Robert_S2_L17

Wed, 1/12 1:33PM 17:07

SUMMARY KEYWORDS

correlation, price, correlation coefficient, car, fuel efficiency, correlated, cleared, determination, correlation matrix, series, calculate, variance, miles, reported, trunk, gallon, scatterplot, positively, mpg, line

SPEAKERS

Robert McKeown

R Robert McKeown 00:05

Hello, welcome. In this video, we're going to continue our analysis of correlation or how to use the correlation coefficient. We're going to visualize correlate correlated series on a scatterplot diagram. And we're going to learn how to interpret the correlation coefficient to understand how much variability in one series is explained by movements, or the variability in another series. And that's called the correlation of determination. Here, we have auto thefts in the city of Toronto for the years 2014 to 2020. We've been given some information about the variance and the covariance, which you can see right here, here and here. That's going to make our work a little bit easier. And we're being asked to find the correlation between reported and cleared auto thefts. So reported crimes are crimes that are reported by the public, they say that a crime has happened. When a reported crime is cleared, that means the police have solved it. One way or the other, maybe they determined that no crime was committed. Maybe they found the perpetrator, and they press charges. Or maybe they found out who did and they don't want to press charges. To find the correlation between reported and cleared auto thefts, we are going to use our with subscript A little r for our reported and little c for clear. And we need the covariance between the two divided by the standard deviation of our multiply by the standard deviation of cleared cases. And we have that information right here. And so we have 21,826.4. That's our covariance, divided by 925.9 multiplied by 119.7. Using my calculator, I find that the correlation is 0.197. How strong a correlation is it between reported cases and cleared cases? Well, it would be a weak correlation if we were using the table than it presented in an earlier video. Now if we want to try and illustrate how two variables move together, one type of graph we could use is a scatter plot. And a scatter plot graph looks like the graph here on the slide in front of you. We have cleared cases on the y axis and reported cases on the Y axis. Notice that our y and x axis do not start at zero, they actually started different values. And each of the years is clearly labeled. So in 20, say 2020, there was a little less than 6000, reported thefts. And the cleared the number of cleared cases looks like it was maybe 425. Or I guess we can just look at this table over here is actually 418. And so we just graphed these points over here on our diagram. And this is what's called a scatterplot. Here is our correlation over here. What kind of correlation do we have? We said it was weak, what kind of weak correlation a weak positive correlation, because the correlation coefficient is greater than zero. Now, if we draw a line, we're going to draw a line on this graph here. That minimizes the

distance between each of the data points, each of these little points here, and that line will look like so a line going upward sloping, it's positively sloped, it's positively sloped, much like our correlation coefficient is positive. And that is kind of an illustration of our correlation by using one of these lines, these fitted lines that try to minimize the distance between Each of these points and the line itself, trying to minimize the distance from the line to each observation that's the graphical interpretation of a positive correlation we put in that line, that fitted line. It's upward sloping. And it tells us that these two series are positively correlated with each other. Here's a data series from the US federal government on different cars and different characteristics of cars that were made in that year. The table you see in front of you, illustrates just a few of the cars believe there was 48 different models that were part of this data set. And here on this table, just have a little bit of the data set, we have the price of the car, and we have miles per gallon. MPG is a measure of fuel efficiency, a gallon is a unit of measure for a fluid. And the more miles is distance traveled. So maybe in Europe, we call it kilometers, or I should say, Yeah, kilometers per liter. But in the United States, its miles per gallon. And so a higher miles per gallon means that the car is more fuel efficient. We also have a measure of the cars weight in pounds, since we're looking at the United States, which is a way of measuring weight. And we have the price in US dollars. Now, I'm not going to show you all the observations because there would be too many to fit on one screen. But when I create a scatterplot of price and miles per gallon, we get the following image illustration, you can see that we have the price in US dollars on the Y axis. And we have fuel efficiency on the x axis. And a whole bunch of different points for different models of cars. Some are very fuel efficient. And some are the opposite of fuel efficient. Some are some are pricey, pricey like these, this these ones up here. They would be high priced, and some are low priced like the cars down here. Now if we want to illustrate a correlation between prices and fuel efficiency, we could do that by drawing a line like we did before a line that will minimize the distance between the line itself and all the different observations. When we do so, we see a line like this green line on the screen in front of you and it is downward sloping. And it is fairly steep. This suggests to us that the correlation between fuel efficiency and price is a negative correlation. Now when you're working on series, you're doing a research paper, you're trying to produce a research report for your boss, you're going to be working with more than two series. And a correlation is between two series. If we have more than two series, then what we typically do is create something called a correlation matrix. And the correlation matrix shows you the correlation between different series. And if we look over here, you can see it's a nice way to present correlations between more than two series. So if we look at a car's weight, right here, we can see that it is positively correlated with the price of the car. And it's substantially, I think we would call it at least medium medium correlated With the price of the car, it's very much negatively correlated with the fuel efficiency of the car. So a heavier car is going to be, according to this correlation matrix is likely to be heavier car is likely to be less fuel efficient. On the other hand, it's also likely to have a bigger trunk. And its correlation with itself is equal to one, of course, the correlation of two of a series with itself is always equal to one. And that's why we have these ones along the diagonal, which you can see here, here and here. Now what about price, we can see that price is correlated perfectly with itself. Price is negatively correlated with fuel efficiency. So the more fuel efficient a car is, the lower the price is going to be, there's a negative correlation between the price of a car and its efficiency. Or I should say its fuel efficiency. The trunk size, and the price of the car also positively correlated, weakly but positively correlated nonetheless. And notice I'm reading down rows, I'm in that first column that's represents price. And I'm moving down the rows to look at how price is correlated with different features of the car, and the price and the way we discussed previously. Of course, we could also go to the middle. So if we're interested in the trunk, and mpg, we're going to jump into the middle of the correlation matrix. And we see that the bigger the trunk size, the less fuel efficient the car is, the fewer miles, you're going to get. For a gallon of gasoline. We can use the

correlation coefficient to calculate the correlation of determination. Correlation determination tells us something a little bit in more detail than just the correlation, the correlation could we could identify whether it was a strong, very strong, weak or very weak based on the numeric value we are given. But the correlation of determination tells us something very specific, it tells us that the percentage of variance in one variable or one series that is explained by the variance and another series. So the correlation of determination is 50%. That means 50% of the changes, and series A can be explained by changes in series B. To find to calculate the correlation determination, we square the correlation coefficient. And it has the notation R squared. So here is our correlation matrix that we were analyzing and discussing on the previous slide. And here is the correlation of determination. And so we can see that if we're looking at the price of a car. And we want to understand the relationship between the weight of the car and the price, we can say according to the correlation of determination that the weight of the car explains 29% of the variance in the price. So the variance in the weight of the car explains 29% of the variance of the price. Now the correlation of determination is just the r squared. And so if we've got the trunk size of point three one squared, that's just going to be equal 2.31 times point three one, which is approximately equal to 0.1 which we see right here. And notice that the negatives are disappearing. So we've got a high correlation between the weight of the car and fuel efficiency. And if we calculate this one, we can see that the negative is multiplied by itself. So it's disappearing. But of course, that still doesn't change the fact that the weight of the car and miles per gallon are negatively correlated. So this correlation of determination is very useful, because it tells us something about how changes in one variable might affect another variable. Given that they're correlated. We have to be a little bit careful about doing too much analysis on this. Because a correlation is not causation. So we haven't shown causation. Although it's probably likely that the weight of the car is causing the miles per gallon, to be lower. That's kind of common sense and understanding of the natural world, that's likely the farmans going on here. But when we're given a correlation coefficient, remember, we can't say anything about causation just based on the correlation coefficient. But this correlation determination is very useful. And it's a useful application of that correlation coefficient. Before you conclude your studies in this section, it's very important that you have a deep understanding, and you really memorize how to calculate the standard deviation and the variance, as well as the correlation and the correlation of determination. These are going to come up again and again and again and in future statistical classes that you take and in this course, so you really want to make sure that you master them, you don't have any doubt on how to interpret them. And you know how to calculate them and recognize them when they appear in another formula. That's really going to help you be successful in your university career and as a person who has an understanding of quantitative methods