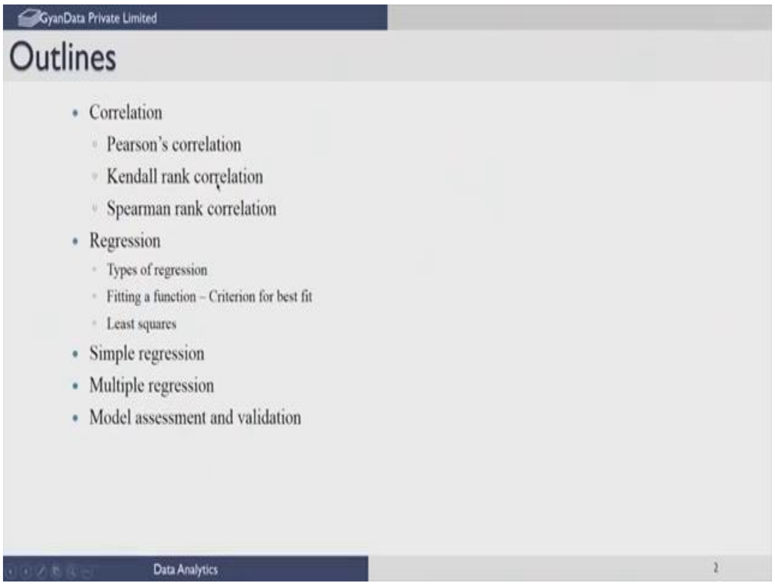
Python for Data Science

Prof. Raghunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 29
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: 00:36)



GyanData Private Limited

Outlines

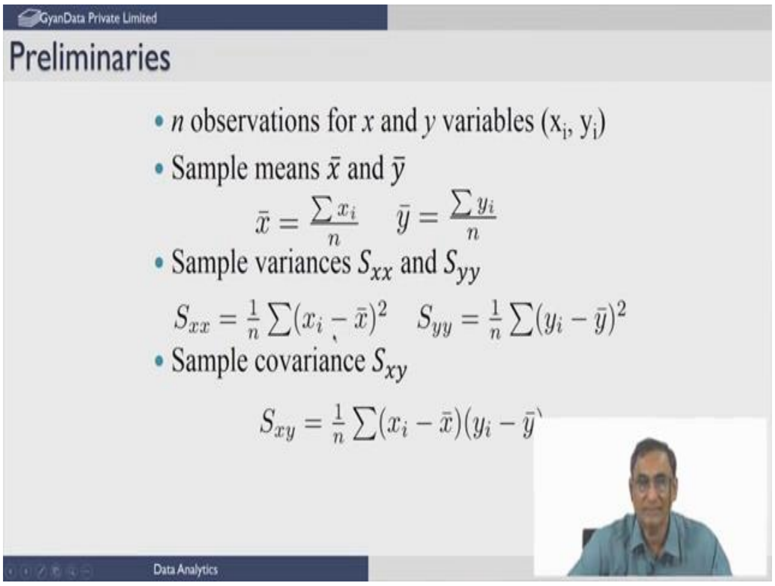
- Correlation
  - Pearson's correlation
  - Kendall rank correlation
  - Spearman rank correlation
- Regression
  - Types of regression
  - Fitting a function – Criterion for best fit
  - Least squares
- Simple regression
- Multiple regression
- Model assessment and validation

Data Analytics 2

Different types of correlation coefficients that have 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 two variables; before taking going through multi linear regression where there are 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. So, let us first look at some measures of correlation, we have already seen one in the basic introductory lectures to statistics.

(Refer Slide Time: 01:22)



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}

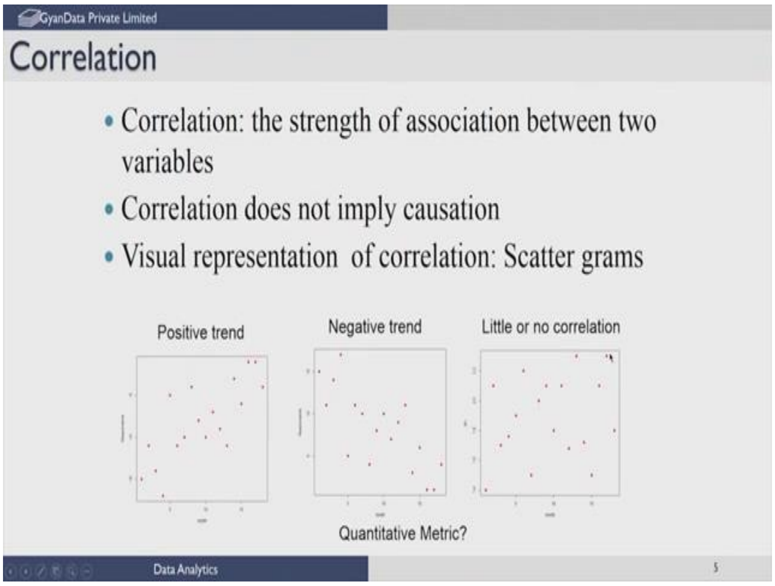
$$S_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}$$

So, let us consider n observations of 2 variables x and y. So, denoted by the samples $x_{i,}$ 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 variances 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 variances S_{xx} and S_{yy} corresponding to the variance of sample variance of x and y. We 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 pairs in the sense that, corresponding to the experimental condition i th experiment we have obtained values for x and y and that is what we have to think we cannot shuffle these values any which way, they have 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: 03:02)



GyanData Private Limited

Correlation

- Correlation: the strength of association between two variables
- Correlation does not imply causation
- Visual representation of correlation: Scatter grams

Quantitative Metric?

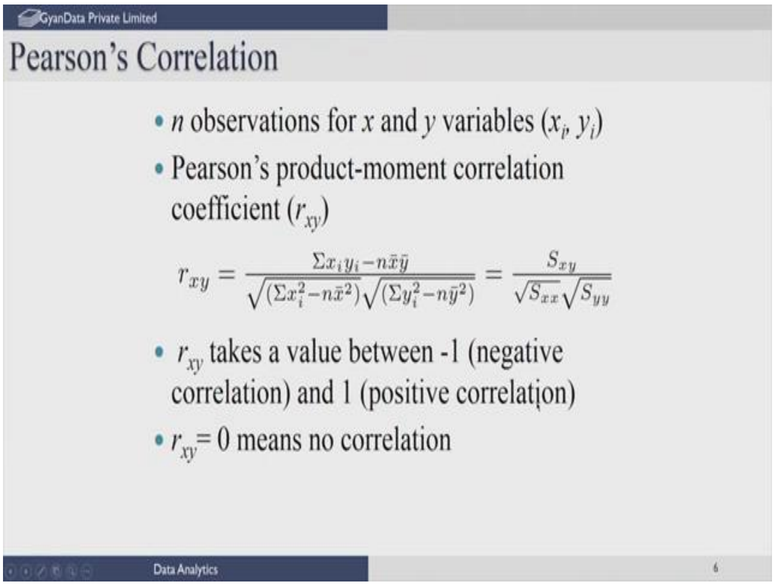
Data Analytics

Now let us define what we call correlation, correlation is nothing but the indicates the strength of association between the two variables; of course, if you find a strong correlation it does not mean that the one variable is a cause and 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 variables is a cause 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 on the 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 xi increases y_i correspondingly increases in the 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 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 sometimes 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: 05:01)



GyanData Private Limited

Pearson's Correlation

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

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

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

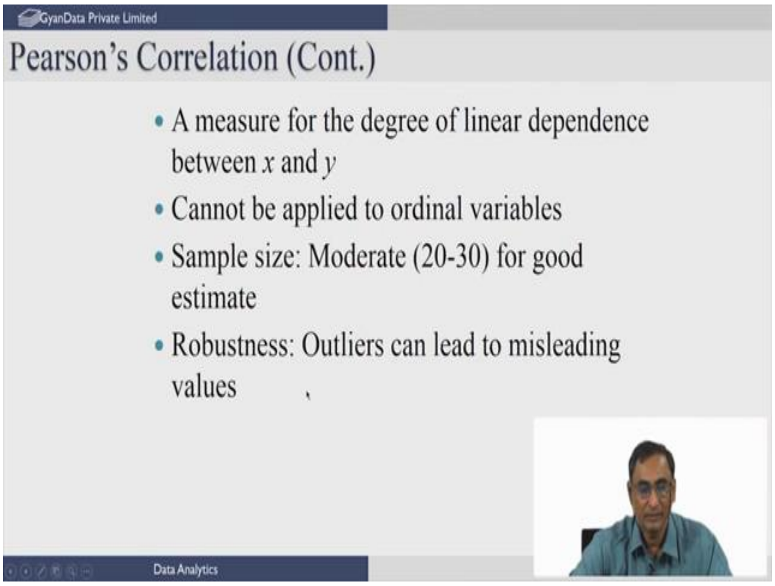
Data Analytics 6

So, let us look at the most common type of correlation, which is called the Pearson's correlation. Here as we started with, we 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 normalization.

So, this value is now bounded, we can show that r_{xy} will take any value between -1 and +1 -1, if it takes a value we say that the two variables are negatively correlated; if r_{xy} it takes a value close to 1, 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 have no correlation between them. Now what how we can use this, let us see.

(Refer Slide Time: 06:30)



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

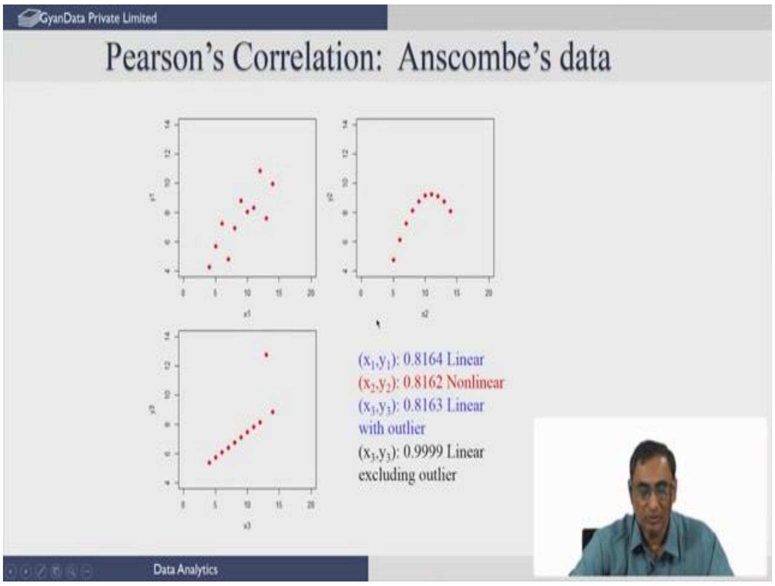
Data Analytics

So, this correlation Pearson's correlation actually is useful to indicate there is a linear dependency between x and y; for example, if y is a linear function of x, then we the correlation coefficient of 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 a 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 them 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 variables. So, suppose you have a variable where you have indicated your scale, on a scale of say 0 to 10 ok; the let us say the course, the those kind of variables are tyπcally 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.

Tyπcally 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 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: 08:23)



Pearson’s Correlation: Anscombe’s data

(x_1,y_1): 0.8164 Linear
(x_2,y_2): 0.8162 Nonlinear
(x_3,y_3): 0.8163 Linear with outlier
(x_3,y_3): 0.9999 Linear excluding outlier

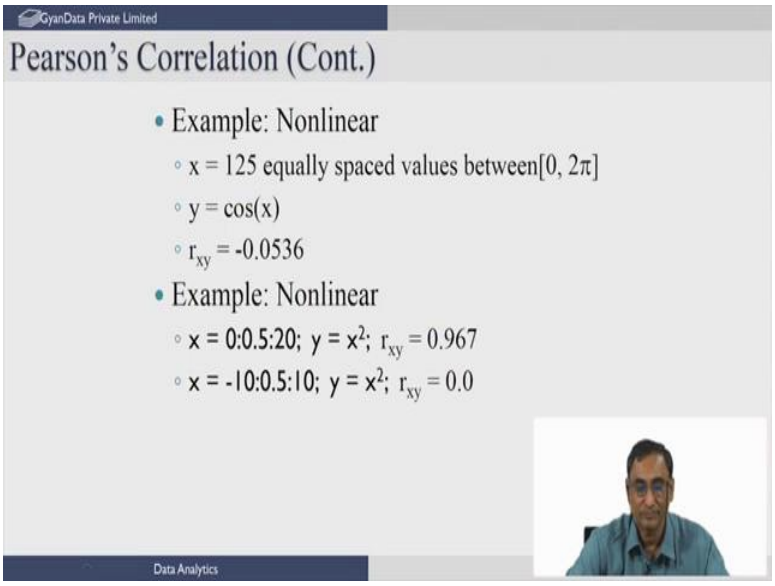
So, let us look at some examples, this is a very famous data set called the Anscombe’s data set. Here there are 11 data points, for each of this there are 4 data sets I have only shown 3 of them. Each of them contains exactly 11 data points corresponding to xi and y i, these points have been carefully selected. In the first one if you look at it, if you plot the scatter plot, you will see that there seems to be a 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 well, there is a perfect linear relationship for all the data points except one which seems to be an outlier, which is indicated we 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; it does not matter whether the you actually apply the first data set, or second data set, the third data set. In fact, the fourth data set there is 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 a high correlation coefficient close to 1 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, we 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 this point.

(Refer Slide Time: 10:14)



GyanData Private Limited

Pearson's Correlation (Cont.)

- Example: Nonlinear
  - x = 125 equally spaced values between[0, 2π]
  - 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 three examples, in the first example I have taken 125 equally spaced values between 0 and 2 π for x; and I have actually computed y as cos(x). So, there is a relationship between y and x, in this case it is a sinusoidal relationship. So, if we apply the Pearson's correlation coefficient, compute the Pearson correlation coefficient for this data set we 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 excess between 0 and π by2 and when it is between π by2 and π and between π and 3π by 2, 3 π by 2 and 2 π 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 compute a y equals x square and then computer Pearson's correlation for this xy data, you find it is a ve ry high correlation coefficient.

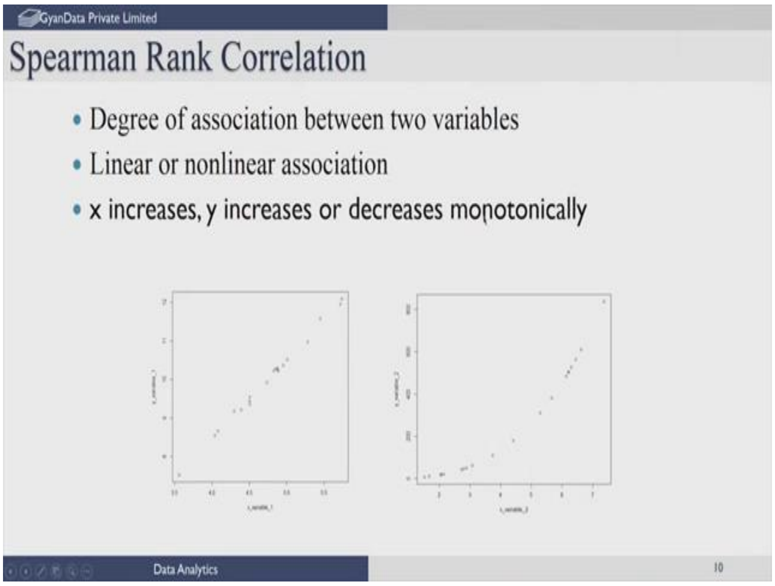
You can not immediately conclude there is a linear relationship between x and y ok, you can only say there is a relationship; perhaps it is linear, maybe it is even long linear we have to explore further. It is close to 0, I would have said there is no linear relationship ok; but it is in this case it 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 ok; for between -10 and 0, $y = x^2$will have positive values; between 0 and 10 y equals 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 value is high, looking at this 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 called not ranked variables, ordinal variables, real valued variables like we have here.

(Refer Slide Time: 13:15)



GyanData Private Limited

Spearman Rank Correlation

- Degree of association between two variables
- Linear or nonlinear association
- x increases, y increases or decreases monotonically

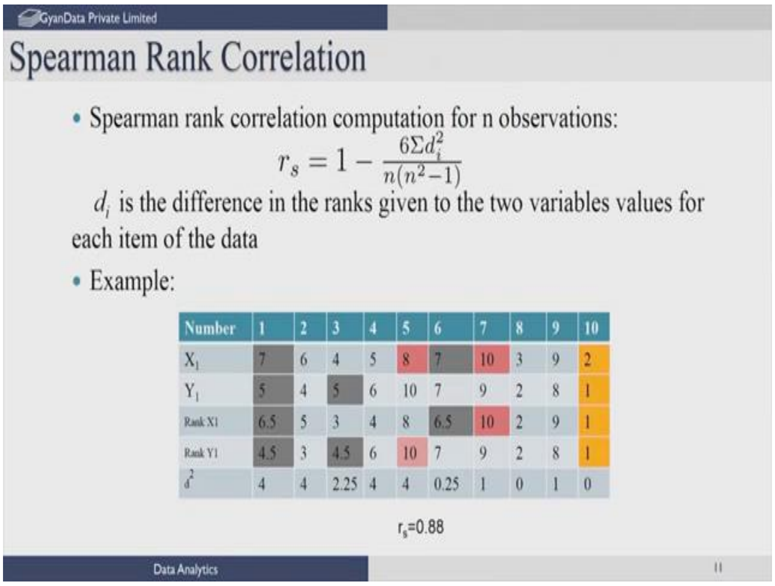
Data Analytics

10

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 again look for degree of association between two 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 is indicates a linear relationship. Let us apply define Spearman's rank correlation, apply it to the same data set and see what happens.

(Refer Slide Time: 13:53)



GyanData Private Limited

Spearman Rank Correlation

- Spearman rank correlation computation for n observations:

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

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_1	7	6	4	5	8	7	10	3	9	2
Y_1	5	4	5	6	10	7	9	2	8	1
Rank X1	6.5	5	3	4	8	6.5	10	2	9	1
Rank Y1	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 11

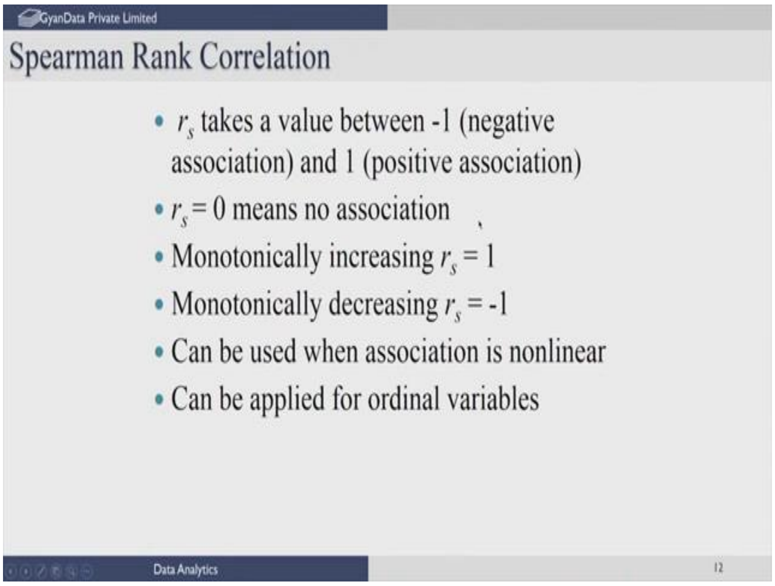
So, in the Spearman’s rank correlation what we do is convert the data, even if it is real value data to what we call ranks. So, for example, let us say i have 10 data points, in this case x_i is like somebody has taken a scale between let us say 2 and 10, 1 and 10 and similarly y is also a value between 1 and 10. So, what we have done is looked at all the individual values of x and assign the 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 have ranked all of these points; notice that the 6th and the 1st 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 than two values which are tied, we take all these ranks and average them by the number of data points which have equal values and correspondingly we obtain the rank.

We also rank the corresponding y values. For example in this case the 10th value as a rank 1 and so on so forth, 8th value as a rank 2 and so on; so we are given a rank in the 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 xi and yi which is 2 and square it and we will 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: 16:00)



GyanData Private Limited

Spearman Rank Correlation

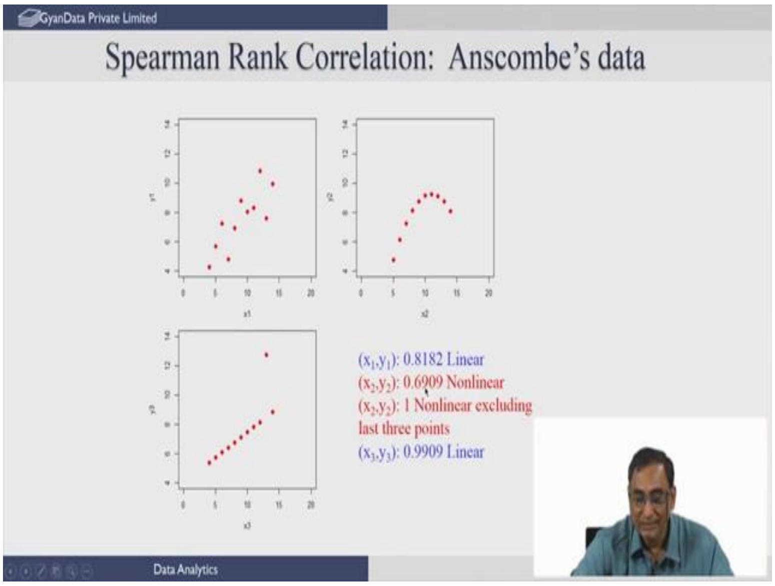
- r_s takes a value between -1 (negative association) and 1 (positive association)
- $r_s = 0$ means no association
- Monotonically increasing $r_s = 1$
- Monotonically decreasing $r_s = -1$
- Can be used when association is nonlinear
- Can be applied for ordinal variables

Data Analytics 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 value a 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 differences 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 ok, it is 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: 16:49)



GyanData Private Limited

Spearman Rank Correlation: Anscombe's data

(x_1,y_1): 0.8182 Linear

(x_2,y_2): 0.6909 Nonlinear

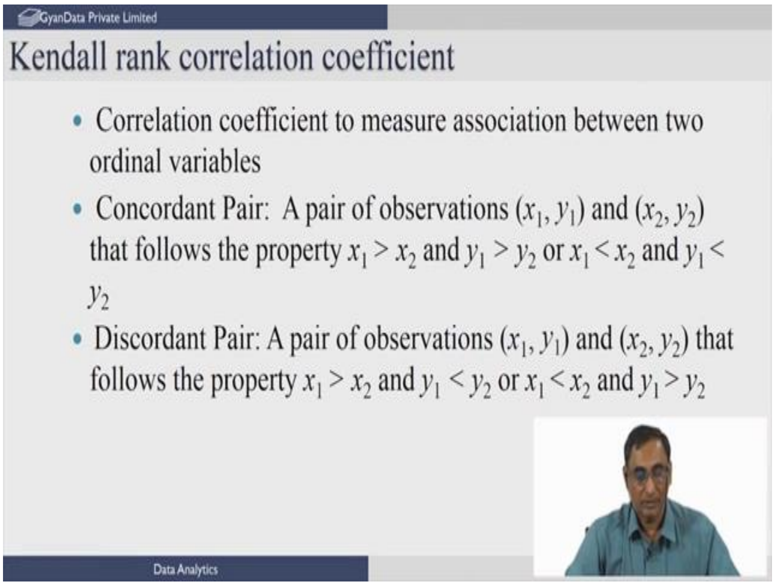
(x_2,y_2): 1 Nonlinear excluding last three points

(x_3,y_3): 0.9909 Linear

Data Analytics

So, let us apply it to the Anscombe's 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 was high; 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: 17:16)



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 $y_1 > y_2$

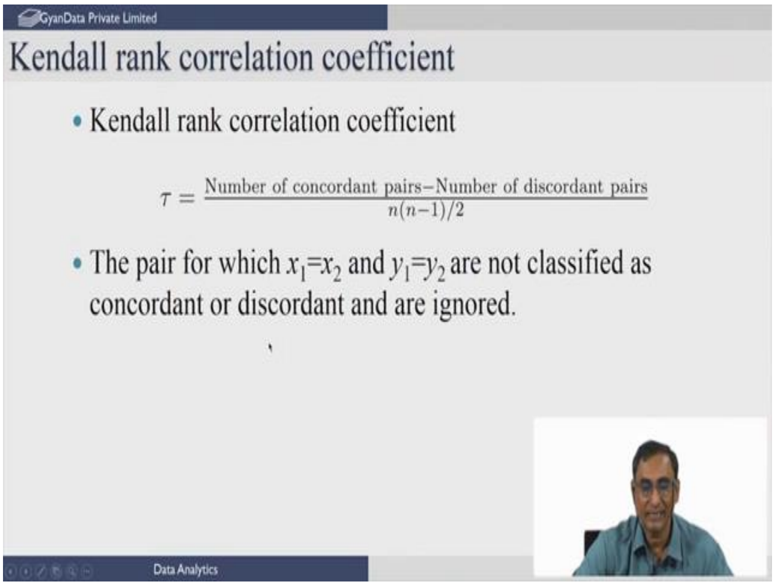
Data Analytics

Suppose we had applied I would suggest that you apply it to the cos x example and y squared x = y squared example; you will find that the Spearman rank correlation for these will be reasonably high. It may not be close to 1, but will be high indicating there is some kind of a non-linear relationship between them, even though Pearson's correlation might be low. So, a 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 variables.

In this case, what we define is a concordant and a discordant pair. If you look at the values compare two observations; let us say x_1 y_1 and sorry here, it should be 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 two data points are known said to be concordant. Similarly if x is increasing and x_1 less than x_2, and correspondingly y, 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 two data pairs are concordant.

On the other hand if there is an opposite kind of relationship. So, if you take two data points x_1 y_1 and x_2 y_2 and we say that you know look, 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.

(Refer Slide Time: 19:09)



GyanData Private Limited

Kendall rank correlation coefficient

- Kendall rank correlation coefficient

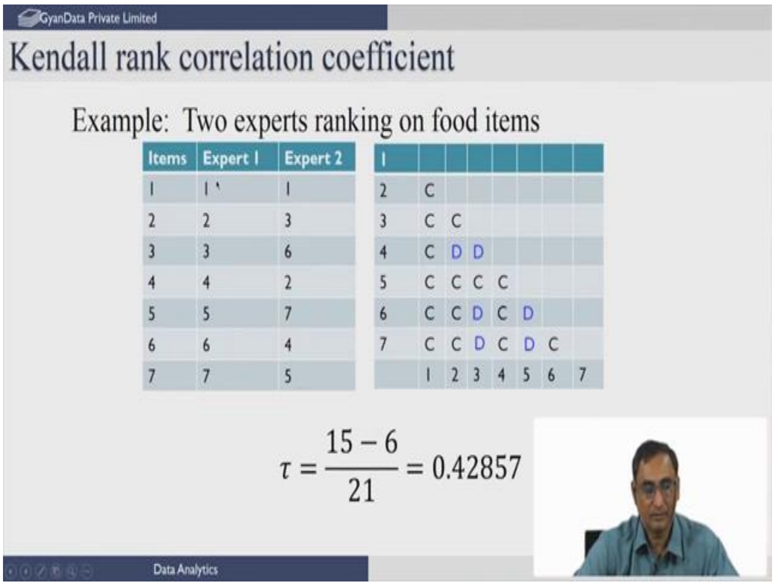
$$\tau = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{n(n-1)/2}$$

- 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

Let us take an example and look at it. So, once we have the number of concordant pairs and number of discordant pairs, we take the difference between then divide by the n into n -1 by 2 and that is called the Kendall stop.

(Refer Slide Time: 19:20)



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

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

Data Analytics

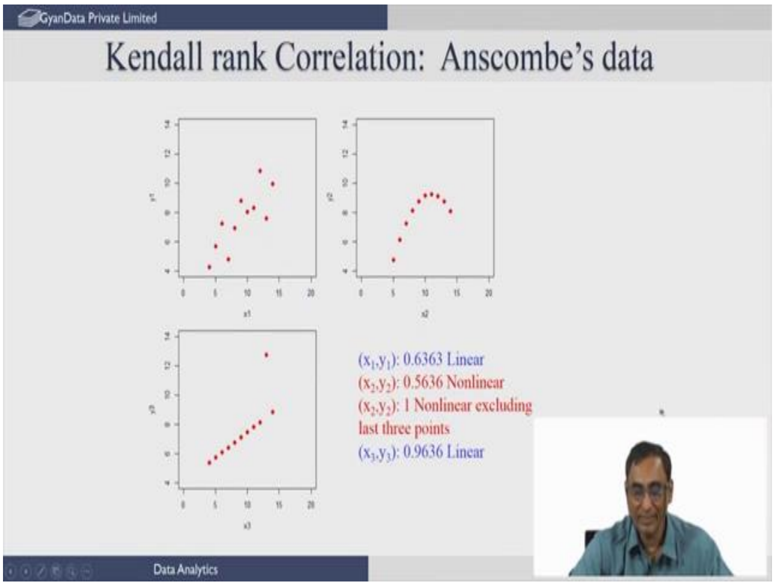
We can take a item here, there are about seven observations; let us say seven different wines or tea or coffee and there are two experts who rank the taste of the tea or coffee or wine on a scale between 1 to 10. For the first, the expert number 1 gives it a rank of 1 and expert 2 also ranks it 1; for the second 1 the expert 1 ranks it 2; while the expert 2 ranks it is in a scale or gives it the value of 3 and so on so forth for the seven different types of thing. Now you compare data point 1 and data point 2, in this case experts oπnion 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 are agree in agreement, 3 is better than 2, 3 is better than 2. So, it is a concordant. Let us look at the 4th and the 1st one, looks like expert 1 says it is better, expert 2 also says better, concordant.

But the 2nd and 4th if you compare, expert 1 says it is better 4th one is better than the 2nd; but expert 2 disagrees, he says the 4th thing is worse than the 2nd one. So, there is a discordant pair of data and 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 ok. 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 or 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 ok. So, the high negative value or high positive value indicates that the two variables x and y in this case are associated with each other. Again this can be used for ordinal variables, because it can worked with ranked values here, as we have seen in this.

(Refer Slide Time: 21:53)



Kendall rank Correlation: Anscombe's data

(x_1,y_1): 0.6363 Linear
(x_2,y_2): 0.5636 Nonlinear
(x_2,y_2): 1 Nonlinear excluding last three points
(x_3,y_3): 0.9636 Linear

So, again if we apply it to Kendall's rank to this Anscombe's dataset, we find that although it has decreased for this linear case; the value has decreased as compared to the Pearson and Spearman's correlation coefficient; it still has a reasonably high value. High in this case tyπcally you should, in experimental data you cannot expect to get a value I mean beyond 0.6 0.7, you should consider yourself fortunate tyπcally; 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 these to get a preliminary idea before you even build the model; of course, for two 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 bid.

And this is the first thing that we do before we jump into linear regression. So, see you next class about how to build the linear regression model.

Thank you.