This tutorial is going to teach you about the coefficient of determination. This is also called R squared, the square of the correlation coefficient.

So the correlation coefficient-- let's make the difference here-- gives a general measure of strength and direction of a linear relationship. So that's a general measurement, whereas R squared gives you a very specific measurement. It tells you the percent of the variation in the Y direction that can be explained by the linear relationship with the X variable. Now, that can be a little confusing to understand. So let's work through an example together.

So the graph here is a dot plot. It's on the y-axis, but it's a dot plot of the seafood prices in 1980. This is going to be our Y variable, but it's not very well contextualized. You might just sit there wondering, well, why is sea scallop so expensive? What would cause that to be so high? Versus why are these prices so low? What we can do is-- when we add a variable we can try and understand why the 1980 price of sea scallops were so high, versus some of the other prices were so low. When we look at it with the 1970 prices we can explain why some of these are high or low or in the middle. Well, the low prices were low in 1970, and the high prices were high in 1970. Looking at this divorced of its previous context doesn't really help to explain why certain prices are high or low. Looking at it with the full context of previous knowledge and its associations helps to explain why this point is high up, and why these points are low. It's high up because it's strongly linearly associated.

The value of R squared in this particular example is 0.935. Now, because it's a 0.935, that means 93.5% of the variation in 1980 prices can be explained by a linear association-- or a linear relationship with 1970 prices. You might be wondering what happened to the other 6.5% of the variation. How is that explained? The reason that it's not 100% of the variation is because these points don't all lie perfectly on a line. If they did, we would be able to explain-- all the reasoning behind the 1980 price would be explained by the 1970 price. But they don't lie exactly on a line.

There are some points that fall conspicuously a little bit below what you would imagine the line to look like. So the remaining 6.5% of variation has to be explained by something else. Maybe some species of fish were overfished, and that raised prices. Or people's tastes changed and the demand for a particular fish fell, and that lowered the price.

Ultimately, R squared is always a positive number, and it does help to measure the strength of the linear association. But it does measure something very, very specific. And it doesn't indicate the direction. It only can indicate the strength. So, for instance, down here. Both of these two scatterplots have the same R squared of

0.81, although clearly this one has a positive association and this one has a negative association. If only R squared is given, what you have to do is you have to take the square root in order to obtain the correlation. But you also have to look at the association, positive or negative, to determine the sign. So this has a correlation coefficient of 0.9. This has a correlation coefficient of negative 0.9.

And so, to recap, the coefficient of determination allows us to understand the percent of variation in the vertical direction that can be explained by the linear association that the two variables have. If you solve for R, from R squared, you need to not only take the square root but also look at the scatterplot to determine the sign. Because R squared can't be negative, but R, in fact, can.

So we talked about the coefficient of determination, also called R squared. Good luck, and we'll see you next time.