2 CORRELATION AND REGRESSION

2.04 Regression – Describing the line

Does eating chocolate make you smart? A recent study indicates that this might be the case. In countries where citizens eat more chocolate, there are more Nobel prize winners per 10 million people. *This* scatterplot shows that a country's annual per capita chocolate consumption positively correlates with the number of Nobel prize winners per 10 million people in a country. The regression line is the straight line that describes the linear relationship between the two variables best.
But how can we describe what this line looks like? This is an important question, because by describing the line with a formula, we can easily communicate our regression analysis to other people, predict the number of Nobel Prize winners in other countries, and identify countries that do not fit the pattern.

Based on the regression line in this scatterplot we would predict that a country with an annual chocolate consumption of 6 kilograms per year per capita would have about 11 Nobel prize winners per 10 million people. Similarly, based on this same line we would predict that a state with an annual chocolate consumption of 11 kilograms per year per capita would have about 25 prize winners per 10 million people. For most countries this prediction will not be completely correct. After all, most countries are not exactly on the line. However, it is the best prediction that we can make based on the information that we have.

There is one simple formula with which we can describe the regression line, and that's this one: $\hat{y} = a + bx$. $\hat{y}$ is not the actual value of y, but it represents the *predicted* value of y. For example, when x equals 12, $\hat{y}$ equals 28. Notice that the actual value of y in this case is 33. However, the predicted value of y is the value of y on the regression line. This means that all the values exactly on the regression line are $\hat{y}$'s. 'a' is what we call the **intercept or the constant**. It is the predicted value of y when x equals 0. It is, in other words, the predicted value of y where the regression line crosses the y-axis and x thus equals 0. In our case that's -5.63. Notice that this value has no substantive meaning. It is impossible to have -5.63 Nobel prize winners per 10 million people. It only has a mathematical purpose, to describe the regression line.

'b' is what we call the **regression coefficient or the slope**. It is the change in $\hat{y}$ when x increases with one unit. In our case we see that when x increases with one unit (for example from 4 to 5), the predicted value of y increases with 2.80 units. Because we have a straight line, the slope of the regression line is the same everywhere. So also if we look at what happens when x increases from 8 to 9, $\hat{y}$ increases with 2.80 units. The regression coefficient in our example is 2.80. This leads to the following regression equation: $\hat{y}$ equals -5.63 plus 2.80 times X.

Take a look at *these* two regression lines. They have the same regression coefficients (or 'b' values). When x increases with 1 unit, the predicted y-values of line one and two increase with the same amount. These lines have different intercepts (or 'a' values), however. After all, they cross the y-axis on different positions. *These* two regression lines have different regression coefficients. When x increases with 1 unit, $\hat{y}$ of line one increases more than $\hat{y}$ of line two. Yet the intercepts of these two lines are the same because they cross the y-axis at the same spot. I have already shown you that we can use the regression line to predict y-values based on given x-values. We can also use the regression formula to make predictions.

Let's take our regression formula: $\hat{y}$ equals -5.63 plus 2.80 times X. We can predict our y-value by using the formula. What if x equals 3.5? We get -5.63 plus 2.80 multiplied with 3.5. That makes 4.17. So, $\hat{y}$ equals 4.17. And what if x equals 10.21? Then you get -5.63 plus 2.80 times 10.21. That makes

$\hat{y}$ equals 22.96. We get the same values when we just look at our regression line. For an x-value of 3.5 we get a predicted y-value of about 4. And for an x-value of 10.21 we get an $\hat{y}$ value of about 23. You can already see that there is one huge advantage of working with the formula: you can make much more precise predictions.

Usually the computer finds the regression line for you, so you don't have to compute it yourself. However, when you know the means and standard deviations of your variables and the corresponding Pearson's r correlation coefficient, you can compute the regression equation by means of two formulas. *This* one and *this* one. The first formula computes the regression coefficient by multiplying the Pearson's r with the standard deviation of Y divided by the standard deviation of X. This shows that the regression coefficient is in fact an unstandardized version of the Pearson's r. When the Pearson's r equals 0, the regression coefficient equals 0. When the Pearson's r is a positive number, so is the regression coefficient. And when the Pearson's r is negative, the regression coefficient is negative as well. These are the means, standard deviations and Pearson's r of our study. So, to find the regression coefficient we multiply 0.93 with 11.87 divided by 3.95. The outcome is 2.79.

The second formula computes the intercept by multiplying the computed regression coefficient with the mean of X and then by subtracting the outcome from the mean of Y. So: 13.17 minus 2.79 multiplied with 6.71. That makes -5.55. The regression equation is -5.55 plus 2.79 times X. The difference with the regression equation found by the computer (which was *this* one) is due to rounding error. I worked with the rounded means, standard deviations and Pearson's r. This has led to a less precise regression equation. So when working with these formulas, try to round off as little as possible.

Congratulations! You are now able to do a regression analysis and calculate predicted values! Understanding the basics of regression is of essential importance to be able to understand inferential regression procedures later on. So watch this video a couple of times. If independent variable X is the number of times you watch this video and dependent variable Y is your knowledge of regression analysis, when you do a regression analysis the regression slope will be a positive number. If you don't understand what that means, re-watch this video immediately!