2 CORRELATION AND REGRESSION

2.02 Pearson's r

This scatterplot shows the relationship between chocolate consumption and body weight. On the horizontal x-axis we see the independent variable chocolate consumption, measured in consumed grams of chocolate per week. On the vertical y-axis we have the dependent variable body weight, measured in kilograms. The cases studied here are 200 female students who are exactly 1m70 tall. They are represented by the 200 dots in this graph. The scatterplot shows at a glance that there is a strong correlation between the two variables: the more chocolate someone eats, the larger her body weight. But *how* strong is this correlation? We will now turn to one of the most often used measures of correlation: **Pearson's r**.

One of the most important advantages of Pearson's r is that it expresses the direction and strength of the **linear correlation** between two variables with a single number. The relation between chocolate consumption and body weight can best be described by *this* straight line. Because all cases cluster closely around the line, we can conclude that this is a rather strong correlation. Another thing to note is that the line goes up: more chocolate consumption is associated with higher body weight. We can therefore also say that there is a positive correlation. Conclusion: we have a strong, positive and linear relationship here.

However, variables could also be correlated in different ways. In *this* graph we see a rather strong positive and linear relation between the variables x and y, just like in the example with chocolate consumption and body weight. But in this graph we have a rather strong *negative* linear correlation. The line goes down, which indicates that when variable x goes up, variable y goes down. In *this* graph we also see a positive linear relationship. However, it is much less strong than the previous ones. We know that because the individual cases are much further removed from the line.
This is a *perfect* negative linear correlation. It is perfect because all cases lie exactly on the line. But the correlation between two variables need not be linear. In *this* graph we also see a relationship between the variables x and y. However, the line that best represents the relationship between the two variables is not straight. Instead, it is a U-shaped line. We call this a **curvilinear relationship**.

A scatterplot helps us to broadly assess whether a correlation is strong or weak, but it does not tell us exactly how strong the relationship is. Pearson's r is a measure that can show us exactly that. More specifically, the Pearson's r tells us the direction and exact strength of the *linear* relationship between two *quantitative* variables. A positive Pearson's r indicates that a correlation is positive, and a negative correlation indicates that it is negative. The size of r expresses how tightly the observations are clustered around the imaginary best-fitting straight line through the cloud of data points. Pearson's r is always a number between minus 1 and 1. Minus 1 refers to a perfect negative correlation. Plus 1 to a perfect positive correlation. And 0 means that there is no correlation at all.

So , how do we compute the Pearson's r? Imagine that the study on chocolate consumption and body weight was not based on 200 but on only 4 individuals. *This* is the data matrix. And *this* is the scatterplot. You can see that every combination of values on the two variables becomes a circle in the graph. This woman consumes 200 grams of chocolate per week and weighs 70 kilograms. She is represented by *this* circle. The other three circles represent the other individuals you see in the data matrix.

To compute the Pearson's r we need *this* formula. What does it mean? Well, first, we change all original scores to z-scores. In other words, we standardize the values. The reason is that we want the Pearson's r to be a number between minus 1 and 1. If we don't standardize, the measure of

correlation will be expressed according to the original metrics. So, first we compute the mean for both variables. This results in the value 162.5 for variable X (chocolate consumption) and 71.25 for variable Y (body weight). Then, we compute the standard deviations for both variables. The results are 110.9 for X and 18.4 for Y. We get the z-scores by applying this formula to every case: we subtract the mean from every value and then divide it by the standard deviation. So: 50 minus 162.5 divided by 110.9 equals -1.01. We do that for every value of the independent variable (chocolate consumption). And for every value of the dependent variable (body weight). In the next step we compute the products of every z-score on X with every z-score on Y. So, for the first case that is -1.01 times -1.15 equals 1.17. *These* are the results for the other three cases. To get *this* part of the formula, we add up all these scores. That's 2.78. To finish we have to divide it by n minus 1. n equals 4, so in our case n minus 1 means 4 minus 1. That's 3. The Pearson's r is 2.78 divided by 3. That equals 0.93.

What does this mean? It means that there is a strong positive linear relationship between chocolate consumption and body weight. One important note. You can always compute the Pearson's r, even if the relationship is not linear. Therefore, it is very important that, before you compute a Pearson's r, you should always check the scatterplot first to see if your variables are linearly related. If not, do not compute the Pearson's r, because it doesn't tell you much about the relationship between your variables. *This* scatterplot, for instance, shows that there is a strong curvilinear relationship between X and Y. If you compute the Pearson's r, you will get a very low value: -0.15. This does not tell you that there is a weak correlation. It only tells you that there is a weak *linear* correlation.

It is rather easy to compute the Pearson's r with only 4 cases. However, you can probably imagine that it becomes an almost impossible task when you have, say, 200 cases. Luckily, every statistical computer program can compute the Pearson's r in no time. Nevertheless, it is important for you to understand what the Pearson r exactly tells you and what the formula means. It helps you to better understand how variables are related, and it might help you to decide for yourself how much chocolate to eat every week.