Print

Interpreting Correlation

Let's examine some important considerations when interpreting correlation coefficients. First, we discuss correlation versus causation. Note that a strong correlation between variables does not necessarily mean that a change in the values of one variable is the cause of the change in the values of the other variable. A strong correlation can happen for a number of reasons. It might be the result of random chance, where the variables appear to be related, but there is no true underlying relationship. There might actually be a cause-and-effect relationship. Or, there might be a third, lurking variable that is not studied but makes the relationship appear stronger (or weaker) than it actually is. A common lurking variable is time. Two unrelated variables might be increasing or decreasing together over time, or one might be increasing while the other is decreasing.

For example, take the relationship between engineering doctorates awarded in the US and cheese consumption in the US. This was created using the Spurious Correlations website, which has many interesting examples. The correlation between these two variables is strongly positive. But it is highly unlikely that cheese consumption is directly related to engineering doctorates, or that eating more cheese leads to more interest in pursuing an advanced degree in engineering. There is likely some other factor, a lurking variable, linking these two variables.

Lurking variables can be difficult to identify. Plotting the data on a line chart can sometimes reveal seasonal trends or similar patterns of change over time. In this example, Year is a possible lurking variable.

Another thing to keep in mind when using correlation is the assumption that the relationship between the variables is linear.

A correlation coefficient of zero or near zero does not necessarily mean that there is no relationship between the variables. It simply means that there is no linear relationship. In both of these examples, there is clearly a relationship between the pairs of variables, but the correlation coefficients are small. There is a curvilinear relationship in the first graph, and a cyclic relationship in the second graph. This illustrates the importance of plotting the data. For both of these examples, correlation is not a useful measure, although there is clearly a relationship.

Another issue in interpreting correlation relates to the influence of unusual observations or outliers. Let's look at an example with one apparent outlier. The correlation coefficient indicates that there is a relatively strong positive relationship between X and Y.

But, when the outlier is removed, the correlation is near zero.

We explore the influence of outliers and unusual observations using a simulation in the next video.

Close