2 CORRELATION AND REGRESSION

2.03 Regression – Finding the line

Although many people like eating chocolate, most people are slightly cautious with their chocolate consumption, because they know that there is a strong correlation between the amount of chocolate you eat and your body weight. However, a recent study shows that it might actually be a good idea to eat a lot of chocolate. *This* scatterplot shows that a country's annual chocolate consumption per person (so, how much chocolate someone eats in a year) is positively related to the number of Nobel Prize winners per 10 million people in a country. Notice that in this scatterplot chocolate consumption is displayed as the independent variable and the number of Nobel prize winners as the dependent variable. The units of analysis in this scatterplot are countries. You can see that the correlation is pretty high. In fact, the Pearson's r correlation coefficient here is 0.93. This suggests that although eating a lot of chocolate might make you fat, it also makes you smart!

The Pearson's r tells you how strong the linear correlation between two continuous variables is. This linear correlation can be displayed by a straight line. In our case that's *this* line. This is what we call the **regression line** and in this video I 'll tell you how we can find the regression line. It is important to know how we find this line. Not only because the regression line shows you at a glance how two variables are related, but also because it forms the basis of many statistical analyses. So, how do we find the regression line?

Imagine that you draw every possible straight line through this scatterplot. So, *this* one, *this* one, *this* one, *this* one, and every other possible line. That's a huge number of lines, so in practice it will be almost impossible to do that. However, for now, imagine that you have superhuman powers and that you are able to do it. Next, you measure for every possible line the distances from the line to every case (so, in this case to every flag in the scatterplot).

Let me give you an example based on a random line. Say… this one. You measure the vertical distance between Japan and the line, the distance between Spain and the line, and so on, until you know the distance to the line of every case in your study. Every distance is called a **residual**. You end up with positive residuals (the distances from cases above the line to the line, displayed in blue) and negative residuals (distances from cases below the line to the line, displayed in red). You measure these residuals for every possible line through the scatterplot. So, not only for *this* line, but also for *this* line, *this* line and *this* line. And for every other possible line through the scatterplot. Eventually, you choose the line for which the **sum of the squared residuals** is the smallest. That's *this* one. Why the *squared* residuals? Because positive and negative residuals cancel each other out: the sum of the length of the positive residuals (the blue lines) is exactly as big as the sum of the length of the negative residuals (the red lines). The best fitting line is called the regression line, and the name of the method of analysis is called **ordinary least squares regression**, which refers to the way we have found the line.

In practice it is almost impossible to draw every possible line and to compute for every single possible line all the residuals. Luckily, mathematicians have found a trick to find the regression line. I won't explain how this trick works here, because it is rather complicated. For now it suffices to know that it is based on minimizing the sum of the squared residuals.

To summarize: you have learned two essential things. First, you now know *how* the computer finds a regression line. And second, you have learned that eating chocolate might well help you to pass this course!