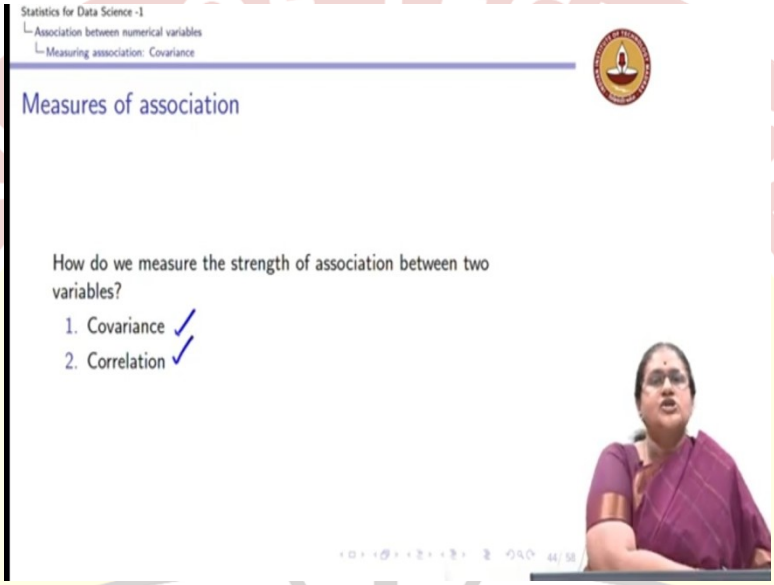
IIT Madras

ONLINE DEGREE

Statistics for Data Science 1
Professor Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture 4.6
Association between Two Numerical Variables: Covariance

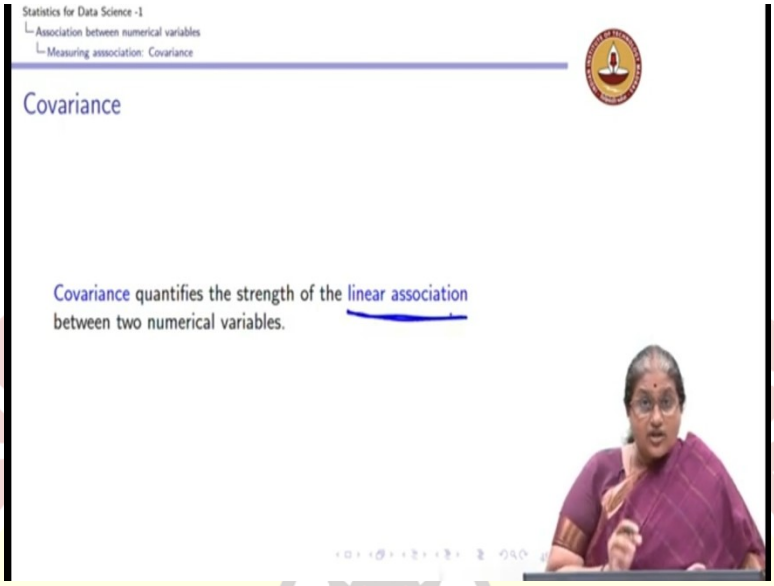
(Refer Slide Time: 00:15)



So, we have seen how to actually describe association using a scatter plot and we will get an idea between whether an association exists between the two numerical variables under consideration or not. So the next question which we want to ask is; can I quantify this association? In other words, since we are discussing and we are describing numerical variables and we have two numerical variables with us.

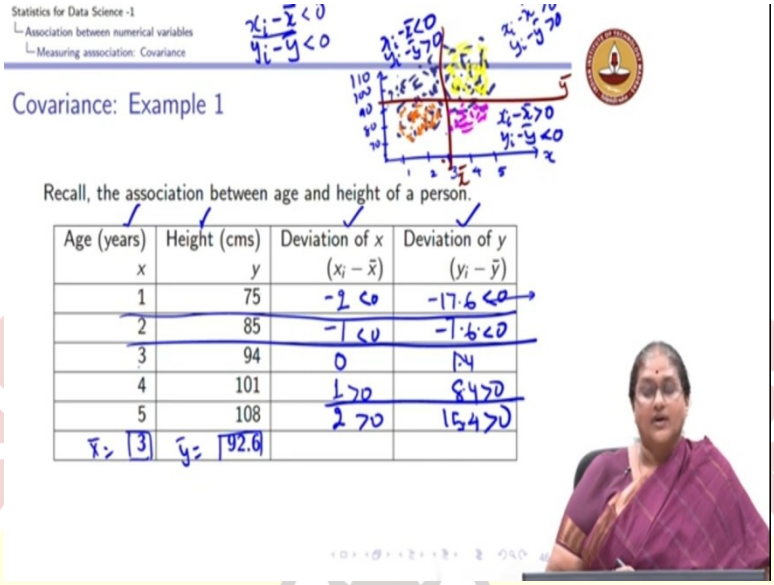
The natural question to ask can I come up with a measure of association between the two variables. The two popular measures of association which is used to describe the association between two numerical variables; the first measure is what we refer to as a covariance and the second measure is what we refer to as correlation. So how do we measure the strength of the association?

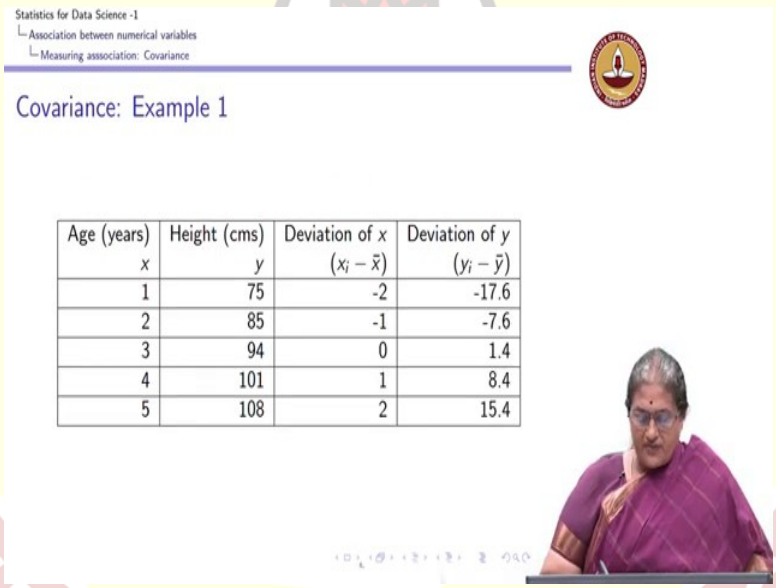
(Refer Slide Time: 01:21)



So let us again go back to a simple example. Covariance actually quantifies the strength of linear association. A word of caution again, I said we are not going to deal with curve data here. We are only going to understand linearly associated variables. So remember covariance and correlation are measures of linear association.

(Refer Slide Time: 01:50)





So let us look at a simple example. Again recall the association between age and height of a person. This is what we have already seen. So now what do we mean by covariance and correlation? So the key observation which we have is the following. So first thing we notice is we want to know, I have the variable here x which is age 1, 2, 3, 4, 5 and I also have the height so I break this. I start with 70, I have 70, 80, 90, 100, 110.

So this is my first point is 70, sorry, this is my first point is 75. Then I have a 85, then I have a 94, then I have a 101 and then I have a 108. This is my scatter plot. Now look at the mean of the first variable. The mean of the first variable is 3 which is nothing but the average of these 5 ages,

which is 1 + 2 + 3 + 4 + 5 divided by 5. So this is the mean of the first variable. So we know this point, so let me cross out all the points.

So if you look at the mean of the first variable, you can see that in the first variable case there are two points which lie to the left of the mean and two points which lie to the right of the mean and there is one point which lies on the mean. Similarly let us look at the mean of the second variable which is 92.6, so I have a 92.6 here. So let us use a different colour.

So I have this here and I have this here. So what this means is this point lies this point is only mean of the first variable but above the mean of the second variable. Now, these two points are below the mean of both the variables and these two points are above the mean of both the variables. So now you can see that I could have data where I could have had points here.

Now, what are these points? These points would be below the mean of the first variable and above the mean of the second variable. Now, these points would be above the mean of the first variable and below the mean of the second variable. So these so in general what is the idea we are trying to say is if in general I have a scatter plot of x and y and this is the scatter plot which I have.

Suppose I have 100 observations and I have scatter plot of this kind I have the same means, assume I have the same mean. So 3 is my mean here and 92.6 is my mean here. This is my $\bar{y}$ and this is my $\bar{x}$. So you can see from the scatterplot, I can divide this entire points into 4 regions and let us give different colours to each region. This, the scatter or the points in the orange region are points which are below the mean of both the variables.

The points which are in the yellow region that is these points are points, which are above the mean for both the variables and the points in pink are above the mean for x variable and below the mean for y variable and the points in blue are above the mean for y variable and below the mean of the x variable. Now, why do we care about this? So if I look at the deviation of the orange points from the mean, what do I mean by deviation?

So you can see that all the orange points are lesser than the mean of both x and y variable. So if I look at the orange points my $x_i - \bar{x}$, which is the difference between the point and the mean and $y_i - \bar{y}$ which is the difference between the point on the y axis and its mean, you can see that

since x_i and y_i are lying below $\bar{x}$ and $\bar{y}$ for orange points, both of them are going to be less than 0.

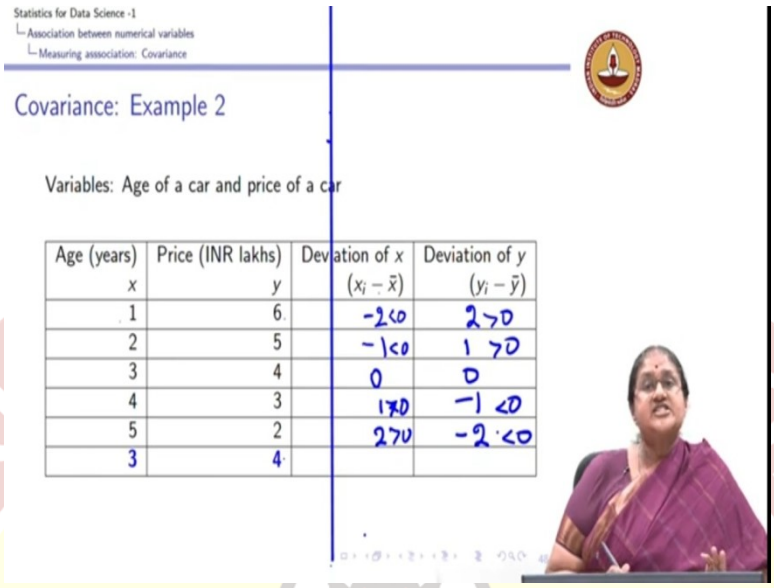
Similarly for the pink points, I will have $x_i - \bar{x} > 0$ and $y_i - \bar{y} < 0$ because y points are lesser than $\bar{y}$. Here I have both $x_i - \bar{x} > 0$ and $y_i - \bar{y} > 0$. Here I have my $x_i - \bar{x} < 0$ and $y_i - \bar{y} > 0$.

So what is this $x_i - \bar{x}$, $x_i - \bar{x}$ is nothing but the deviation of observation from its mean. Recall, I have two variables. So x is the first variable, y is the second variable, $\bar{x}$ is equal to 3, $\bar{y}$ is equal to 92.6. So my $x_i - \bar{x}$ is 1 - 3, which is - 2, 2 - 3 which is -1, 3 - 3 which is 0, 4 - 3 which is a 1, 5 - 3 which is 2 the deviation of x_i from its mean.

Similarly, if I look at the deviation of, I can look at 75 - 92.6 which is again -17.6 and similarly I have -7.6 I will have 1.4 and then I have 8.4 and I have 15.4. So you can see that this is the deviation of the points from their respective means. Now why is this again of any interest to us? Now the key point so I have written the deviation of the points from their respective means.

Now what is the interest in this? So here you can notice that the deviation of the first observation on both the variables have the same sign. Second observation again has the same sign, the fourth has the same signed by same sign, I want to say that both of them are greater than 0. Here both of them are less than 0.

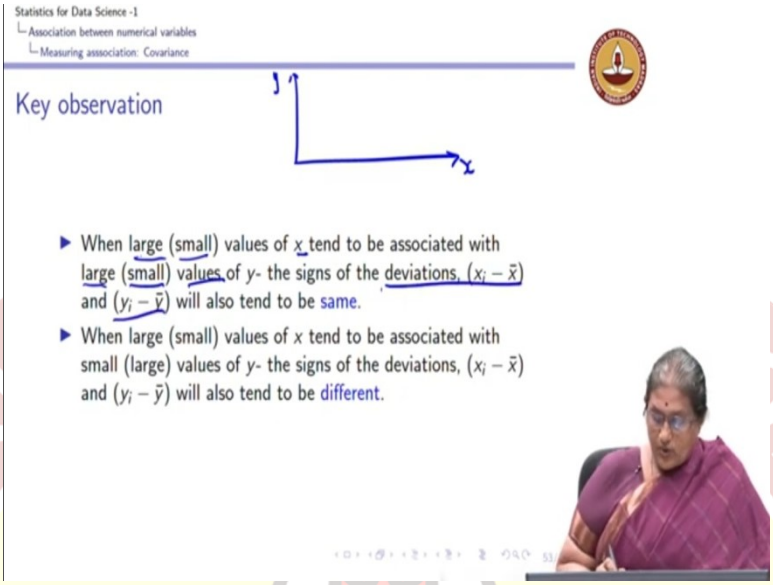
(Refer Slide Time: 10:02)



Now let us look at another example and do the same exercise. So let us look at the deviation here. You can see that again I have 1 - 3 which is - 2; 2 - 3 which is - 1, 0, 1, 2, age; the deviation does not change, 6 - 4 is a 2, 5 - 4 is a 1, 4 - 4 is a 0, 3 - 4 is a -1, 2 - 4 is - 2 and if you look at the deviation of x and y in this example, you can see that this is greater than zero and this is greater than zero, this is greater than zero. When this is greater than zero, I have this less than zero, this less than zero.

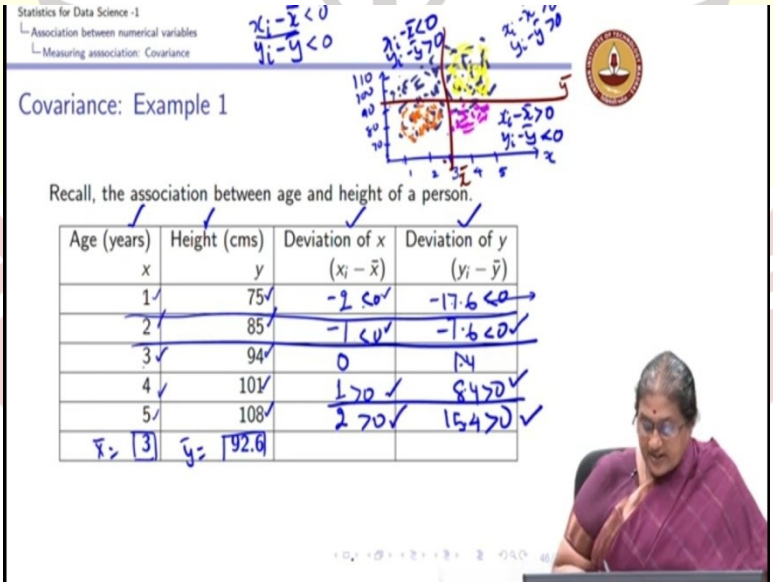
So these are the observations which I can make from these 2 examples. So that leads us to a key question that what is the key question we are asking.

(Refer Slide Time: 11:09)



So you can look at the key observations that if I am plotting x versus y, if large values of x are associated with large values of y, similarly if small values of x are associated with small values of y, the signs of the deviation, so what do we mean by signs of the deviation?

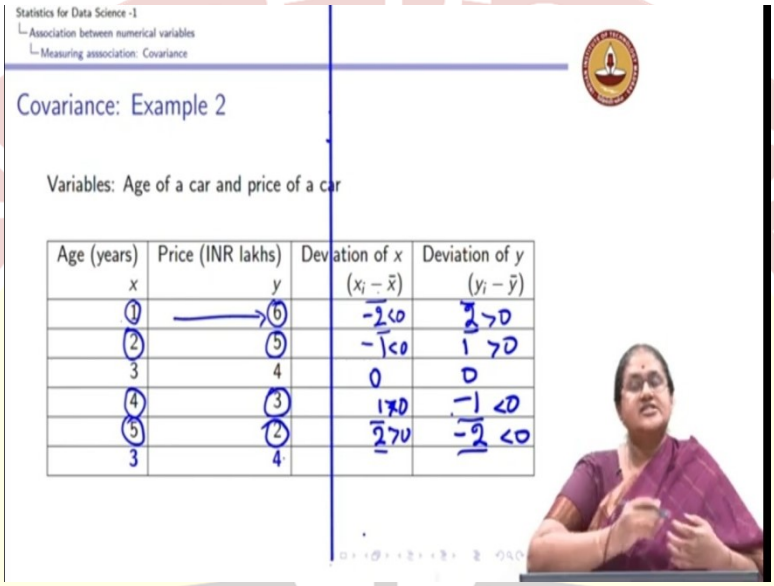
(Refer Slide Time: 11:38)



If you look at this case the first example you saw the case. In the first example, we saw that large values of age. What are the large values of age? 4 and 5 large values of height 101, 108; 3, 4, 5 can be looked as large values of age, and they are associated with large values of height. Small

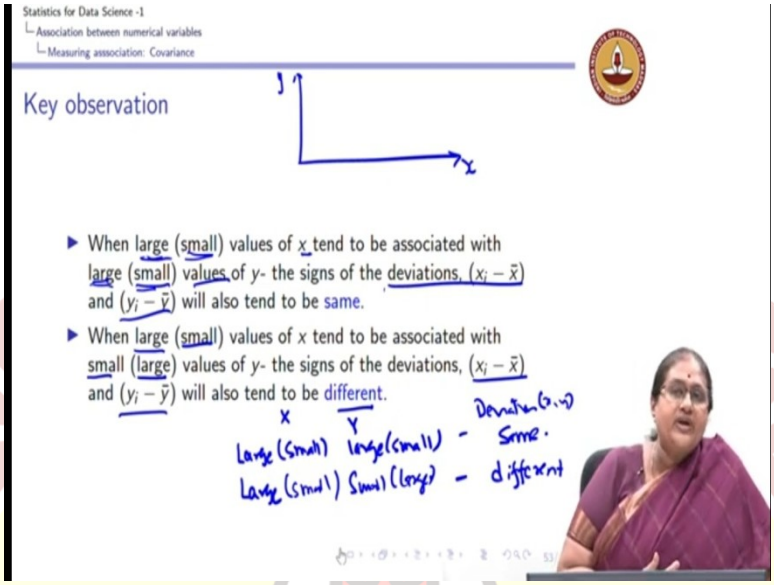
values of age are associated with small values of height. Hence, you can see that the deviation of both the variables have the same size. Here, the deviation is less than 0. Here also, the deviation less than 0. Here again it less than 0. It is less than 0. This is greater than 0, greater than 0, greater than 0, greater than 0.

(Refer Slide Time: 12:33)



Whereas in the next example, you can see that the small value is associated with the large value of price and the large value is associated with small value. Similarly 2 is associated with the next larger value and 4 is associated with a small value and as a consequence, you can see that the deviation of x is negative but the deviation of y is positive. Similarly when the deviation of x is positive, the deviation of y is negative. In other words the deviation of the variables are of different signs. So, how do we use this observation?

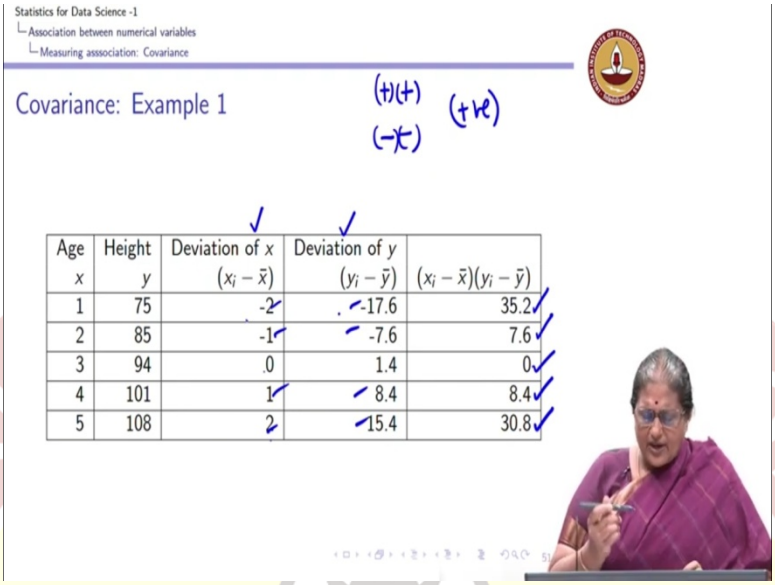
(Refer Slide Time: 13:20)



Remember we wanted to quantify the association. So when we look at this, so when the large value tends to be associated with small or small tends to be associated with large, we see that the signs and the deviation are different whereas if the large is associated with large and small is associated with small, we see the signs of deviation to be same.

So to summarize if the large, so x large is associated with y large or x small is associated with y small. The deviation of x and y are same, the signs. Similarly if x large is associated with y small and x small is associated with y large, we find the deviation signs to be different. This is one key observation which we have. So the question is now if the deviation signs are the same, then if I take a product of these deviation, it could give me a measure of the covariance or association. So how do we look at that?

(Refer Slide Time: 14:49)

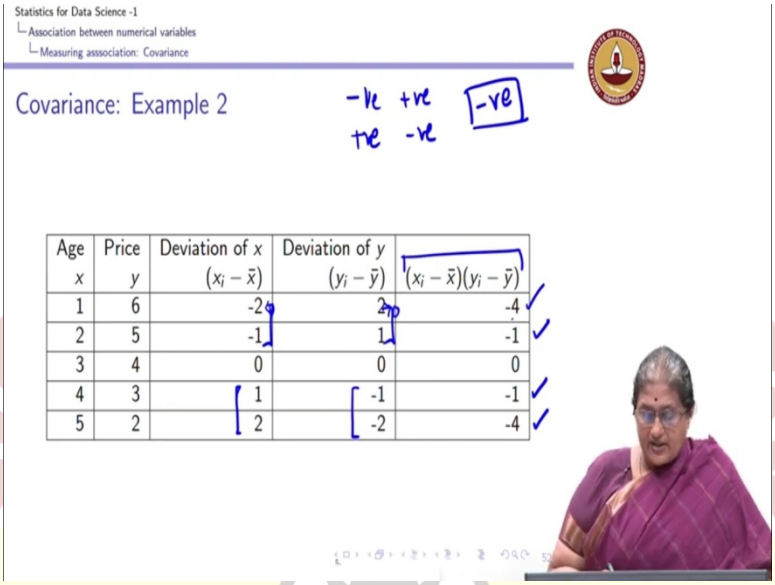


So if I have for example, look at my first example. I saw that the signs of the deviations are same. In other words, I could have if x is a positive deviation, my y is also a positive deviation. If x is a negative deviation my y is also a negative deviation. You can see both are negative in this case, both are positive in this case.

So if I look at the product of the deviations, I am going to have a positive sign. If I look at the product of the deviation because both the deviation $x_i - \bar{x}$ and $y_i - \bar{y}$ are of same size. When I look up the product of these deviations, I will get a positive sign and you can see that (- 2) × (-17.6) is 35.2, (-1) × (-7.6) is 7.6. Here, I have a 0, I have 1 × 8.4 is 8.4, 2 × 15.4 is 30.8.

So I have the product of deviations. All my deviations is positive in this case. There be a chance of one being positive or negative, there could be a chance but for in this simple example, I can see that the product of all my deviation is positive.

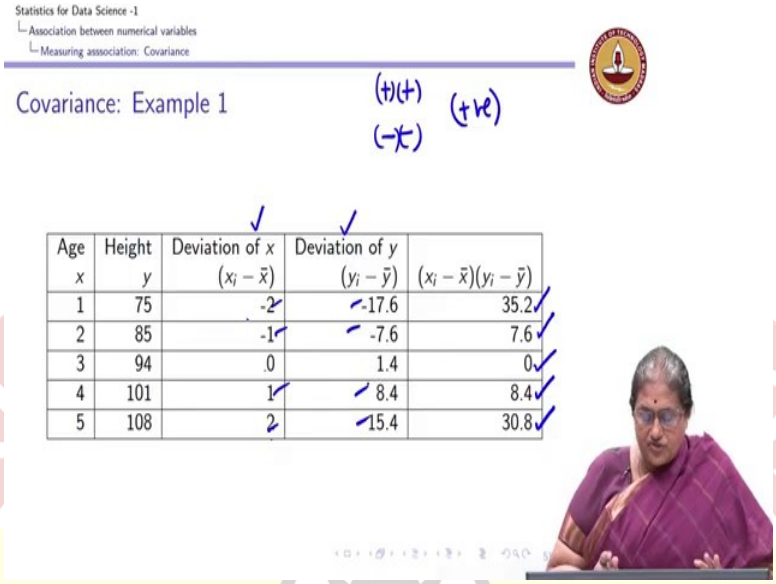
(Refer Slide Time: 16:21)



Similarly, let us look at the other extreme example. Here my x deviation is negative, this is positive so the product, so I either have a negative x and a positive y deviation or a positive x and a negative y deviation. So here these 2 are negative x with a positive y deviation (whether these two) whereas these 2 are positive x deviations with negative y deviations.

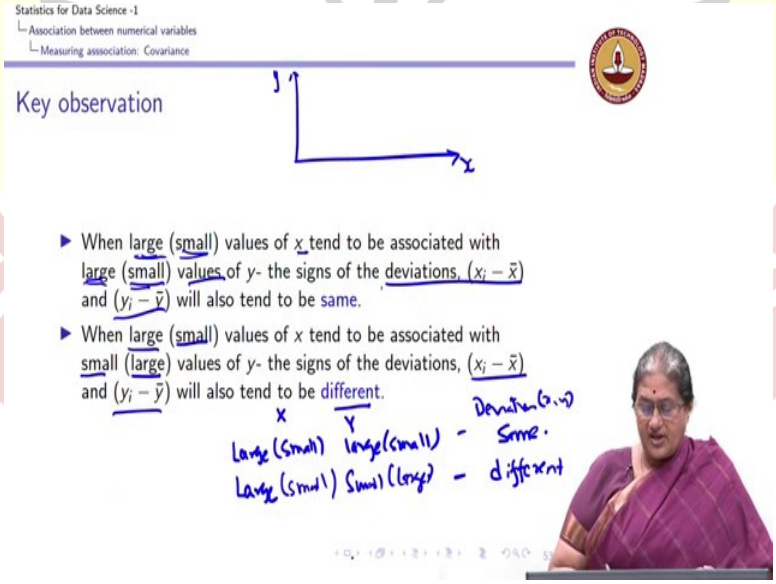
I know the product of these deviations is always going to have a negative sign. So you can see that - 2 × 2 is a - 4. I have a - 1 which is - 1 × 1, 1 × - 1 is again - 1, 2 × - 2 is - 4. So I can say that the key observations we are trying to make is when I am coming up with a product of deviations.

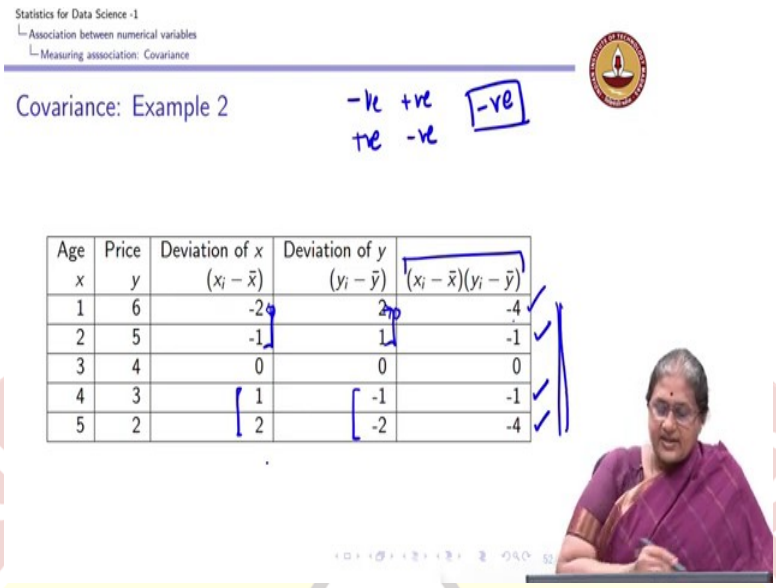
(Refer Slide Time: 17:29)



When I have data of this kind, this is one extreme and the other extreme is I have that data where my deviations are when it is positive, it is positive and when the deviation is negative, it is negative. That is, I have a match of deviation. There could be a case where I could have a negative deviation and a positive deviation. We look at it very soon.

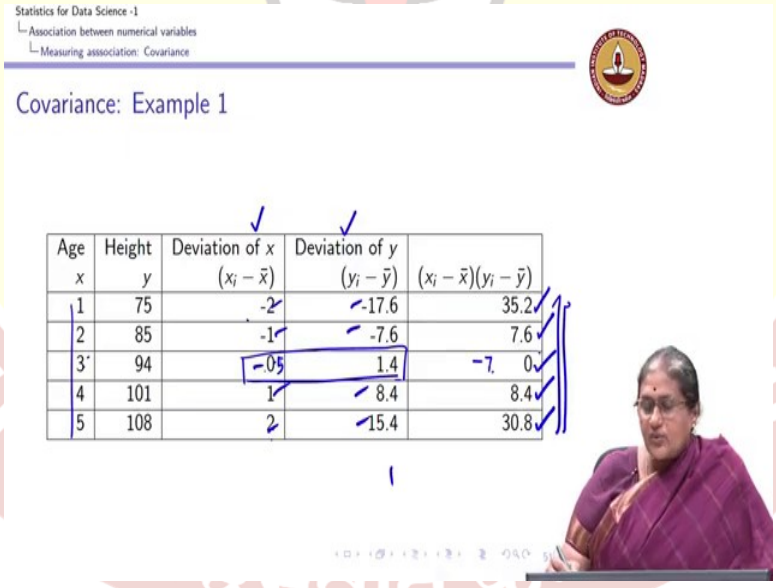
(Refer Slide Time: 17:54)





So now you can see that the key idea we have defined is, when I have this deviation the product of the deviation could be positive or it could be negative. In extreme cases here, I have all the product of the deviations negative.

(Refer Slide Time: 18:15)



And in the earlier example, I had all the products to be positive, but I could have a situation where for example if this was instead of 3, if it was 3.5, sorry, if it was a different age, I could have an example where just a hypothetical situation that this instead of 0 was or say some 0.5, we should make appropriate changes here. In that case I could have so this was or a - 0.5, I could

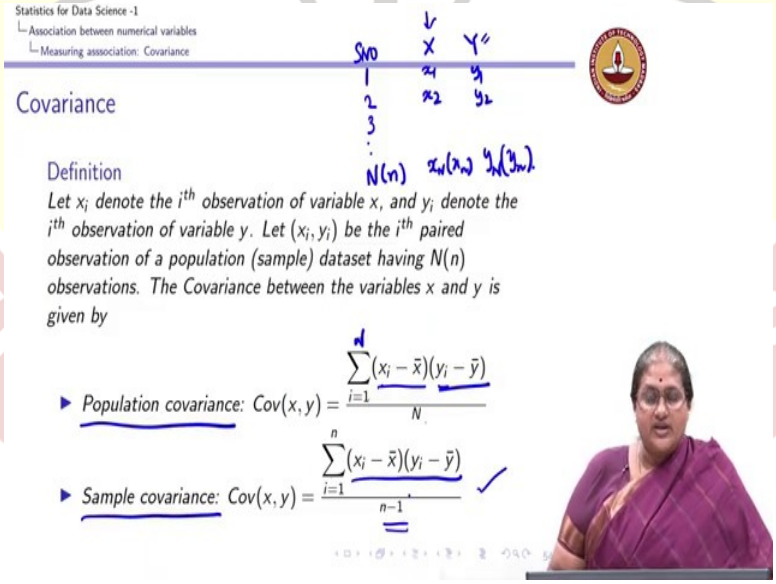
have here a case where the product of - 0.5, so I will have a 1.4 × 0.5 which is about 7, but it would have a negative sign.

This could be rare but mostly I would find cases which are positive. I might have a few cases which are negative product deviation. So how would I measure the strength of an association? Remember by strength of an association we are seeking the answer to: do both the variables increase together, do they decrease together is the association linear, do I have outliers? These are the questions which we are trying to answer.

So one way is, if as in the earlier example, if this was a -7 then you can see that this negative would have cancelled out with the positives, if everything were positive, then I could have told a story. So one way to quantify this measure is take the sum of all the deviations, the sum of these deviations will tell us what is the strength of the association.

Now the sum over, so I am just not looking at the sum of the deviations, I need to look at the sum of the deviations which would be my numerator and I have to divide it by the number of observations.

(Refer Slide Time: 20:36)



So we formally define what is a covariance measure. So if I have my dataset where I have my 1, 2, 3, these are the number of my observations. I have a N if I refer to a population this is from our earlier lectures. I would have a small n if I am referring to my dataset which comes from a

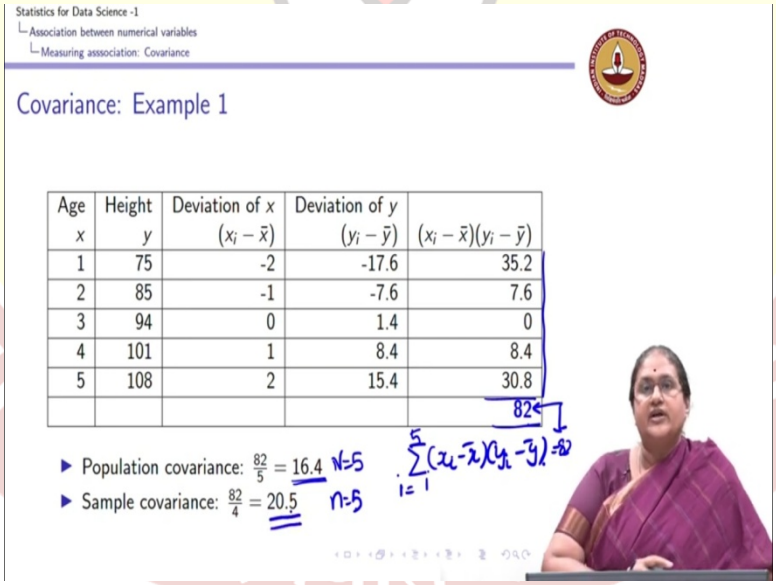
sample. I have my x variable. I have my y variable, x_1, x_2,..., x_N or x_n are the N observations of the first variable, y_1, y_2,..., y_N or y_n are the observations from the second variable.

Then $x_i - \bar{x}$ is the deviation of my first variable from its mean, $y_i - \bar{y}$ is a deviation of my second variable from its mean. This is the product of deviation. I sum up the product of deviations over all possible values and divide by the total number of values I get what I refer to as the population covariance.

Similarly, if I sum up the product of deviations over all the sample values, n is the sample value I divide it by n - 1. Recall when we defined the sample standard deviation and the sample variance also we divided it by n - 1 and at that point of time, I said when you are referring to a population you divide it by N, if you are referring to a sample you divide it by n - 1.

We use the same thing when we refer to a sample covariance. We divide the numerator with n - 1 observations and this quantity is referred to as the sample covariance. So how do I compute the sample covariance?

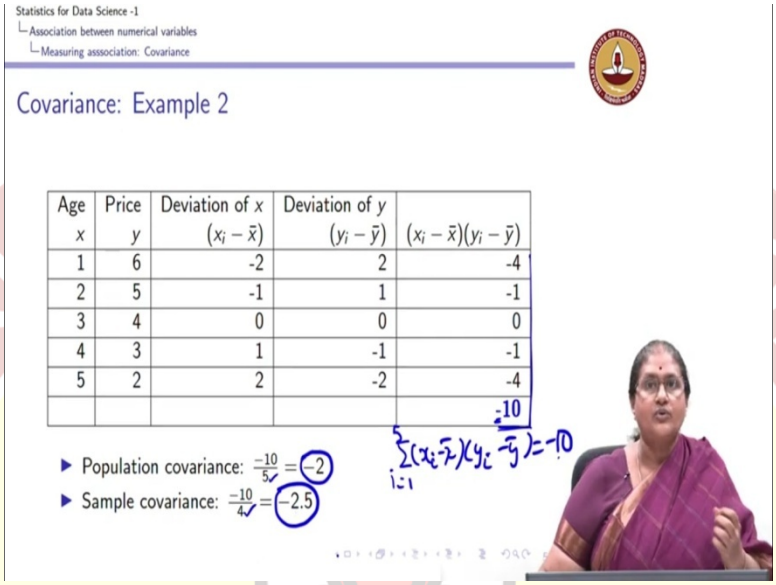
(Refer Slide Time: 22:50)



So now you can see that for the first example, I add all my quantities. I get 82. What is this 82? $82 = \sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})$. That is what is 82 of these 5 observation is equal to 82, which is given by this. So if I am interested in finding the population covariance my N is equal to 5. I divide 82 by 5 and I have 16.4 which will give me my population covariance whereas if I want to

find out my sample covariance, I know again n equal to 5. I divide 82 by n - 1 which is a 4 I get a sample covariance of 20.5.

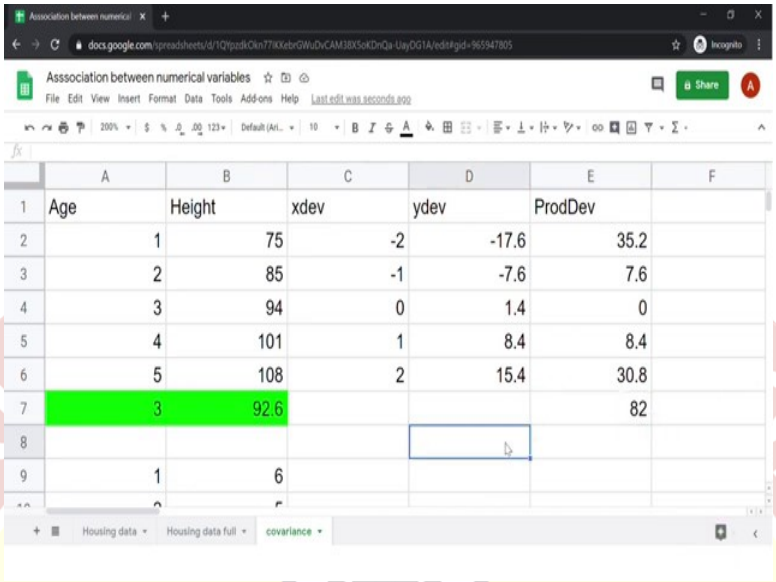
(Refer Slide Time: 23:57)

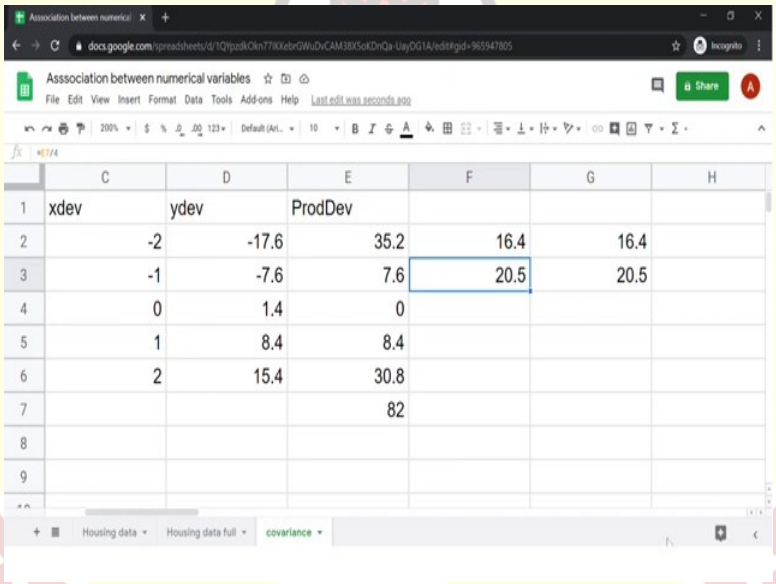


So you can see the same thing for the second example, you can see that when I add up my deviations, I have -10 as the sum. So $\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})$ = -10. The sign is very important. Again, if I am interested in that population covariance, I divide it by 5. If I am interested in the sample covariance, I divide it by 4, you can again see there is a sign the population covariance in the second example is -2 where a sample covariance is -2.5.

So what is the sample covariance give? It gives us a quantified measure when two variables are moving in the same direction, the covariance is a positive measure whereas when two variables are moving in opposite direction, what do we mean by opposite direction? As one variable increases the other variable decreases then my sample covariance is negative. So this is how you quantify the covariance. How does Google Sheets give you the covariance?

(Refer Slide Time: 25:18)





Let us go back to Google Sheets. Go to this sheet here. So let us look at the first example here. So I have my first example, which is given by this quantity here. So you can you recall. This was again my age in years and this was the price, sorry, this was the height in centimetres. This is 3 is my $\bar{x}$, so I can find my deviation of x.

So I will call that x deviation or x dev. That is what I am going to refer to and I will also refer to as y deviation, $y_i - \bar{y}$ as y dev. So x deviation, I can find out what is my x deviation, x deviation is going to be x_i which is my age - my mean.

Since I am going to have the same mean I am just going to, this is what I have and you can see that I have $x_i - \bar{x}$ and $y_i - \bar{y}$. These are my deviation. This is precisely what we had in our, you can see this is for the first example I have -2, -17.6. So that is what my Google Sheets give me. I am doing the example right from beginning. I can find the product of the deviation and that is nothing but I can find the product of the deviation.

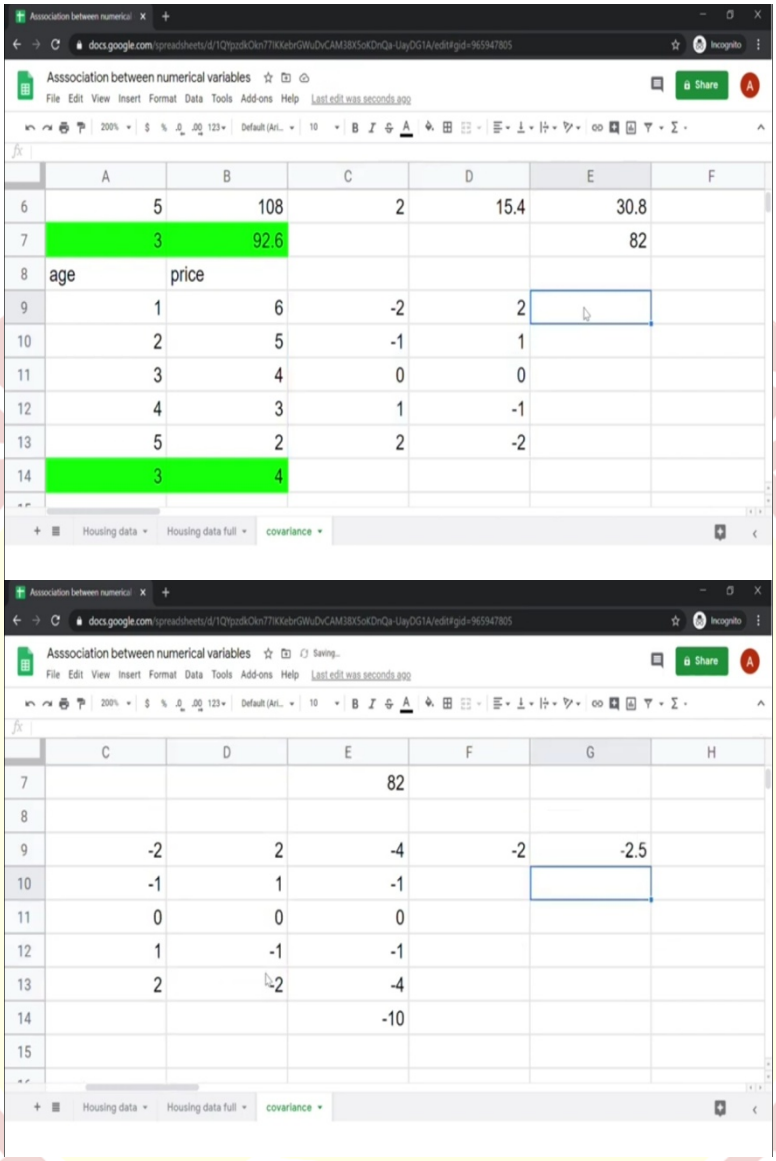
The product of the deviation, I can write that down and you can see that this product of the deviation is precisely what we have here, 35.2 up to 30.8. Now I can find out what is the sum of these deviations and if I divide the sum by I can divide this sum by 5, I can divide this by 5 I get my population covariance. I divide this sum by 4, I get my sample covariance.

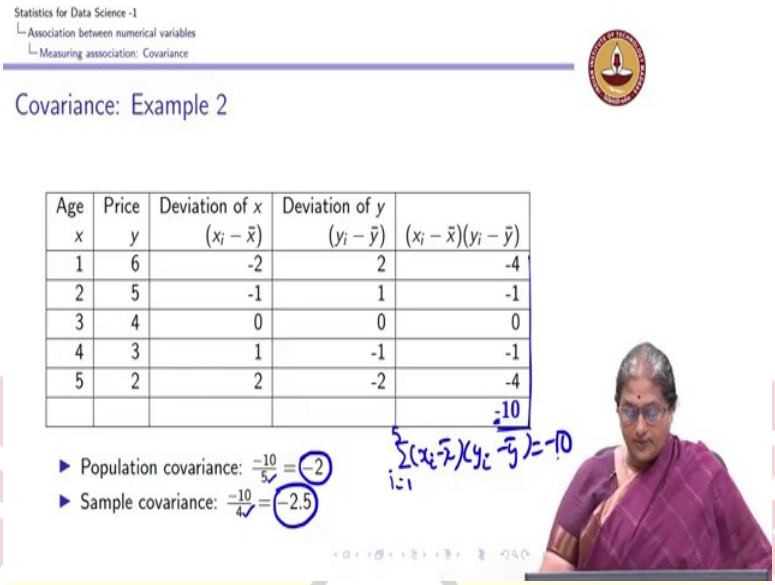
Now, there is a function n in our Google Sheets. If you go to a Google sheet and type covariance, you can see that there are 3 functions available which is covariance of a dataset, covariance.P, which is covar and covariance P, you make a population covariance of the datasets whereas covariance.S gives me the sample variance of a dataset. So let us see what covariance.P gives me. So I have covariance.P. So to find out the covariance, my first variable, I choose the first variable here.

The second variable is height. I choose that and then you can see that the covariance or covariance.P of my variables gives me precisely the population covariance which we have worked out from first principles. Similarly, the covariance sample covariance of both the dataset so again, I choose my age and I choose the height.

The sample covariance is equivalent to dividing this 82 by n - 1 which is 4, gives me the sample covariance. I can do the same thing for my next example also. We will just quickly go about it because we have already done it.

(Refer Slide Time: 29:39)

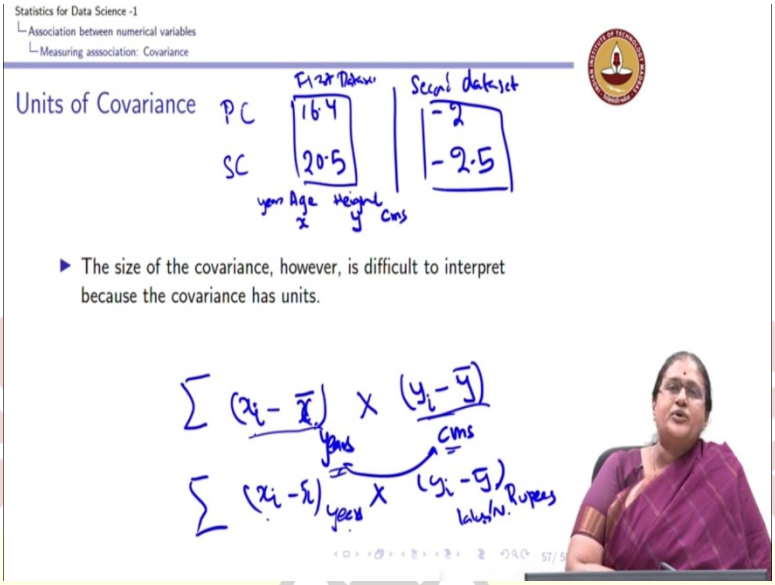


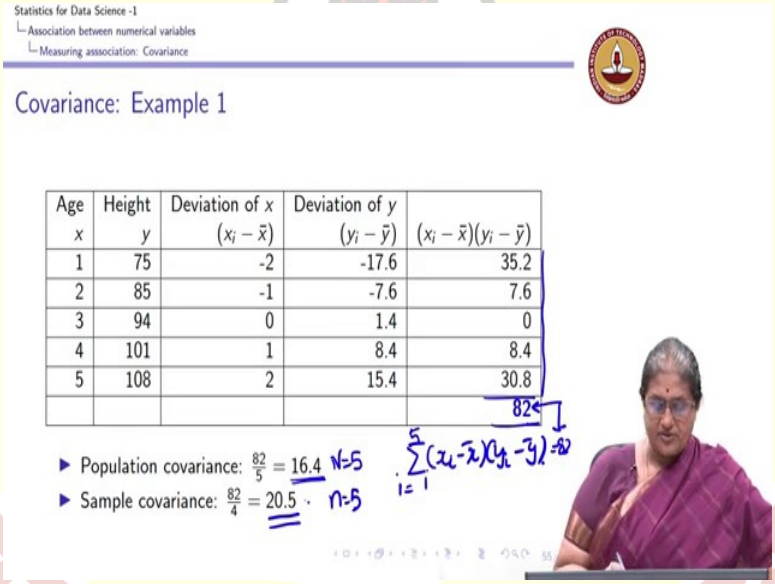


This is again this is my age of a car and this is the price of a car. So I can find out what is my, I will just do the deviations quickly. So this is going to be $x_i - \bar{x}$. I can freeze this. So we notice here again the deviations of x and y are of different size. I take the product of these 2 and I sum it up. I have a -10. Again this is very, this is what we did manually and you can see that, that is equal to -10 here.

So now what is my population covariance? My dataset is going to be age and price which you can see is -2 and -2 = -10/5 where 5 is the population size and similarly my sample covariance is again going to be again, if I compute my sample covariance for the same pairs of variable, I get it - 2.5 which is -10/4 and you can see that my sample covariance is - 2.5 which is consistent and the same would be obtained when we did it manually.

(Refer Slide Time: 31:29)





So now the question is what the covariance measure gives us is the size of an association. So in earlier case I had two sizes. So when we are looking at population or so for now, let us restrict ourselves to sample covariance. So for my second dataset and my first dataset, my population covariance and my sample covariance. So if you go back, my population covariance in this first dataset was 16.4 and 20.5 whereas here this was a - 2 and - 2.5.

So you can see that this is a positive measure whereas here this is a negative measure. Now the question is when you look at the first dataset my variables here was again age and height. My x variable was age, this was measured in years, my height y, it was measured in centimetres. If x is

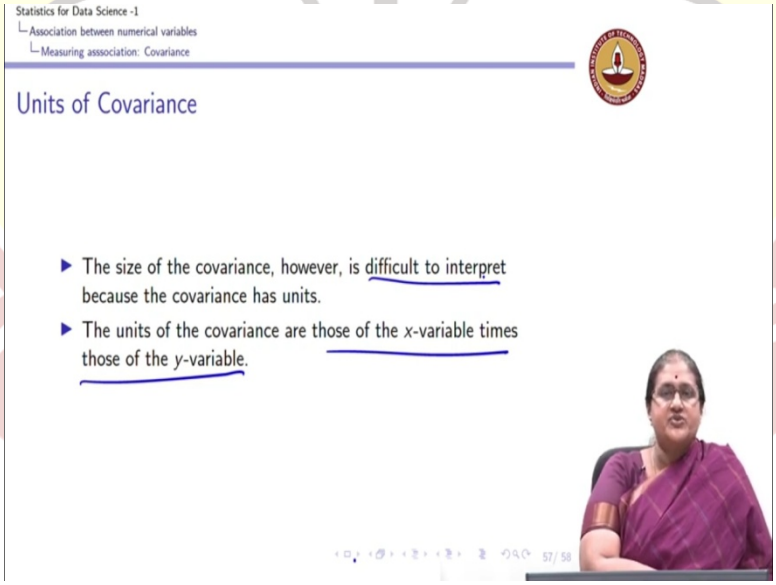
measured in age, I know $\bar{x}$ is also takes the same units as my original variable so $\bar{x}$ is also in years.

If $\bar{x}$ is in years, $x_i - \bar{x}$ is also in years. Similarly if each y_i is in centimetres, $\bar{y}$ is also in centimetres so $y_i - \bar{y}$ is in centimetres. So this $x_i - \bar{x}$, this deviation is in years and $y_i - \bar{y}$ is in centimetres. Hence, the product of these two variables is actually going to, the units are going to be a product of the units. There is a product of the units actually is a difficult thing to articulate.

Similarly when I look at this second dataset, my x_i is again in years. So my $x_i - \bar{x}$ is in years whereas my $y_i - \bar{y}$ for my second dataset is in currency or is it in lakhs of rupees, Indian national rupees in lakhs. So again, the product is going to have a unit of measure. Each product has a unit of measure. The summation over all the observations would also take the same units of measure.

So because of this it actually is very difficult to interpret the covariance measure. So the next natural thing to ask is can I have a unit-less measure so that it becomes easier for me to interpret the strength of association between two variables. The answer is yes.

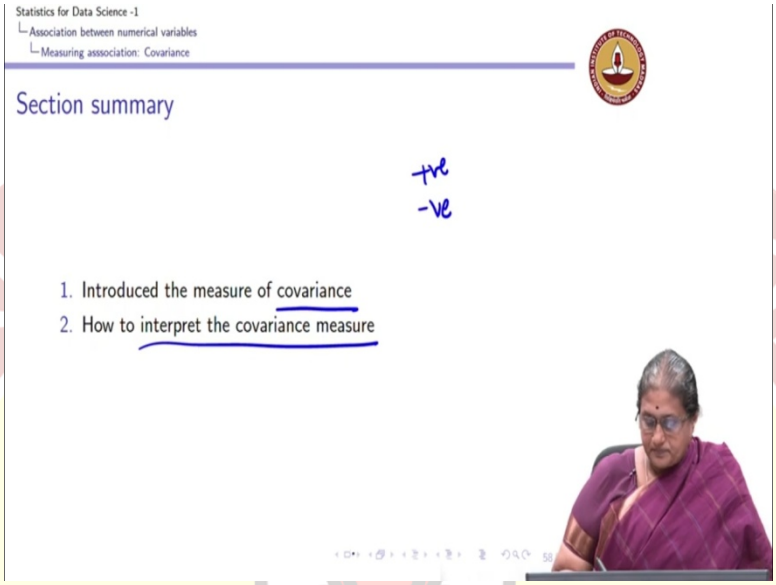
(Refer Slide Time: 34:49)



So the question is, is there a unit of measure since the units of covariance are those of the variable of x times the variable of y and it becomes difficult to interpret. Thought it gives a good measure of strength of association, very natural strength of association, the question is do I have

another measure which can help in interpreting the strength of association better. Answer is yes and that is the measure which we refer to as a correlation measure which we will be seeing next.

(Refer Slide Time: 35:26)



So in summary what we have learned so far is we introduced the measure of covariance. We saw what was the inclusion behind coming up with this measure of covariance and we saw how to interpret this covariance measure in a sense if it is positive, then in a sense we say that two variables might be moving up in the same direction and we assume linear because both covariance and correlation are measures of linear association.

And if the covariance is negative, it means that if 1 variable is moving in an up direction, the other variable is moving in the down direction, but however interpretation is difficult, hence, we seek another measure and that measured is what we refer to as the correlation measure.