2 CORRELATION AND REGRESSION

2.06 Correlation is not causation

When we do a regression analysis, we assume that the independent variable X explains the dependent variable Y. Building on that assumption, we can make a scatterplot and let the computer draw the line that best describes the linear relationship between the two variables.
With this line and the corresponding regression equation we can predict the values of the dependent variable based on the values of the independent variable. Moreover, with r-squared we can also assess how well the line fits our data.

However, for at least two reasons, we need to be very careful when we interpret the results. The first reason is that on the basis of a regression analysis, we can never prove that there is a causal relationship between two variables. We can, in other words, never be certain that one variable is the cause of another variable. This translates to one single and not very complicated, but extremely important message: **correlation is no causation**.

For instance, research suggests that eating a lot of chocolate makes you fat. *This* scatterplot shows that the more chocolate people eat, the larger their body weight tends to be. However, we need to be careful here. It might also be the case that causality runs in the opposite direction. The correlation between the two variables could also have another reason. It might for instance be the case that people with more body weight are more hungry and therefore eat more chocolate. This means that your X variable becomes your Y variable, and your Y variable becomes your X variable. This changes your scatterplot and your regression equation. *This* is the old scatterplot and *this* is the new one. *This* is the old equation and *this* is the new one.

The most likely explanation of the relation between chocolate consumption and body weight, however, is that causality runs in both ways. The more chocolate you eat, the heavier you get, and the heavier you get, the more you crave chocolate.

However, something else might be going on. A study has shown that the inclination to be hungry is genetic. This means that some people are because of biological reasons more likely to eat more chocolate and also more likely to eat other things and thus to gain body weight. There is thus another variable, variable Z, that explains both chocolate consumption and body weight. We say that chocolate consumption and body weight are **spuriously related** to each other. In this example we would call genetics a **confounding** variable if it we had included it in our study. If it was not included, but could have the *potential* for confounding, we would call it a **lurking variable**.

One more example. *This* scatterplot shows that in countries with more income inequality, people tend to be, on average, more dissatisfied with politics. The assumption here would be that income inequality is the cause and dissatisfaction is the consequence, because people will blame the political establishment when they perceive large differences between the rich and the poor. However, something else might be going on. It might well be the case that the degree of corruption in a country predicts *both* income inequality *and* political satisfaction. The idea then would be that when there is a lot of corruption, the rich will try to further enrich themselves at the expense of the poor, leading to more inequality. At the same time, corruption will make people more cynical about the political process, as a result of which they become more dissatisfied. It is not X leading to Y, but a secret variable, Z, which explains the correlation between X and Y.

A second reason why we should be very careful when interpreting regression is that influential **outliers** can have strong effects on the results of an analysis. *This* scatterplot shows that there is a

very strong positive correlation between chocolate consumption and body weight. But look what happens when we add another case with rather extreme values on both the dependent and the independent variable. *This* person eats 500 grams of chocolate per week and weighs only 40 kilograms. Adding this case to our analysis strongly changes the result. The slope of the regression line now is not positive anymore, but negative, indicating that the more chocolate you eat, the less you weigh… Especially when you only have few cases in your analysis, outliers can have a large impact on your findings. If you see outliers you should always investigate what's going on. If you have good reasons to suspect that the outlier is the result of wrong measurement, you might want to decide to delete this case from your sample.

So always be very careful when you conduct a regression analysis. First, you always need to remember that with a regression analysis you never *prove* that there is a causal relationship between two variables. Second, you should always check for influential outliers, especially when you are working with a small sample. These outliers could have a large impact on the results of your study.