Robert_S3_L07

Thu, 2/17 4:01PM 13:19

SUMMARY KEYWORDS

scatterplot, correlation, outliers, relationship, variable, correlation coefficient, nonlinear, weaker, fact, draw, quantitative variables, linear, equal, observations, calculate, line, data, variability, cloudy, satisfied

SPEAKERS

Robert McKeown

R Robert McKeown 00:04

Let's continue our analysis, we're looking at scatterplot diagrams and trying to determine if the conditions for an accurate correlation coefficient are satisfied, namely that the data is quantitative, the variables or quantitative variables, the relationship is linear, and there are no outliers. Let's take a look at scatterplot. See, we can see looking at scatterplot see that as the X variable increases, the Y axis variable is also increasing. So we have a positive, sometimes known as a direct relationship. Looking at the data, it seems that it is quantitative variables, we've got numbers on both the axes. And the data is spread out in such a fashion that we can get a sense of the relationship itself. It doesn't take on some strange characteristics that it might have worked categorical data. And I like to draw a line. And if I try to draw a line here, hope you can kind of see that it's not that easy. I'm having a difficult time drawing a line where this area down here, excuse me, this area down here, looks like that area up there and this area down here. And in fact, if I get rid of my scribbles here, it kind of looks like this relationship isn't so much a straight line as a curve. And in fact, it looks almost like an exponential curve, it could be something like Y is equal to, AEX, something like that. And so this relationship does not look linear. It doesn't look linear. And that means our correlation coefficient might not be, well, it's not going to be as accurate. It could overestimate or underestimate the true relationship between the X and Y variable. And we'll summarize that a little bit later in this video. So I would guess that this is some kind of exponential function, possibly of the form that I'm suggesting there, although it could be something different. An alternative, maybe it's just Y is equal to A times X squared, or something like that, or X, it could even be X cubed. Now let's take a look at scatterplot D. So I guess maybe we can say before we move on is not just to be clear, that condition for linearity is not satisfied in scatterplot. See, now looking at scatterplot D, as the x axis variable increases, our Y axis variable decreases. So it's a negative relationship. But again, if I try to draw a line through the middle of this thing, it's hard to do so in such a way, that all at all points on the line, the relationship between the line and the actual observations themselves, is the same. And if I were to remove this, these scribbles that I've made, I would suggest that the relationship here is actually a nonlinear one, possibly have the form y is equal to one over X, or maybe some coefficient A over X, something like that. Where as X is getting bigger, Y is actually getting smaller. And so this is not linear. So the condition of an area is not satisfied in scatterplot D, just as it was not satisfied and scatter plots. Let's take

a look at scatter plot E. Looking at scatterplot E, it's actually very similar to the scatterplot we just saw where it doesn't seem to be linear, so it's not linear either. It also has another condition that's not satisfied, which is these outliers here. It seems to have outliers Remember, an outlier is an observation that's very far away from the mean, or that I said was unusual. And so this observation rate here, this one right here is unusual, because the X value is very low, but the X is high. So we have a low X value. We have a high y value. And that seems to suggest that it's an outlier. So are these observations over here, they're potential outliers. In a situation where you're trying to write a school paper, or you're trying to understand the data better, you might want to look at these outliers. And ask yourself, what is it about them that makes them so much so different? They could just be, there could actually be a good reason why they're outliers. Something about those observations and makes them unlike these other ones down here. So we've got nonlinear, and it contains outliers. So those are two reasons why the correlation coefficient for scatterplot II is likely to or could be inaccurate. Looking at scatterplot F over here, simply saying we have that nonlinear sort of decaying function here, it could be something like Y is equal to A over X. And we also have outliers. So it's not linear. It's a negative relationship. But it's a nonlinear one, nonlinear one, it's not linear. And it contains outliers as well. So here are some situations where the correlation conditions are not satisfied, and you want to be skeptical of the correlation coefficient that you calculated. Now let's make sure that we can tell the difference between a strong correlation and a weak correlation just by looking at a scatterplot diagram. Here we have a scatterplot G. And you can see that it is a positive relationship. It's linear, and there don't seem to be any outliers. And looking at the axes, it seems that it is using quantitative variables, as well as the distribution of dots on the scatterplot. Now notice that the dots are very close to the line, so the dots are close to the line. And if the dots are closer to the line, that suggests that the correlation is stronger. In fact, if we look at the correlation of scatterplot, ci, we can see that the correlation is 0.98. It's almost perfectly correlated, right? A perfect correlation would be 1.0. And if we wanted to understand how much variability in one series is explained by the other series, we could look at the correlation of determination, the R squared, and the R squared tells us that 96% of the variation in one variable is explained by the variation in the other. So very strong correlation here that we see in scatterplot G. Now let's look at a weaker correlation. We've got scatterplot h, this is going to be a weaker correlation. I could try to draw draw a line through this cloud. And notice that I'm using the word cloud it looks like Cloud or maybe, maybe more cloudy. I think I think there is a There is indeed a positive correlation the way I've drawn it, we're gonna find out in a moment when we look at the correlation. But it seems to be a weaker, a weaker relationship. If I'm trying to predict y values based on the X value, you know, I might end up with a Y down here. But I could also end up with a Y, up here. So, I hope you can see that it's a weaker relationship. If we look at the correlation coefficient, it's not zero, but it is a weaker and falls within the, what we defined as a weak correlation, not a zero correlation by any means, but a weak correlation. In fact, if we're looking at the variability in one of the variables, only 7% of the variability in one variable can be explained by the other. That's our correlation of determination right there. Here, you see that we can use the scatterplot to actually determine the strength of the correlation without even really calculating any numbers. Looking at scatterplot I, our last scatterplot, we can see that it looks even more cloudy than our previous scatterplot. Ah, I'm kind of afraid to even try and draw a line through this. In fact, I won't even try to draw a line through this. I guess, well, you know, maybe it goes like that, but could also convince me that it goes like that, and it's hard for me to argue against you. So let's take a look, what is the correlation of this very cloudy, you know, looks like a cloud. Cloudy picture. And the correlation is 0.04. So it's almost a zero correlation. Typically, when you have two variables, you're very rarely going to actually calculate a zero correlation. Just because of machines and calculations, the fact that this series have a variance remember, if the series have more variants, all else being equal, we're going to get a higher correlation coefficient. But

we see here that the correlation coefficient is very small, it's quite close to zero, it would be a very weak relationship, maybe even zero. And if we tried to calculate the correlation of determination, you know, 0.4 to one times itself, is a very small number. So it's going to be less than 1% of the variability in one series explained by the other. So these are essentially, for all our purposes. This looks like the two series the series on the X axis and the Y axis are uncorrelated with each other.