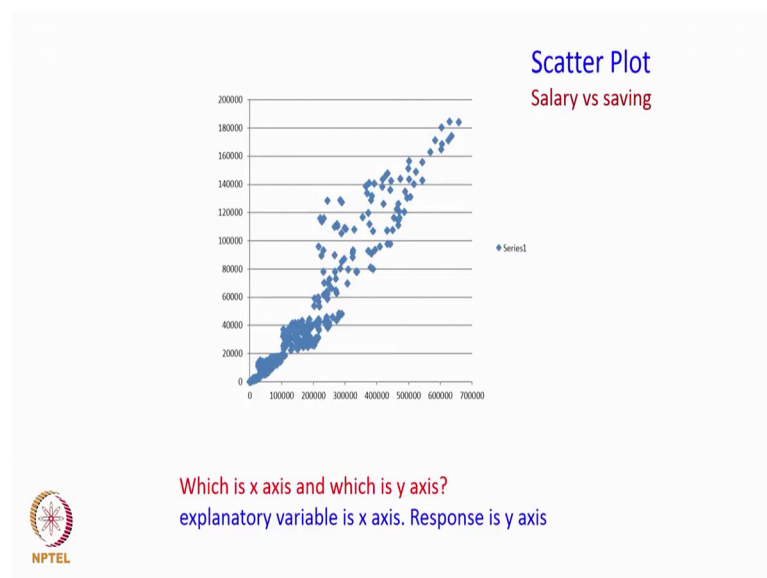


Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 10
Association Between Numerical Variables

In this lecture, we study Association between Quantitative Variables or numerical variables. In the previous couple of lectures, we looked at association between Categorical variables. Now, we extend; we look at measures like Chi square and Cramer's V for categorical variables and we will try to find out what are the equivalent measures, if we look at quantitative or numerical variables.

(Refer Slide Time: 00:46)

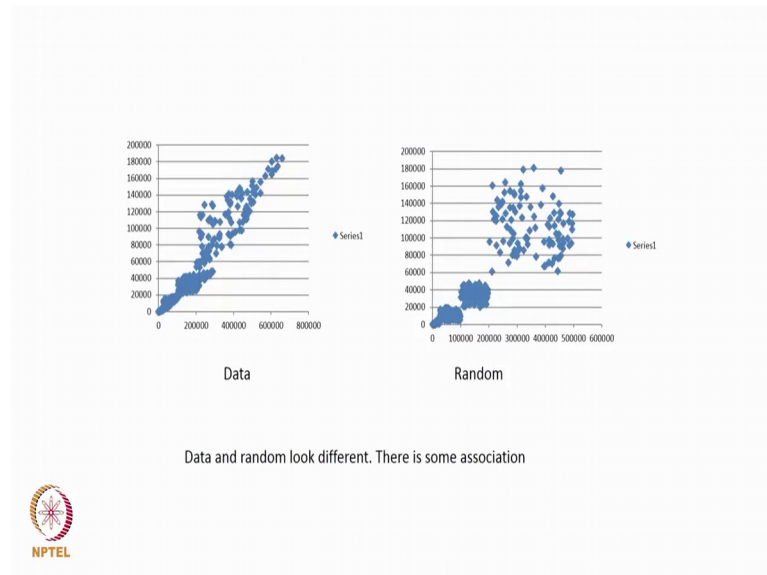


Now, let us look at some data and let us say that we are trying to look at the association between salary and saving. So, we have 2 variables; one is the Salary of the person and the other is the saving of the person. So, one of the variables acts as the x variable or the x axis variable and the other variable acts as the y variable or the y axis variable. So, generally the explanatory variable is the x axis variable and the response variable is the y axis variable.

So, in this example we can assume that the salary is the reason for saving because people get income and salary from which they save. Therefore, the saving becomes the y axis variable and the salary becomes the x axis variable. So, let us assume that we have some

100 pieces of data for salary and saving for a pair of x , y ; where, x is the salary and y is the saving. And let us assume we have plotted this and this is what we get if we plot these 100 pieces of data each data having an x and a y value.

(Refer Slide Time: 02:00)




Now, the same pictures is shown on the left hand side and let us look at this. Now, let us do something else and then, we started with hundred sets of data each data having an x and each data having a y . Now suppose, we quickly randomize the x and y ; in the sense, we just quickly make a make a random sort of x and then random sort of y which means now with the new data the x and y are not exactly as they were paired in this data and we still get 100 pairs with different x and different y . Let us say kind of randomly chosen and then, we plot that using this and get this kind of a plot.

So, this is the data looks and this is the random. Now the first impression is that this the data looks very different from the random and therefore, the first impression is that there is association between these 2 variables, if they look alike look reasonably similar to the eye. Then, one could say that there is no association. In this case they look different and therefore, we can say that there is association. So, how different and all those depends on how we understand the pictures. But I am sure, most of us would agree that these 2 pictures are not similar; they look different and therefore, there is some association between the salary and the saving.

(Refer Slide Time: 03:30)

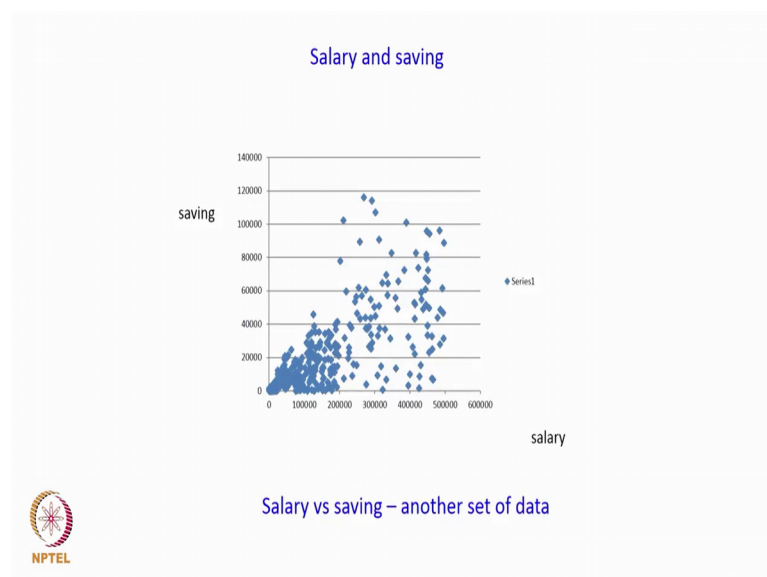
Describing association

1. *Trend* – upward or downward?
2. *Curvature* – Is it linear or does it show a curve?
3. *Variation* – Are points tightly clustered along the pattern?
4. *Outliers and surprises* – Are there outliers?



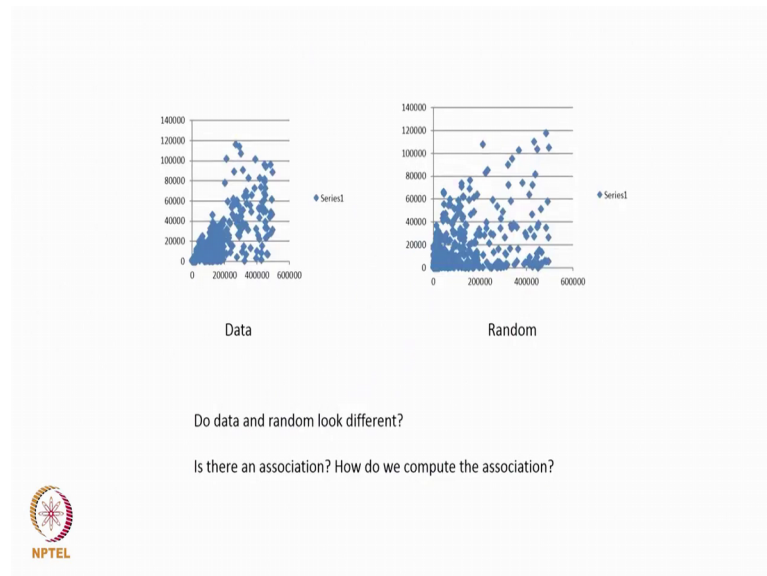
Now, how do we describe association in terms of many things? The Trend so, is there an upward or a downward association which means as x increases; thus, y increase or as x increases; thus, y decrease. We could look at Curvature which means is it linear; it is a straight line or does it show a curve and then, we look at Variation are points tightly clustered along the line. Are they are further away and then, are there outliers; are there points that should not be belonging completely away; are there surprises and so on? So, we will try to look at some of these in this lecture.

(Refer Slide Time: 04:09)



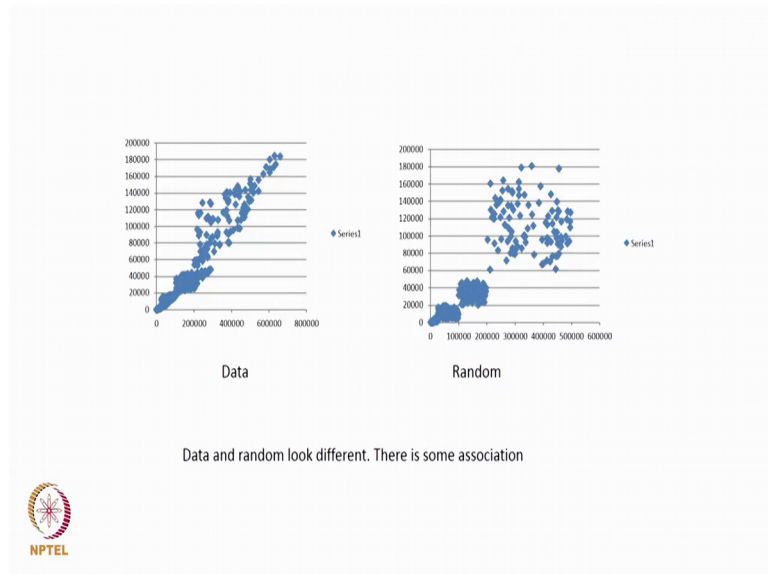
Now, if we look at another set of data on salary versus savings and let us say this is how the data looks with x as the salary and y as the response variable which is the saving. Now if we try to do the randomized picture of this, where we still have 100 points, but the x and y are now sorted completely randomly which means they do not have the old x y pair and when we do a similar exercise, this is how the original data looked.

(Refer Slide Time: 40:38)



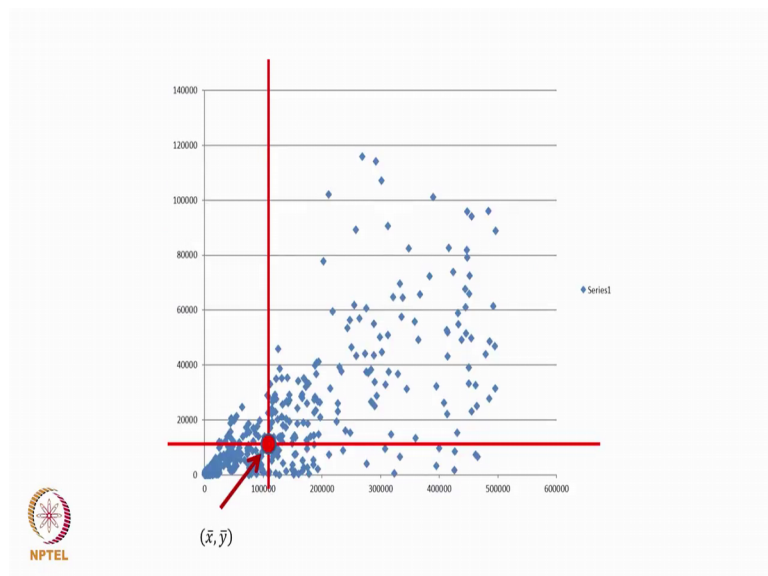
This is how the randomized data looks. And do they look different? May be they do not; I mean if we look at it very very carefully one might say they look different; but to the eye, one might get a feeling that both of them look a little cluttered and here, we might conclude that there is actually no association between this or very little association between this.

(Refer Slide Time: 05:09)



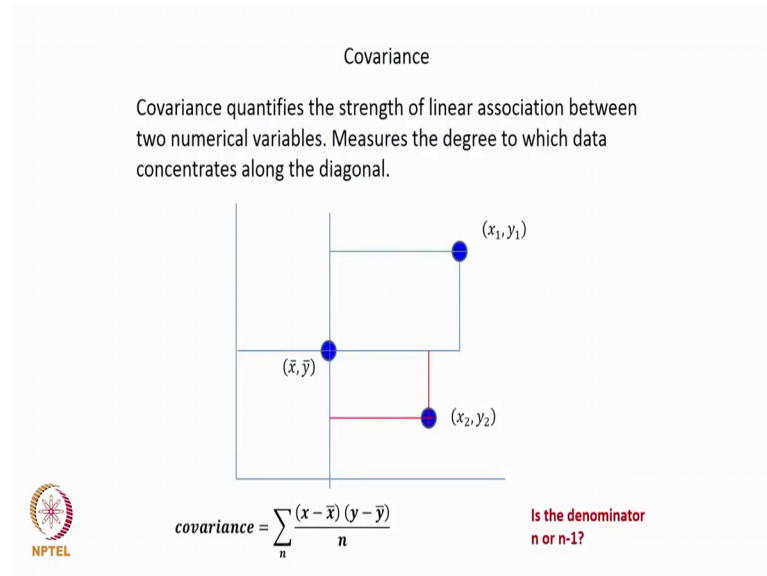
So, we have if in the in the previous example, we saw that there is a vast difference. So, we said there is association here there is not so much of a difference. So, we said may be there is not so much association. But then, how do we compute is there a measure or is there a metric that tells us there is association or there is no association among these variables; we will see those matrix as we move along.

(Refer Slide Time: 05:32)



Now, to do this let us take this kind of a data and then, we try to plot the \bar{x} and \bar{y} and that is shown in this picture. This point is our \bar{x} \bar{y} that we can calculate.

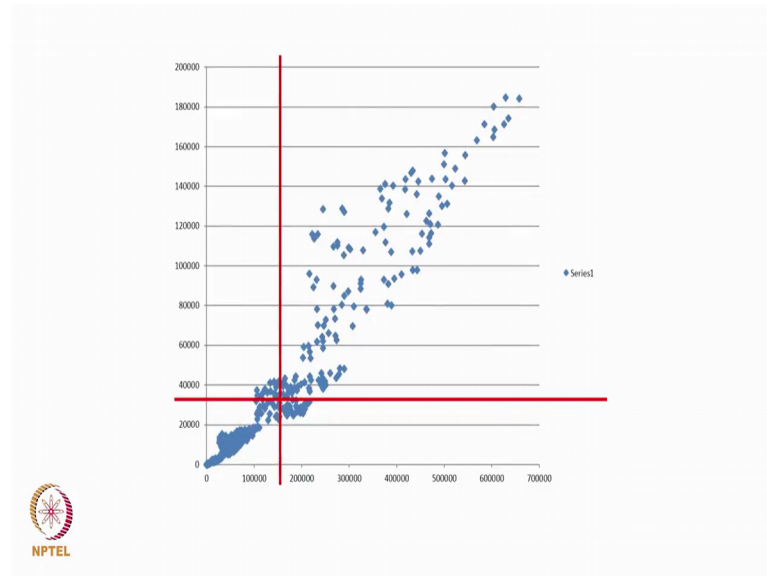
(Refer Slide Time: 05:45)



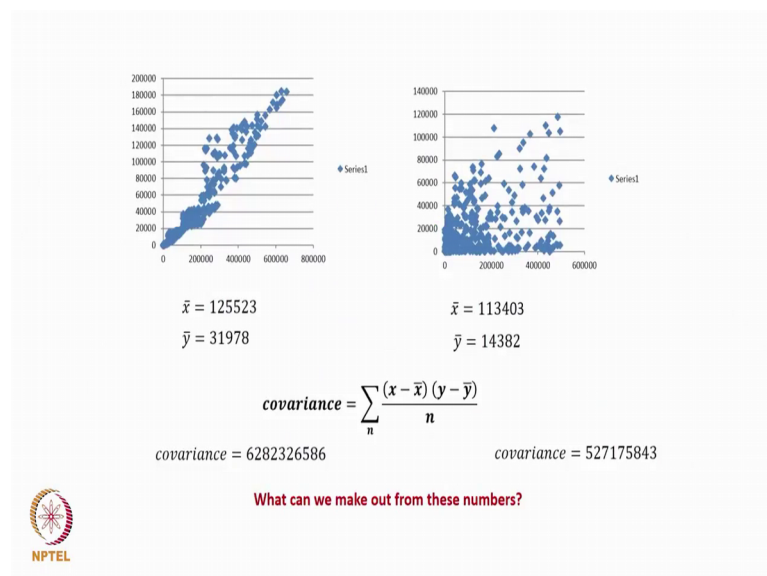
Now, we have already seen this measure called Covariance. So, we visit the Covariance again. Covariance quantifies the strength of linear association between two numerical variables. It measures the degree to which data concentrates across the diagonal.

So, given x_1, y_1 and x_2, y_2 and an \bar{x}, \bar{y} the covariance is given by $(x - \bar{x})(y - \bar{y})$ by n . I have also raised a question is it n or is it $n - 1$? We can assume n and later when we find out correlations and other measures, we consistently use the same denominator so that there is no bias in the calculation. So, we use n in this case. $(x - \bar{x})(y - \bar{y})$ by n is the covariance between these two.

(Refer Slide Time: 06:41)



(Refer Slide Time: 06:44)



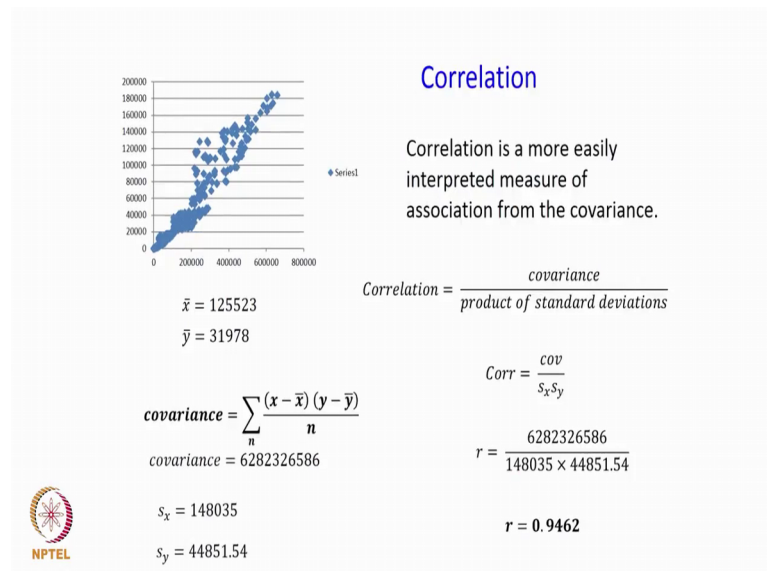
So again, in the same picture, where the averages are shown so, when we do this the first picture when we have this picture which we go back to this, this is the original data and this is the random. So, we get back to this picture and then, we try to find out x bar and y bar. So, in this picture x bar is 125523 and y bar is 31978. In the random picture x bar is 113403 and y bar is 14382. This is not the random picture. This is the second set of data. So, let us go back the, this is one set of data; this is data set 1 and let us say this is data set 2 and you can see these 2 pictures here.

So, this is data set 1 which is reflected here and this is our data set 2 which is here that is expanded to this. Therefore, they show different x bars and y bars, if there had been the same data and the random the x bar and y bar would not change because the same 100 values would be used. Therefore, these two represent two different sets of data. This is called data set 1; this is called data set 2.

Now, x bar is 125523; x bar is 113403 because the data set is different. y bar is 31978; y bar is 14382. So, if find the covariance of both sigma x minus x bar into y minus y bar n; summations for all the 100 values, the covariance in the first case become 6282326586. The covariance in the second case is 527175843. So, which has a higher covariance?

We quickly calculate the number of digits and then, realize that there are 10 digits in this number and there are 9 digits in this number. Therefore, this shows higher covariance compare to this data. So, what we make out from these numbers? Generally, what we can make out is if the covariance is higher, there could be some association with the data and between comparable data sets the one that has higher covariance seems to have association.

(Refer Slide Time: 09:15)

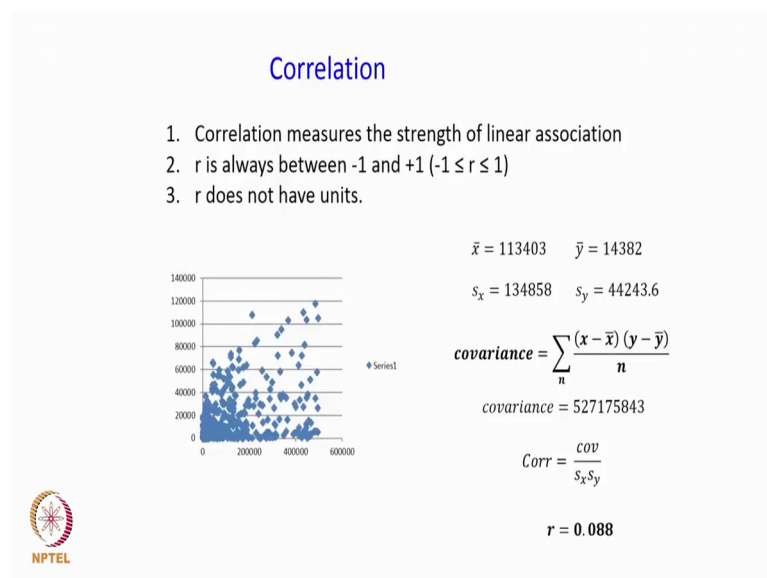


So, for the first set of data x bar is 125523; y bar is 31978. Covariance is 6282326586 and then, we calculate standard deviation of x which is S x is 148035; S y is 44851.54. So, standard deviation of x and standard deviation of y are also shown here. Now, we compute the correlation. So, correlation is equal to covariance divided by the product of

standard deviations. So, covariance by $S_x \times S_y$; so, 6282326586 divided by 148035 into 44851.54 which is 0.9462. So, correlation for this is 0.9462. We have already studied the correlation coefficient and its computation.

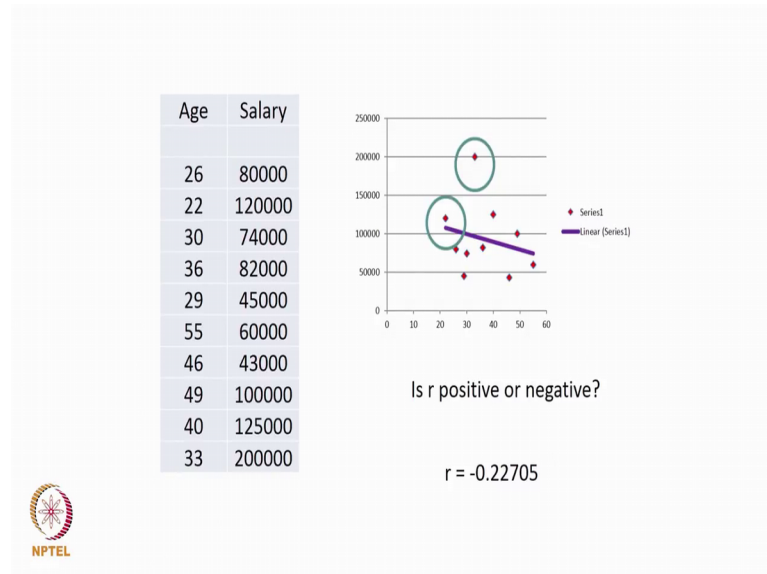
So, we are now doing it one more time to show that correlation is 0.9462. Also remember that covariance can be negative, we saw that in an earlier lecture; whereas, standard deviations are strictly non-negative, there either 0 or positive. And because, correlation coefficient is covariance divided by a positive quantity, we would have correlation as a negative quantity as well, as well as the positive quantity. The range is between minus 1 and plus 1 and in this case, we have a correlation of 0.9462 which is very close to 1 and then, we could say that the data is indeed associated.

(Refer Slide Time: 11:09)



Now, if we look at this data the second set of data. Before that let us look at these correlation measures the strength of linear association between numerical variables. What is important is linear association between numerical variables. r is always between minus 1 and plus 1 and r does not have units. Standard deviation has units. So, S_x has a unit in this case its money or rupees.

(Refer Slide Time: 11:45)



S y also has unit which is money or rupees and then, we go back we do the product of S x and S y. So, it is rupees square or money square. Covariance is $(x - \bar{x})(y - \bar{y})$ by n summation. So, it also has the unit money square and therefore, correlation does not have a unit. It is a unit less quantity which is between minus 1 and plus 1. So, when we look at the second set of data \bar{x} is 113403; \bar{y} is 14382; S of x is 134858; S of y is 44243.6. Covariance is 527175843 and when we compute the correlation, we get point 0.088. So, correlation is closed to 0 and therefore, there is no association or very very little association between salary and savings in this case.

Now, let us look at another example, where we have age of the person versus salary and let us say we have set of these 10 points and then, we first plot these 10 points. So, these points these 10 points are plotted. We get a scatter plot of these 10 and then, let us say we also fit a line through a software that we can do and you want to check this and we then want to ask a question is there a positive association or is there a negative association from this data?

The line seems to say that there is looks like at least for the data that we have looked at, there is a negative association between age and salary. Because which is also given by a computed correlation coefficient of minus 0.22705. Now, there is an outlier that we can think of which is well away and outside, it answers all the 4 questions that we looked at. So, there is an outlier in this example.

(Refer Slide Time: 13:47)

Age	Height	Weight
11	152	38
12	153	40
13	160	43
14	168	52
15	170	61
16	183	76
17	176	72
18	180	78
19	178	81
20	180	69

Correlation Matrix

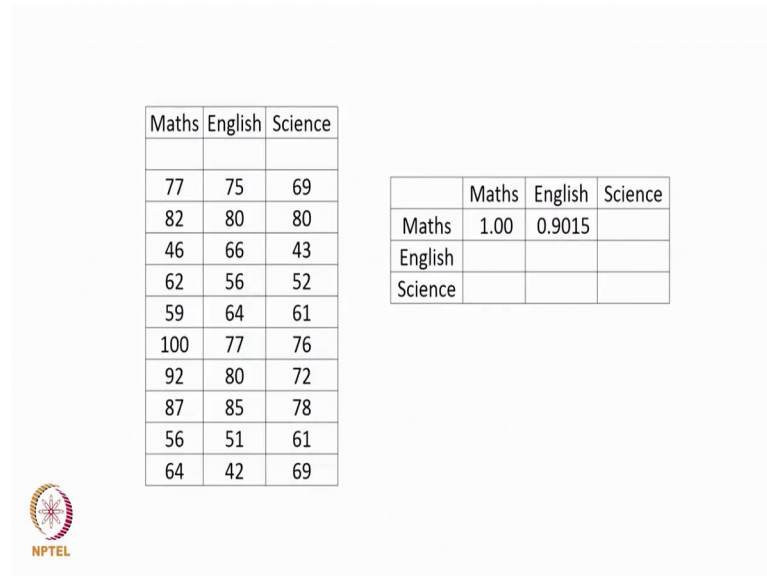
	Age	Height	Weight
Age	1.00	0.9015	
Height			
Weight			



It is also possible to find the correlation when we have more than 2 numerical variables. So, we have age, we have height and we have weight. Let us say we have this data for boys in the age group of 11 to 20 and let us say we picked up one person with age 11; one person with age 12 and so on and that is the height and weight respectively. So, correlation between age and age, age and itself is 1.

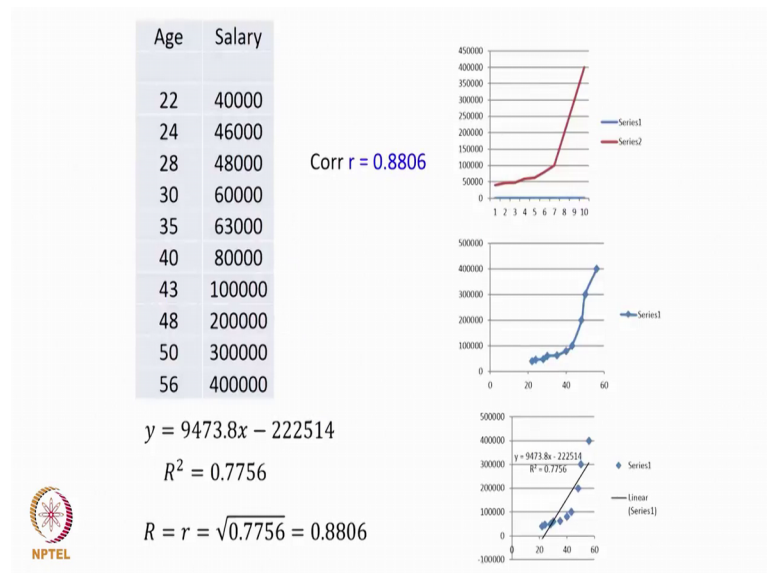
So, in this example, we can compute the correlation between age and height which works out to be 0.9015 and you can do the rest of it correlation between height and height is also 1 and correlation between weight and weight is also 1. Correlation between age and height is the same as correlation between height and age and therefore, this matrix will be a symmetric matrix with 1 across the diagonals. So, effectively it is enough to find only 3 numbers; age versus height, high versus weight and age versus weight.

(Refer Slide Time: 14:55)



Similar exercise, you can do that let us say the marks obtained in 3 subjects; Mathematics, English and Science by high school students are given. So, we could do that and we could complete this table.

(Refer Slide Time: 15:08)

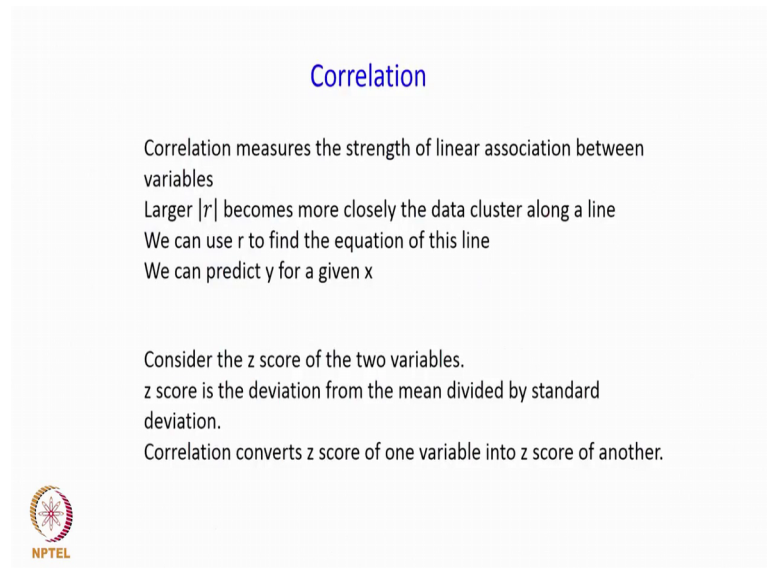


Now, let us look at another set of data. So, here we look at the age that is given here and the salary is also given here. Another set of data. So, we did correlation and we get correlation is equal to 0.8806. So, if we fit a line, we show the line that is fit which is shown here and this line as I mentioned also in an earlier lecture has something called r

square which is a goodness of fit which is 0.7756 and then, we also realize that this r square is actually square of the correlation coefficient of 0.8806.

So, 0.8806 square becomes 0.7756. But all this is true only if we decide to fit a straight line this association holds. If we fit a curve then, we have to have different types of association. We already saw that covariance or correlation is a measure of linear association between numerical variables.


(Refer Slide Time: 16:15)



Correlation

Correlation measures the strength of linear association between variables
Larger $|r|$ becomes more closely the data cluster along a line
We can use r to find the equation of this line
We can predict y for a given x

Consider the z score of the two variables.
 z score is the deviation from the mean divided by standard deviation.
Correlation converts z score of one variable into z score of another.




So, correlation measures the strength of linear association between variables. Larger r absolute value of r becomes more closely the data clusters along the line, we can use r to find the equation of this line and we can predict y for a given x . We can do all this when we have the correlation coefficient. And if we consider the z score, we have not yet come in detail about z score, but z score is the deviations from the mean divided by the standard deviation correlation converts the z score of one variable into the z square of another variable.

(Refer Slide Time: 16:48)

Correlation

$$z_x = \frac{(x - \bar{x})}{s_x} \quad z_y = \frac{(y - \bar{y})}{s_y}$$

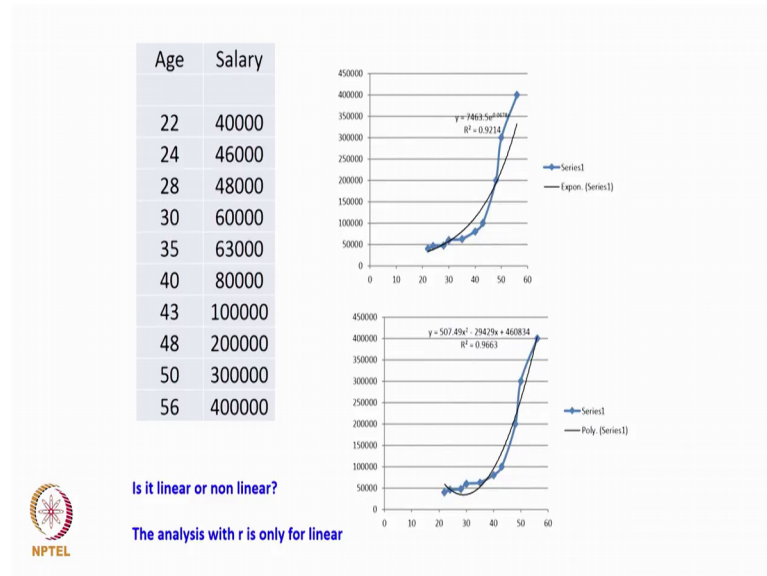
The equation of the line is $\hat{z}_y = rz_x$

$$\frac{(\hat{y} - \bar{y})}{s_y} = r \frac{(x - \bar{x})}{s_x}$$
$$\hat{y} = \bar{y} + \frac{rs_y(x - \bar{x})}{s_x} = \left(\bar{y} - \frac{rs_y\bar{x}}{s_x} \right) + \frac{rs_y}{s_x}x$$
$$\hat{y} = a + bx$$
$$a = \bar{y} - b\bar{x} \quad b = \frac{rs_y}{s_x}$$


So, there are these mathematics that I have given the equations that I have given. So, z_x is x minus \bar{x} by S_x ; z_y is y minus \bar{y} and then, we can have the equation of the line is \hat{z}_y is equal to r into z_x and from this we can get a and b that are associated with the line. And we are just showing these computations. So, again this data correlation is 0.8806.

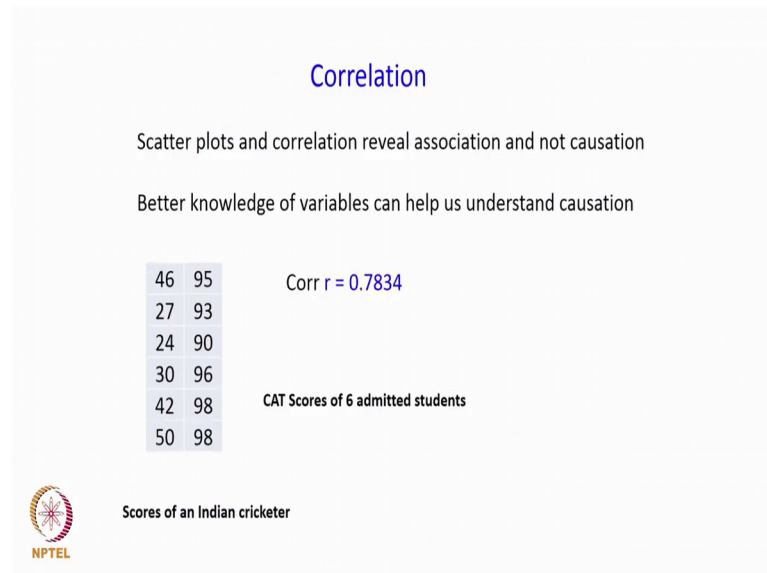
So, you can use this calculation from \bar{x} , \bar{y} , S_x , S_y and r and from this, we can quickly get this b which is 9473.8 and then, we can get a which is minus 222514 and if we actually fit that using the line on a using a software, we would get 9473.8. We got 9473.28 in our calculation, a small approximation minus 222514 was calculated as 222495. r^2 was 0.7756 correlation is point 0.8806.

(Refer Slide Time: 17:58)



A shown similar pictures and here, I have actually shown how curves are there and how these curves are also fit for this kind of data. But we will restrict ourselves to linear association. But we have to understand that the analysis for r is only for linear association among these.

(Refer Slide Time: 18:18)



Now, we also have to understand one thing with through which we look at it through another set of data. Now, let us show 2 sets of data. Let us assume that this r CAT scores

are exam scores of 6 students and let us say these are data which correspond to scores made by a cricketer in 6 innings.

So, we could treat one of them as a x variable and we could treat the other as a y variable and then, if we only apply the Math and try to find the correlation we get r is equal to 0.7834. So, if I had not told you that this set of variables represent let us say CAT scores and this set of variable represents runs made by a cricketer and if I simply had withheld this and had simply asked find the correlation coefficient? You will get 0.7834 and if I had asked the question is there an association? Then, you might say yes, there is a reasonably high correlation. Maybe there is an association between this x and this y.

But the moment we say that this represents say CAT score and this represents the runs made by a cricketer. So, we realize that there need not be and will not be an association. So, better knowledge of variables can help us understand causation. So, scatter plots and correlations reveal association. They donot tell us causation. For example, they donot tell that if this is the x variable; then, I calculate the y variable, I can apply the math and calculate a number.

But how well I interpret the number will actually depend on the variables that we are studying and it is important to know those variables first, before we even attempt to find is there a cause and effect between these 2 variables? But just by themselves without defining what these variables are, if there is an association; yes, there is an association with the high value of r.

So, with this we complete this lecture and in the next lecture, we would look at some numerical examples of association among numerical variables after which we will start studying probability in further detail.