was given to 10 patients. Its quantity (in mg.) and the time (in hours) taken in showing the affect was recorded as follows:
 We want to know the effect of medicine on time taken to affect.

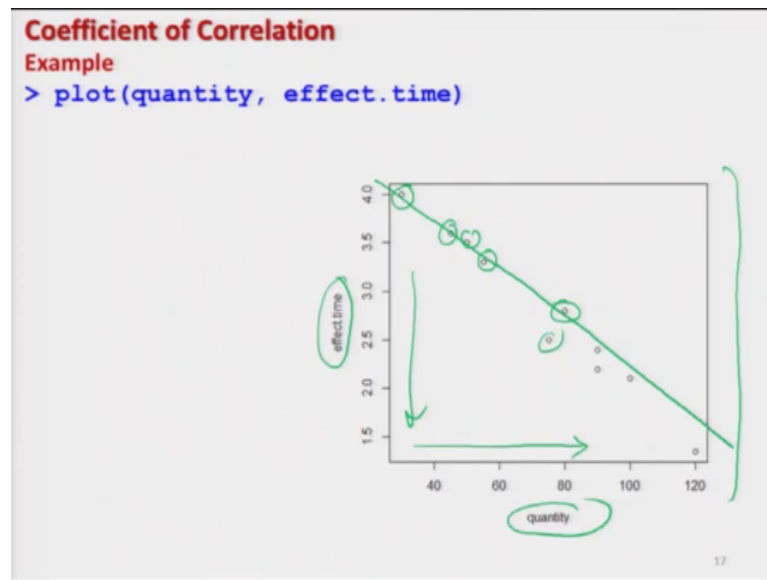
Quantity (in mg.)	30	45	80	120	90	75	55	90	50	100
Time (in hours)	4	3.6	2.8	1.35	2.4	2.5	3.3	2.2	3.5	2.1

> `quantity = c(30, 45, 80, 120, 90, 75, 55, 90, 50, 100)`
 > `effect.time = c(4, 3.6, 2.8, 1.35, 2.4, 2.5, 3.3, 2.2, 3.5, 2.1)`

Now, I will try to take one more example and try to illustrate something more. Suppose there are 10 patients and those patients are given some medicine and the quantity of the medicine is measured in milligrams mg and the time say in hours is recorded and to show that when the medicine is started showing the effect and this data is compiled here say quantity of medicine and the time in say hours. So, this is indicated here, so this is the data set of patient or the person number 1 or patient number 1.

So, this is indicating that 30 mg of medicine was given and it took 4 hours of time. Similarly for the second data set this is the patient number 2 and a 45 mg of medicine was given and he took 3.6 hours of time and so on. So, this data on the quantity is stored here inside a variable quantity and time is stored here in say effect dot time. One thing please do not use their variable time because time is used by the R software also, so be careful. So, now I will try to take this data set and I will try to first make a plot here.

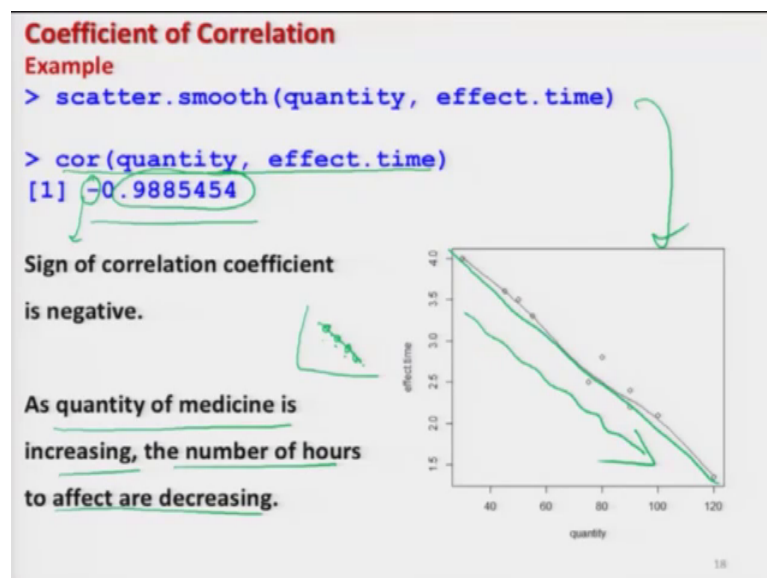
(Refer Slide Time: 20:09)



So, you can see here this is the screenshot and you can see here this is on the x axis it is quantity and y axis this is time, that is the effect dot time and you can now see here these observations are coming out to be like this. So, you can see here that as the value of x's are increasing the values of y's are decreasing.

So, this shows here a sort of negative trend right and this information we would like to verify with the covariance function or the correlation function but now let me plot here the trend line also.

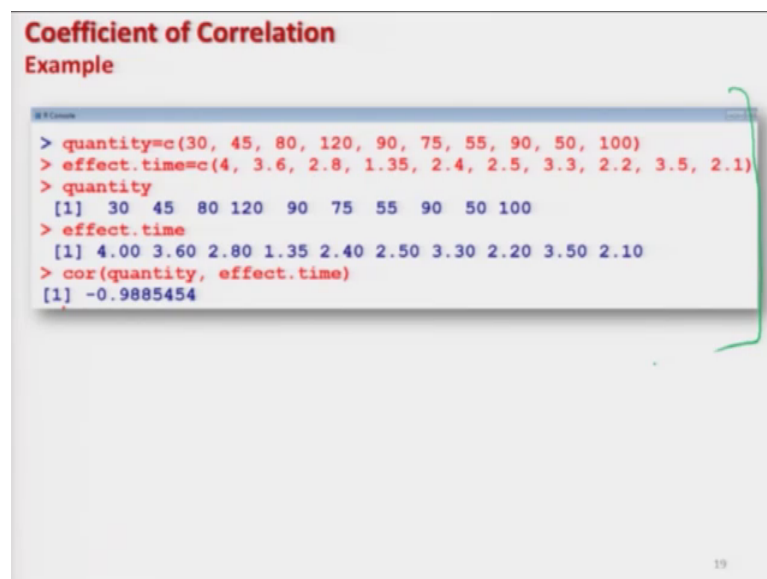
(Refer Slide Time: 20:43)



So, you can see here that the trend line is also indicating that the relationship is almost linear and we can safely use the concept of correlation coefficient right, so this is the outcome. And now I try to find out the correlation between quantity and effect dot time, so this comes out to be here minus 0.9885454, well you can control the number of digits also. And now if you try to see this is showing the value negative, so the sign of negative is indicating that this relationship is decreasing, as the values of x's are increasing the values of y's are decreasing and the magnitude here is 0.9885454 close to 0.99.

So, close to 0.99 means the relationship is going to be nice and linear and the degree of linear relationship is pretty high, the maximum value is 1 in the case of 1 what will happen that all the observation will be lying exactly on the line and whereas, in this case this is just 0.988, so it is very close to the line. So, now, this give us a information and confidence that we can use here the concept of correlation coefficient and this degree is coming out to be 988.

(Refer Slide Time: 22:20)



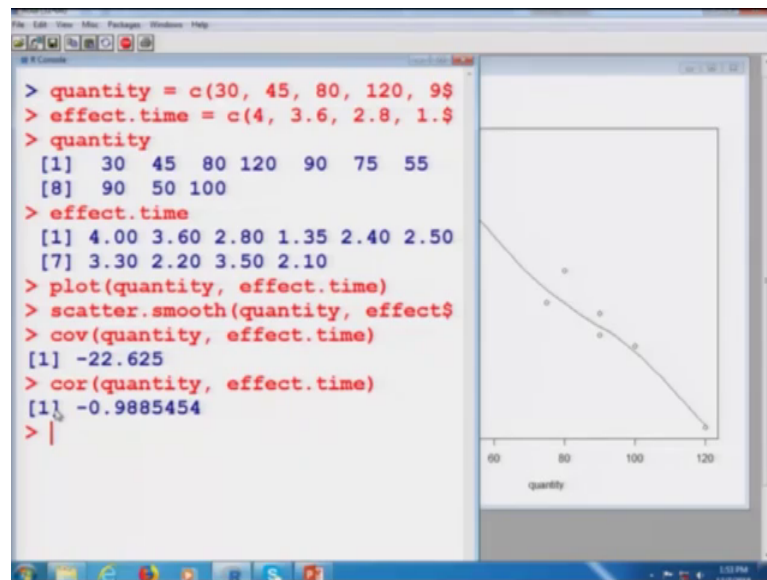
The screenshot shows an R console window titled "Coefficient of Correlation Example". It displays the following R code and output:

```
> quantity=c(30, 45, 80, 120, 90, 75, 55, 90, 50, 100)
> effect.time=c(4, 3.6, 2.8, 1.35, 2.4, 2.5, 3.3, 2.2, 3.5, 2.1)
> quantity
[1] 30 45 80 120 90 75 55 90 50 100
> effect.time
[1] 4.00 3.60 2.80 1.35 2.40 2.50 3.30 2.20 3.50 2.10
> cor(quantity, effect.time)
[1] -0.9885454
```

A green bracket on the right side of the console output highlights the correlation coefficient result.

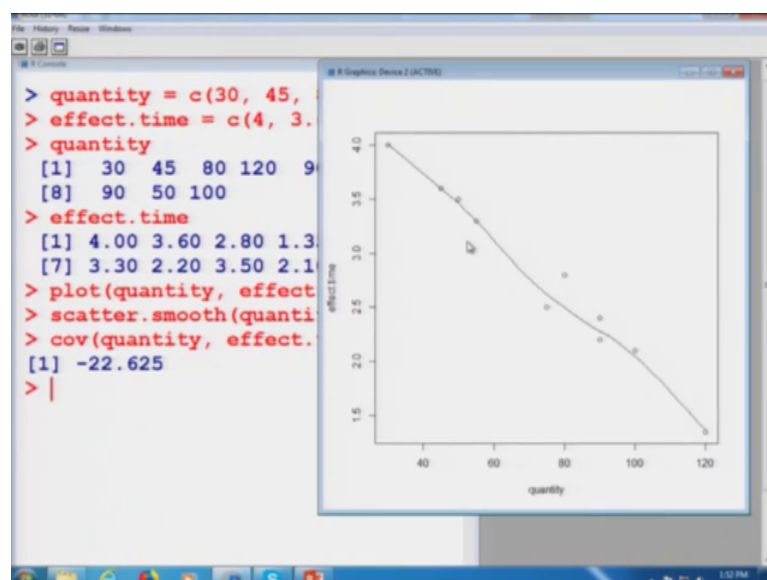
So, now, I can conclude that as the quantity of medicine is increasing, the number of hours to effect are decreasing and I would try to show you this thing on the r console also and this is the screenshot what I have done ok. So, let me first try to copy this data the earlier data was there because I had used it earlier.

(Refer Slide Time: 22:39)



So, this here is quantity first let me clear all the things clear the screen by control l, try to get this quantity and then I try to copy the effect dot time variable and the data content in it. So, this is here effect that time, so you can see here quantity is obtained here, effect dot time is obtained here. Now, I would like to know what is the nature of relationship. So, I try to make here a plot between see here quantity and effect dot time.

(Refer Slide Time: 23:15)



And you can see here this is the plot this is the same plot that we have just obtained in the slides and now if you want to make it here I scatter a smooth plot by adding a trend

line. So, I have to use the command here scattered dot smooth and this gives me here this graph which is the same points, but with a trend line. And now first I try to find out here the covariance between the two because covariance will assure us that what is the sign of the direction whether this is positive or negative. So, you see this covariance comes out to be minus 22.625.

So, this minus sign is indicating here that the relationship is negative and this is verified here in this graph also, if you try to observe the direction of my cursor on the screen this is decreasing right. Now, I try to find out the correlation coefficient between the quantity and effect dot time and this comes out to be here minus 0.988. So, once again this sign is coming from the covariance and it is indicating that the relationship between quantity and effect of time they are negatively correlated, as the quantity increases that time taken to affect is decreasing right.

And this is quite obvious also we know from our experience that when we try to increase the dose of the medicine, then it acts faster and the time to react becomes smaller and smaller well up to certain extent after certain limit that may say that is not advisable and you have to depend on that doctors advice right. So, now in this lecture I have shown you that how to attempt or how to take a decision whether you want to compute the correlation coefficient or not.

So, the steps are first to try to find out the scatter plot or the smooth scatter plot whether trend line try to look at the trend. In case if you are convinced yes there can be a linear trend or the relationship between the two variable is approximately linear this can be positive or negative, then you decide that well in this case I can use the concept of correlation coefficient to measure the degree of linear relationship and then you try to use the formula for the coefficient of correlation.

So, I stop here and I would request you that you please try to take some more data sets and try to plot them and then try to compute the values of correlation coefficient and see what do you get. So, you practice and I will take the topic of rank correlation in the next lecture and then I will see you in the next lecture.

So thank you very much and see you soon, goodbye.