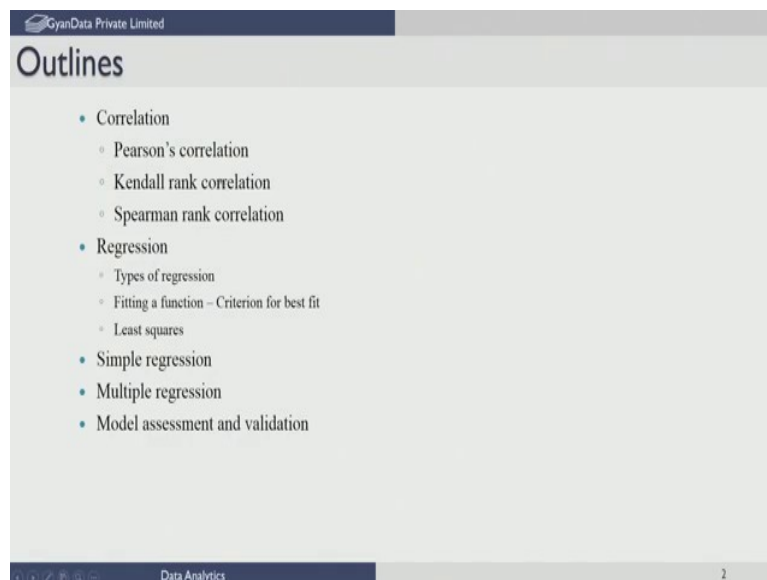


**Data Science for Engineers**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Madras**

**Lecture- 31**  
**Module: Predictive Modelling**

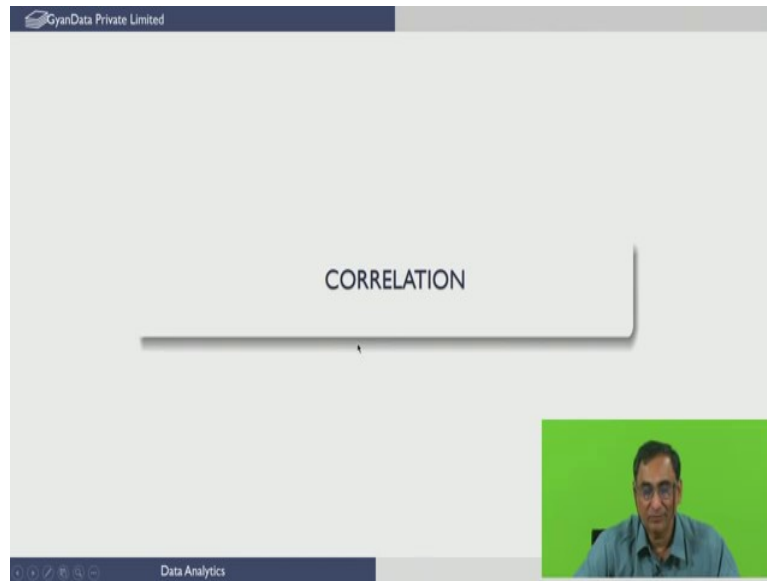
Welcome to the lectures in Data Analytics. In this series of lectures I am going to introduce to you model building; in particular I will be talking about building linear models using a techniques called regression techniques. So, let us start with some basic concepts. We are going to introduce the notion of correlation.

(Refer Slide Time: 00:38)



Different types of correlation coefficients that are been defined in the literature what they are useful for this is a preliminary check you can do before you start building models. Then I will talk about regression specifically linear regression and I will introduce the basic notions of regression and then take the case of 2 variables before taking going through multi linear regression where the several input variables and one dependent output variable. Finally, after building the model we would like to assess how well the model performs, how to validate some of the assumptions we have made and so on. So, this is called model assessment and validation.

(Refer Slide Time: 01:16)



So, let us first look at some measures of correlation. We have already seen one in the basic interactive lectures to statistics.

(Refer Slide Time: 01:26)

**Preliminaries**

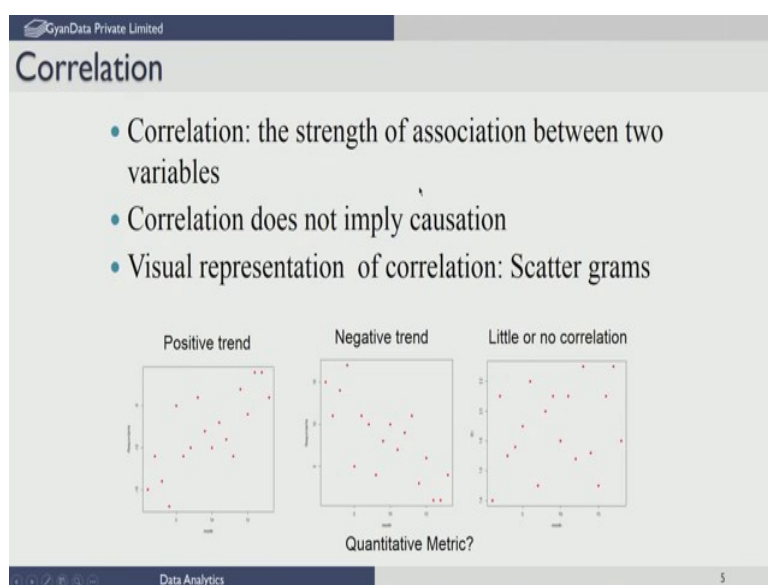
- $n$  observations for  $x$  and  $y$  variables  $(x_i, y_i)$
- Sample means  $\bar{x}$  and  $\bar{y}$   
$$\bar{x} = \frac{\sum x_i}{n} \quad \bar{y} = \frac{\sum y_i}{n}$$
- Sample variances  $S_{xx}$  and  $S_{yy}$   
$$S_{xx} = \frac{1}{n} \sum (x_i - \bar{x})^2 \quad S_{yy} = \frac{1}{n} \sum (y_i - \bar{y})^2$$
- Sample covariance  $S_{xy}$   
$$S_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

So, let us consider  $n$  observations of 2 variables  $x$  and  $y$ . So, denoted by the samples  $x_i$  comma  $y_i$  and we can of course, compute the sample means we have seen this in statistics which is just the summation of all values divided by  $n$  for  $x$  which is denoted by  $\bar{x}$  and similarly we can do the sample mean of  $y$  which is denoted by  $\bar{y}$ . We can also define sample variance, which is nothing but the sum square

deviation of the individual values from the respective means  $x_i - \bar{x}$  whole square summed over all the values divided by  $n$  or  $n - 1$  as the case may be.

So, we define these sample variance  $S_{xx}$  and  $S_{yy}$  corresponding to the variance of sample variance of  $x$  and  $y$ . You can also define the cross covariance which is denoted by  $S_{xy}$  which is nothing but the deviation of  $x_i$  from  $\bar{x}$  and the corresponding deviation of  $y_i$  from  $\bar{y}$  and the product of this we take and sum over all values and divide by  $n$ . Notice again that this  $x_i$  and  $y_i$  order test in the sense that corresponding to the experimental condition  $i$ th experiment; we have obtained values for  $x$  and  $y$  and that is that is what we have to take you cannot shuffle these values any which way they are known assumed to be corresponding to some experimental conditions that you have set. So, there are  $n$  experimental observations you have made.

(Refer Slide Time: 03:03)



Now, let us define what we call correlation. Correlation is nothing but the indicates the strength of association between the 2 variables. Of course, if you find the strong correlation it does not mean that a the one variable is a causation, the other is an effect you cannot treat correlation as a causation because there can be a third variable which is basically triggering these two and therefore you can only find the correlation, but cannot assume that one of the variable is a causation and the other is an effect.

We can also before we actually do numerical computation, we can check whether there is an association between variables by using what is called the scatter plot, we have done this before. So, we can plot the

values of  $x_i$  on the x axis  $y_i$  under y axis and for each of these points and we can see whether these points are oriented in any particular direction. For example, the figure on the left here indicates that the  $y_i$  increases as  $x_i$  increases there seems to be a trend in this. In particular we can say there is even a linear trend as  $x_i$  increases  $y_i$  corresponding the increase is in a linear fashion.

This is a positive trend because when  $x_i$  increases  $y_i$  increases. In the next figure the middle figure we show a case where  $x_i$  is as  $x_i$  increases  $y_i$  seems to decrease and again there seems to be a pattern association between  $x_i$  and  $y_i$  and this is a negative trend.

Whereas, if you look at the third figure the data that we find seems to be having no bearing on each other. That is  $x$   $y_i$  values do not seem to depend in any particular manner on the  $x_i$  values. When  $x_i$  increases maybe  $y_i$  increases for some cases and  $y_i$  decreases that is why it is spread in all over the place. So, we can say there is little or no correlation. This is a qualitative way of looking at it, we can quantify this and there are several measures that have been proposed depending on the type of variable and the kind of association you are looking for.

(Refer Slide Time: 05:03)

**Pearson's Correlation**

- $n$  observations for  $x$  and  $y$  variables ( $x_i, y_i$ )
- Pearson's product-moment correlation coefficient ( $r_{xy}$ )

$$r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}} = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

- $r_{xy}$  takes a value between -1 (negative correlation) and 1 (positive correlation)
- $r_{xy} = 0$  means no correlation

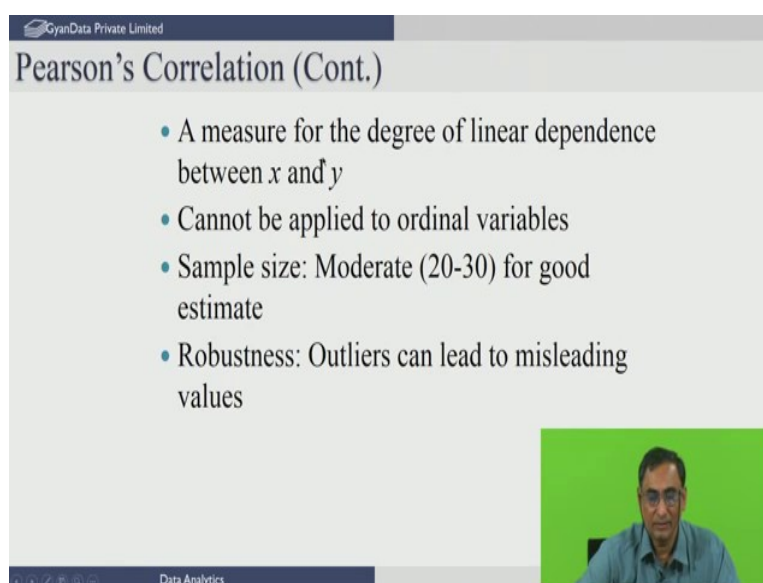
Data Analytics 6

So, let us look at the most common type of correlation which is called the Pearson's correlation. Here as we started with you have  $n$  observations for the variables  $x$  and  $y$  and we define the Pearson's correlation coefficient denoted by  $r_{xy}$  or sometimes denoted by  $\rho_{xy}$  by this quantity defined. Where we are essentially taking same thing that we did before the covariance between  $x$  and  $y$  divided by the standard deviation of  $x$  and the standard deviation of  $y$ .

The numerator represents the covariance between  $x$  and  $y$  can also be computed in this manner we can expand that definition we have for the covariance and we can find that it is nothing but the product of  $x_i y_i - n$  times the mean of  $x$  and the mean of  $y$  which represents the covariance of  $x$  and  $y$  and the denominator represents the standard deviation. We can look at this division by the denominator as what is called normalisation.

So, this value is now bounded. We can show that  $r_{xy}$  we take a any value between  $-1$  and  $+1$ .  $-1$  if it takes a value we say that the 2 variables are negatively correlated if  $r_{xy}$  takes a value close to one we say they are positively correlated, on the other hand if  $r_{xy}$  happens to value close to 0 it indicates that  $x$  and  $y_i$  have no correlation between them. Now, what how we can use this we will see.

(Refer Slide Time: 06:34)



GyanData Private Limited

### Pearson's Correlation (Cont.)

- A measure for the degree of linear dependence between  $x$  and  $y$
- Cannot be applied to ordinal variables
- Sample size: Moderate (20-30) for good estimate
- Robustness: Outliers can lead to misleading values

Data Analytics

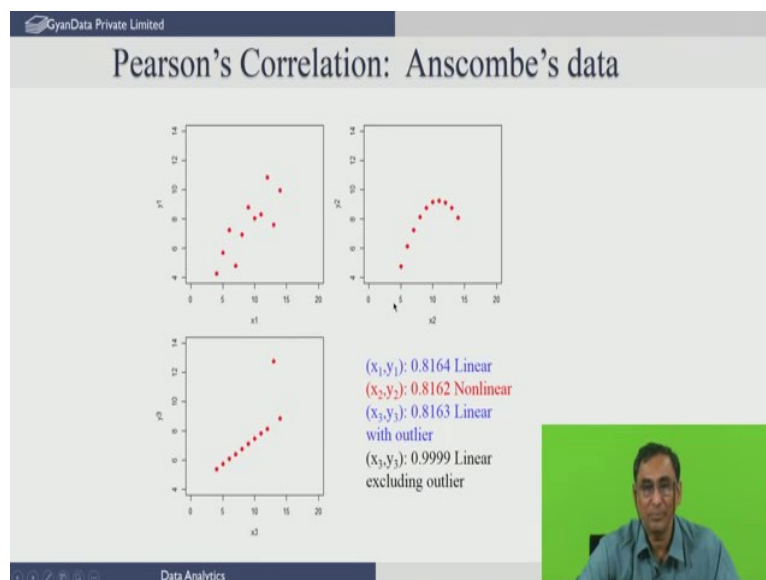
So, this correlation, Pearson's correlation, actually is useful to indicate there is a linear dependency between  $x$  and  $y$ . For example, if  $y_i$  is a linear function of  $x$  then the correlation coefficient or the Pearson's correlation coefficient will turn out to be either close to  $+1$  or close to  $-1$  depending on whether  $y$  is increasing with  $x$  then we say it is a positive correlation as we saw in a figure, if  $y$  is decreasing with  $x$  linearly then we say it is the correlation coefficient will be close to  $-1$ .

On the other hand if the correlation coefficient is close to 0 all we can conclude from then is perhaps there is no linear relationship between  $y$  and  $x$ , but perhaps there is non-linear association so, we will come to that a little later. The way it is defined we can also not apply it to ordinal variables which means ranked variable. So, suppose you

have a variable where you have indicated your scale on a scale of say 0 to 10 the let us say the course the those kind of variables are typically you do not apply a Pearson's correlation there are other kinds of correlation coefficients defined for what we call ranked or ordered variable ordinal variables.

Typically in order to get a good estimate of the correlation coefficient between y and x you need at least 20 - 30 points. That is generally recommended and then like the your sample mean or the sample variance standard deviation if there are outliers if there is one bad data point or experimentally you know experimental point which is wrongly recorded for example, that can lead to misleading values of this correlation coefficient. So, it is not robust with respect to outliers just like the sample mean and sample variance we saw earlier.

(Refer Slide Time: 08:24)



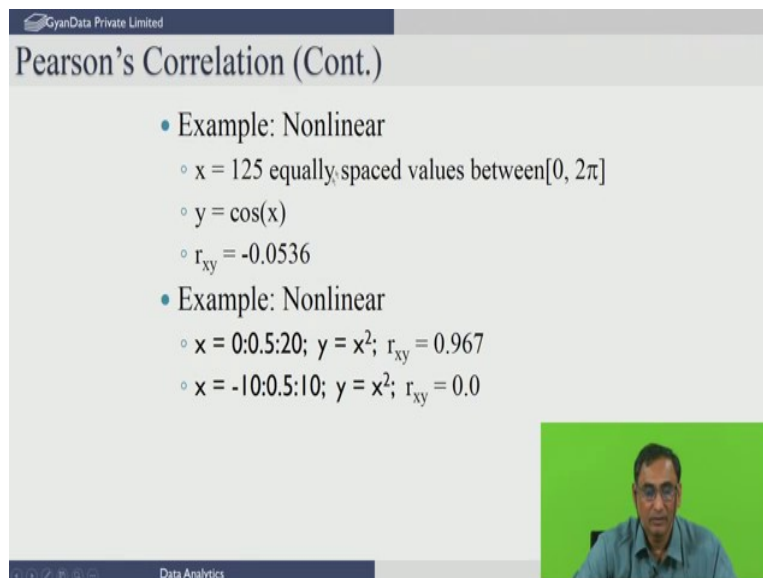
So, let us look at some sample examples this is a very famous data set called the Anscombe's data set. Here there are 11 data points for each of this there are 4 data set I have only shown 3 of them, each of them contains exactly 11 data points corresponding to  $x_i$  and  $y_i$  these points have been carefully selected. In the first one if you look at if you plot the scatter plot you will see that there seems to be linear relationship between y and x in the first data. In the second data if you look at this figure you can conclude that there is a non-linear relationship between x and y and the third one you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

So, if I apply Pearson's compute Pearson's correlation coefficient for each of these data sets we find that it is identical, does not matter

whether the, you actually apply into first data set or second data set or the third data set. In fact, the fourth data set has no relationship between  $x$  and  $y$  and it turns out to be they have the same correlation coefficient. So, what it seems to indicate is that if we apply the Pearson's correlation and we find the high correlation coefficient close to one in this case.

It does not immediately you cannot conclude there is a linear relationship. For example, this is a non-linear relationship and still gives rise to a high value. So, it is not confirmatory in the sense there is a linear you can say it is one way if there is a linear relationship between  $x$  and  $y$  then the correlation Pearson's correlation coefficient will be high. When I say high it can be  $-1$  or  $+1$ , but if there is a non-linear relationship between  $x$  and  $y$  it may be high or it may be low. We will see some data sets to actually show this illustrate point.

(Refer Slide Time: 10:18)



GyanData Private Limited

### Pearson's Correlation (Cont.)

- Example: Nonlinear
  - $x = 125$  equally spaced values between  $[0, 2\pi]$
  - $y = \cos(x)$
  - $r_{xy} = -0.0536$
- Example: Nonlinear
  - $x = 0:0.5:20$ ;  $y = x^2$ ;  $r_{xy} = 0.967$
  - $x = -10:0.5:10$ ;  $y = x^2$ ;  $r_{xy} = 0.0$

Data Analytics

Here are 3 examples. In the first example I have taken 125 equally spaced values between 0 and  $2\pi$  for  $x$  and I have actually computed  $y$  as  $\cos$  of  $x$ . So, this is a relationship between  $y$  and  $x$  in this case is a sinusoidal relationship. So, if you apply the Pearson's correlation coefficient compute the Pearson correlation coefficient for this data set you get a very low value close to 0 indicating as if there is no association between  $x$  and  $y$ , but clearly there is a relationship because it is non-linear.

In fact, it is symmetric the points above the 0 line when it is  $x$  is between 0 and  $\pi/2$  and when it is between  $\pi/2$  and  $\pi$  and between  $\pi$  and  $3\pi/2$  and  $3\pi/2$  and  $2\pi$  they all seem to cancel each other out and

finally, give you a correlation coefficient which is very small, does not indicate imply that there is no relationship between  $y$ . All you can conclude from this is perhaps there is no linear relationship between  $x$  and  $y$ .

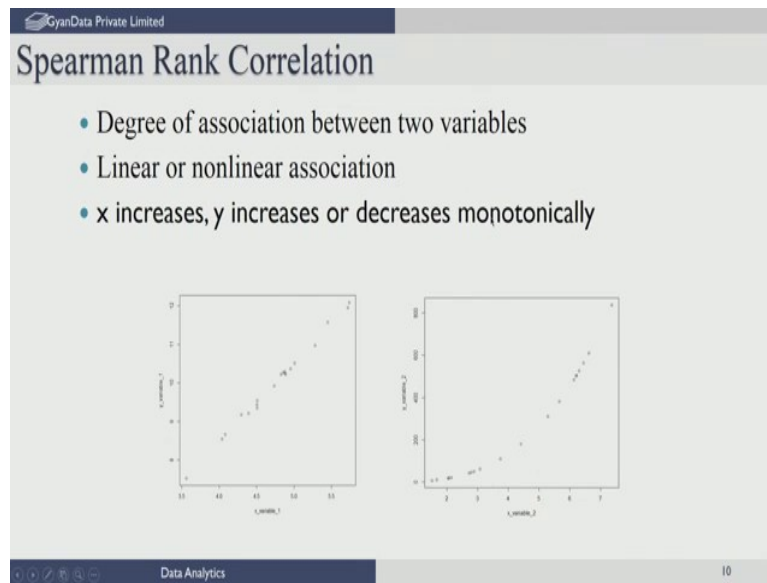
Similarly, let us look at another case where this non-linear  $y = x^2$ , where I have chosen  $x$  between 0 and 20 with equally spaced point of 0.5 each 40 points you get and I will computed  $y = x^2$  and then compute a Pearson's correlation for this  $x$   $y$  data. You find it is a very high correlation coefficient you cannot immediately conclude there is a linear relationship between  $x$  and  $y$ , you can only say there is a relationship perhaps it is linear may be it is even non-linear we have to explore further.

If it is close to 0 I would have said there is no linear relationship, but it is in this case it is very high. So, there is some association, but perhaps the association you cannot definitely conclude it is linear it may be non-linear. On the other hand if the data if I chosen my  $x$  data between - 10 and 10 symmetrically for between - 10 and 0  $y = x$  square will have positive values between 0 and 10  $y = x$  square will have positive values again although  $x$  is positive  $y$  is positive in this range and between negative values for  $x$   $y$  is still positive. So, these will cancel each other out exactly and will turn out that the correlation coefficient in this case is 0 although there is a non-linear relationship between  $y$  and  $x$ .

So, all we are saying is this if there exists a linear relationship between  $y$  and  $x$  then the Pearson's correlation coefficient will be either close to 1 or - 1 perfect relationship linear relationship. On the other hand if it is close to 0 you cannot dismiss a relationship between  $y$  and  $x$ . Similarly if a value is high looking at just the value we cannot conclude that there definitely exists a linear relationship between  $y$  and  $x$ , you can only say there exists a relationship between  $y$  and  $x$ . So, let us actually look at other correlation coefficients, you should note that Pearson's correlation coefficient can be applied only to what we call not ranked variables ordinal variables real value variables like we have here.



(Refer Slide Time: 13:18)



So, let us look at other correlation coefficients that can be applied even to ordinal variables. Now here is a case where we only look at the degree of association between 2 variables, but this time the relationship may be either linear or non-linear if x increases y increases or decreases monotonically then the Spearman's Rank Correlation will tend to be very high.

So, here is a case when x increases y increases this also is a case when x increases y increases monotonically, but in this case the right hand figure is a non-linear relationship the left hand figure indicates a linear relationship. Let us apply the Spearman's rank correlation to a same data set and see what happens.

(Refer Slide Time: 13:57)

GyanData Private Limited

## Spearman Rank Correlation

- Spearman rank correlation computation for n observations:  

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$
 $d_i$  is the difference in the ranks given to the two variables values for each item of the data
- Example:

Number	1	2	3	4	5	6	7	8	9	10
$X_i$	7	6	4	5	8	7	10	3	9	2
$Y_i$	5	4	5	6	10	7	9	2	8	1
Rank $X_i$	6.5	5	3	4	8	6.5	10	2	9	1
Rank $Y_i$	4.5	3	4.5	6	10	7	9	2	8	1
$d^2$	4	4	2.25	4	4	0.25	1	0	1	0

$r_s = 0.88$

Data Analytics

So, in the Spearman's rank correlation what we do is convert the data even if it is real value data to what we call ranks. So, for example, let us say i have 10 data points in this case  $x_i$  is like a somebody has taken a scale between let us say  $y_2$  and 10, 1 and 10 and similarly y is also a value between 1 and 10. So, what we have done is looked at all the individual values of x and assigned a rank to it for example, the lowest value in this case x value is 2 and it is given a rank 1 the next highest x value is 3 that is given a rank 2 and so on and so forth.

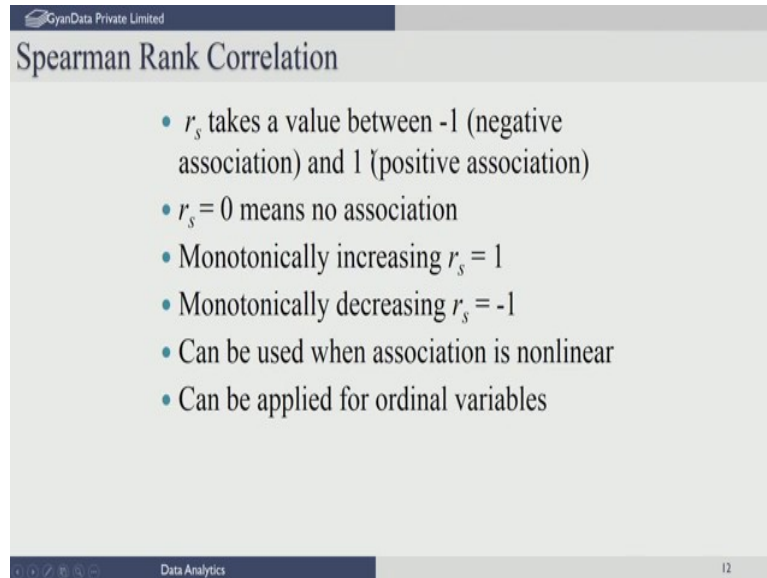
So, we are ranked all of these points notice that the sixth and the first value both are tied. So, they get the rank 6 and 7 which is the midway, the half of it. So, we have given it a rank of 6.5 because there is a tie. Similarly if there are more more than 2 values which are tied we take all these ranks and average them by the number of data points which have equal values and correspondingly you have to in the rank. We also ranked the corresponding y values for example, in this case the tenth value has a rank 1 and so on so forth, eighth value has a rank 2 and so, on.

So, we have given a rank in a similar manner now once you have got the rank you compute the difference in the ranks. So, in this case the difference in the rank for the first data point is 2 and we square it, similarly we take the difference in the second data point in the ranks between  $x_i$  and  $y_i$  which is 2 and square it we get 4.

So, like this we take the difference in the ranks square it and we get the final what we call the d squared values we sum over all values and then we compute this coefficient. It turns out that this coefficient also will be lie between - 1 and + 1 and - 1 indicating a negative association

and + 1 indicating a positive association between the variables and in this particular case the rank the Spearman rank correlation turns out to be 0.88.

(Refer Slide Time: 16:03)



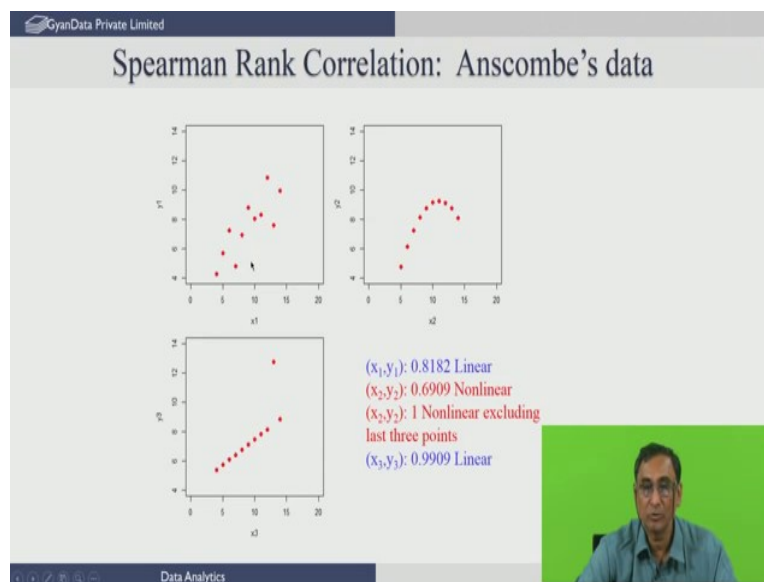
The slide is titled "Spearman Rank Correlation" and is part of a presentation by GyanData Private Limited. It lists several key properties of the Spearman rank correlation coefficient,  $r_s$ .

- $r_s$  takes a value between -1 (negative association) and 1 (positive association)
- $r_s = 0$  means no association
- Monotonically increasing  $r_s = 1$
- Monotonically decreasing  $r_s = -1$
- Can be used when association is nonlinear
- Can be applied for ordinal variables

The slide footer includes navigation icons, the text "Data Analytics", and the page number "12".

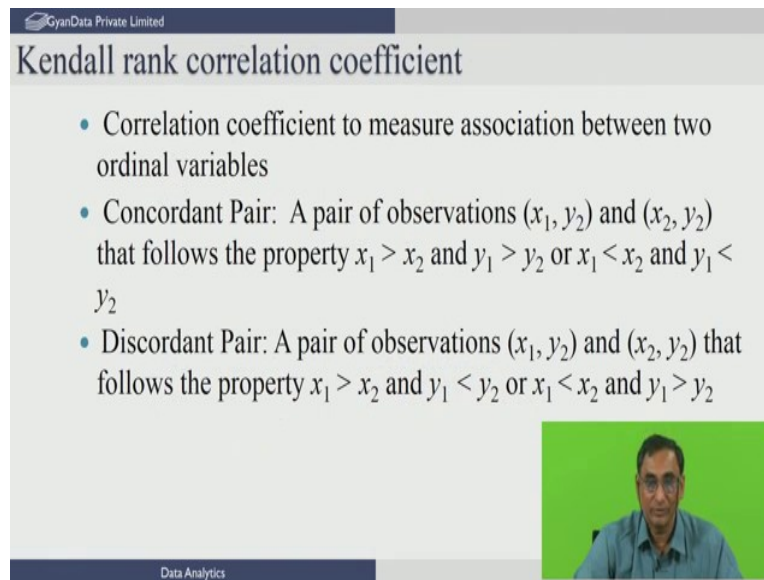
Let us look at some of the things as I said 0 means no association. When there is the positive association between y and x then the r s values or the Spearman's thing will be + 1 like the Pearson's correlation and similarly when y decreases with x then we say that you know the Spearman's rank correlation is likely to be close to - 1 and so, on. The difference is between Pearson's and Spearman is not only can it be applied to ordinal variables even if there is a non-linear relationship between y and x the spearman rank correlation can be high it will not likely to be 0, it will have a reasonably high value. So, that can be used to distinguish maybe to look for the kind of relationship between y and x.

(Refer Slide Time: 16:51)



So, let us apply it to the Anscombe data set in this case also we find that the, for the first one the Spearman rank correlation is quite high in the second one also reasonably high. In fact, the Pearson also a sign notice that the Pearson was same for all this and the third one also is fairly high 0.99. So, all of these things it is indicating that there is a really strong association between x and y.

(Refer Slide Time: 17:18)




GyanData Private Limited

### Kendall rank correlation coefficient

- Correlation coefficient to measure association between two ordinal variables
- Concordant Pair: A pair of observations  $(x_1, y_1)$  and  $(x_2, y_2)$  that follows the property  $x_1 > x_2$  and  $y_1 > y_2$  or  $x_1 < x_2$  and  $y_1 < y_2$
- Discordant Pair: A pair of observations  $(x_1, y_1)$  and  $(x_2, y_2)$  that follows the property  $x_1 > x_2$  and  $y_1 < y_2$  or  $x_1 < x_2$  and  $y_1 > y_2$

Data Analytics



Suppose we had applied I would suggest that you apply it to the  $\cos x = y$  example and  $y = x^2$  example you will find that the Spearman rank correlation for these will be reasonably high it may not be close to one, but it will be high indicating there is some kind of a non-linear relationship between them even though Pearson's correlation might be low. So, third type of correlation coefficient that is used for ordinal variables is called the Kendall's rank correlation and this correlation coefficient also measures the association between ordinal variable. In this case what we define is a concordant and a discordant pair.

If you look at the values. Compare 2 observations let us say  $x_1, y_1$  and  $x_2, y_2$ , if  $x_1$  is greater than  $x_2$  and the corresponding  $y_1$  is greater than  $y_2$  then we say it is a concordant pair; that means, if  $x$  increases and  $y$  also correspondingly increases then these 2 data points are known said to be concordant. Similarly if  $x$  decreases if  $x_1$  is less than  $x_2$  and  $y_1$  is less than  $y_2$  then also we say it is a concordant pair; that means, when  $x$  increases  $y$  increases or  $x$  decreases or  $y$  decreases then we say these 2 data pairs are concordant.

On the other hand if there is an opposite kind of relationship, so, if you take 2 data points  $x_1, y_1$  and  $x_2, y_2$  and we say that you know we look at the data points and find that if  $x_1$  is greater than  $x_2$ , but the corresponding  $y_1$  is less than  $y_2$  or if  $x_1$  is less than  $x_2$ , but  $y_1$  is greater than  $y_2$  then we say it is a discordant pair. So, we take every pair of observations in your sample and then assign whether there is a concordant or discordant pair let us take an example and look at it.

(Refer Slide Time: 19:11)

GyanData Private Limited


## Kendall rank correlation coefficient

- Kendall rank correlation coefficient

$$\tau = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{n(n-1)/2}$$

- The pair for which  $x_1 = x_2$  and  $y_1 = y_2$  are not classified as concordant or discordant and are ignored.

Data Analytics



So, once we have the number of concordant pairs and number of discordant pairs we take the difference between them divide by  $n$  into  $n - 1$  by 2 and that is called the Kendall's  $\tau$ .

(Refer Slide Time: 19:25)

GyanData Private Limited

## Kendall rank correlation coefficient


Example: Two experts ranking on food items

Items	Expert 1	Expert 2
1	1	1
2	2	3
3	3	6
4	4	2
5	5	7
6	6	4
7	7	5

	1	2	3	4	5	6	7
1							
2	C						
3	C	C					
4	C	D	D				
5	C	C	C	C			
6	C	C	C	D	D		
7	C	C	C	C	D	D	

$$\tau = \frac{15 - 6}{21} = 0.42857$$

Data Analytics



We can take a item here there are about 7 observations let us say 7 different wines or tea or coffee and there are two experts who ranked the taste of the tea or coffee or wine on a scale between 1 to 10. For the first the expert number 1 gives it a rank of 1 and expert 2 also ranks it 1 for the second one the expert 1 ranks it 2 while the expert 2 ranks it in a scale or gives it the value of 3 and so on so forth for the 7 different types of thing.

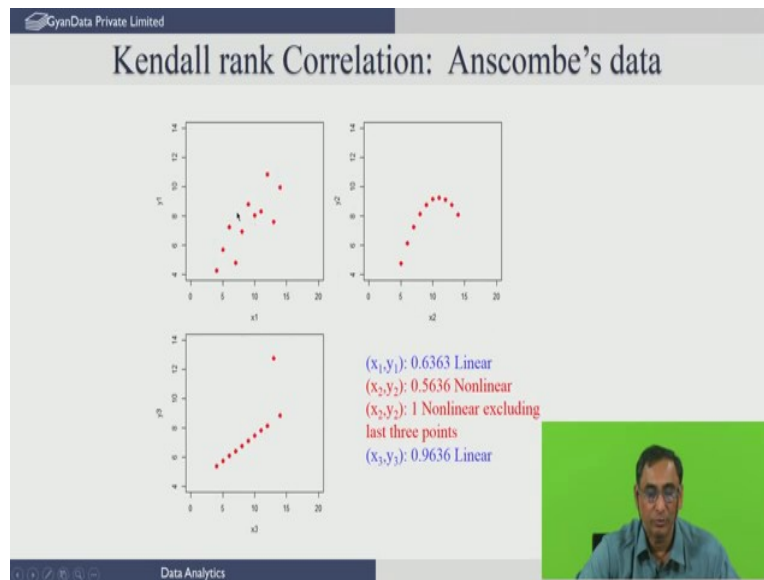
Now, you compare data point 1 and data point 2. In this case experts opinion is that 2 is let us say better than 1, expert 2 also says 2 is better than 1. So, it is a concordant pair. So, 1 and 2 are concordant that is what is indicated here. Similarly if I look at the data point 1 and 3 expert 1 says it is better 3 is better than 1, expert 2 also says 3 is better than 1. So, it is a concordant pair similarly if you look at 2 and 3 both agree in agreement 3 is better than 2, 3 is better than 2 so, it is a concordant.

Let us look at the fourth and the first one looks like expert 1 says it is better, expert 2 also says it is better concordant, but the second and fourth if you compare expert 1 says it is better fourth one is better than the second, but expert 2 disagrees he says the fourth thing is worse than the second one. So, there is a discordant pair of data that is indicated by D. So, 4 and 2 are discordant 4 and 3 are discordant.

Similarly here it says 5 and 1 are concordant 5 2 are concordant and so, on so, forth. So, between every pair  $n$  into  $n - 1$  by 2 pairs you will get and we have classified all of these pairs as either concordant or discordant and we find there are 6 discordant pairs and 15 concordant pairs and we can compute the Kendall's  $\tau$  coefficient. This basically says if this is high then there is broad agreement between the two experts right.

So, basically we are saying  $y$  and  $x$  are associated with each other also there is a strong association. Otherwise it is not strongly associated or completely if the expert 2 completely disagrees with expert 1 you might get even negative values. So, the high negative value or high positive value indicates that the 2 variables  $x$  and  $y$  in this case are associated with each other. Again this can be used for ordinal variables because it can work with ranked values here as we have seen in this example.

(Refer Slide Time: 21:56)



So, again if we apply it to Kendall's rank to this Anscombe data set we find that although it has decreased for this linear case the value has decreased as compared to the Pearson and Spearman correlation coefficient, it still has a reasonably high value. High in this case typically you in experimental data you cannot expect to get a value I mean beyond 0.6 or 0.7. You should consider yourself fortunate typically because we rarely know the nature of the relationship between variables we are only trying to model them.

So, in this case non-linear relationship you find again a reasonably high correlation coefficient for Kendall's rank and the last one again it is linear perfectly linear. So, you are getting very high association between them. So, really speaking you can actually use this to get a preliminary idea before you even build the model. Of course, for 2 variables it is easy you can plot it you can compute these correlation coefficient and try to get a preliminary assessment regarding the type of association that is likely to be and then try go ahead and choose the type of models you want to build and this is the first thing that we look before we jump into linear regression.

So, see you next class about how to build the linear regression model.

Thank you.