2 CORRELATION AND REGRESSION

2.08 Example Pearson's r and regression

Social scientists have shown that a leader's physical height is related to his or her success. Suppose you want to test if you can replicate this result. To do that, you look at the heights and the average approval ratings of the four most recent presidents of the United States. You employ *this* data matrix and your goal is to answer 4 related questions: (1) is there a linear relationship between the two variables? (2) what is the size of Pearson's r correlation coefficient? (3) what do the regression equation and the regression line look like? And (4) what is the size of r-squared?

Let's start with the first question: is there a linear relationship between the two variables? To answer that question, we make a scatterplot. To make a scatterplot, you must first decide what's the dependent variable and what's the independent variable. In this case it's more likely that a leader's physical height influences his or her approval ratings than that approval ratings affect a leader's height. After all, we don't expect a leader to become taller once his or her approval rates get better… So, the independent variable height goes on the x-axis, and the dependent variable approval rating on the y-axis. Based on the minimum and maximum values of our variables we scale our axes. Our independent variable height ranges from 182 centimeters to 188 centimeters. We therefore use a scale from 180 to 190 centimeters. Our dependent variable ranges from 47 through 60.9. We therefore scale this axis from 45 through 65. Next, we decide, based on our data matrix, where we should position the 4 presidents. Obama is 185 centimeters tall and has an approval rating of 47, so he should be positioned *here*. Bush junior has a physical height of 182 centimeters and an average approval rating of 49.9, so we position him *here*. Clinton and Bush senior are located here.

Now we can answer the first question: Yes, there seems to be a linear relationship between a leader's height and his approval rating. The line describing this relationship goes up, which means that the correlation between the two variables is positive.

The second question is what the value of Pearson's r is. To compute Pearson's r we need *this* formula. To start with, we need to compute all the z-scores of both our independent and our dependent variable. To do that we need the means and standard deviations of these variables. I assume that you know how to compute them, so I will just give them to you. The mean of the independent variable height is 185.75 centimeters and the standard deviation is 2.87 centimeters. The mean approval rating (the dependent variable in our study) is 53.23 and the standard deviation is 6.12.

First we compute the z-scores for our independent variable by subtracting the mean from every original score and then dividing the outcome by the standard deviation. We do that *here*: 185 minus 185.75 divided by 2.87. That makes -0.26132. We also do that for the other scores. *Here* are the results. We then repeat that for the dependent variable. 47 minus 53.23 divided by 6.11 makes -1.01964. And we do that for the other cases too. The next step is to multiply the z-scores of every case with each other. For the first case this results in -0.26132 multiplied with -1.01964. That makes 0.266456 and so on. We have now finished *this* part of the formula. Next we have to add up all these values. That makes 2.202649. Finally, we have to divide by n minus 1. Our n is 4, so n minus 1 equals 4 minus 1 is 3. The result, rounded up, is 0.73. That is our Pearson's r. This indicates that there is a rather strong and positive linear correlation between a leader's body height and his average approval rating.

The next step is to find the regression equation. The computer finds the regression line by looking for the line that minimizes the sum of the squared residuals. You do not have to do this yourself.

Luckily this complicated procedure boils down to two rather simple formulas. One formula to compute the regression coefficient (*this* one), and one formula to compute the intercept (*this* one). Together these formulas give you your regression line. We already have all our necessary ingredients. So now we can use the formulas. The regression slope is 0.73 multiplied with 6.11 divided by 2.87. That makes 1.56. The intercept is 53.23 minus 1.56 multiplied with 185.75. That makes -237.11. The regression equation is $\hat{y}$ minus 237.11 plus 1.56 times X. The intercept indicates that the predicted y-value is minus -237.11 when x is 0. This number has no substantive meaning because a physical height of 0 meter is impossible. The intercept only serves mathematical purposes: it makes it possible to draw the line.

With the regression equation found, we can predict the value of our dependent variable when our independent variable equals 182 centimeters (the minimum value in our sample). That's -237.11 plus 1.56 times 182. That makes 46.81. That's *here*. We can also do that for our maximum value: that's *-237.11 plus 1.56 multiplied with 188. That's 56.17. That's *here*. We can now draw the regression line. This line is the straight line that best represents the linear relationship between X and Y. It is the line for which the sum of the squared residuals is the smallest. We can, of course, predict y-values for every possible x-value. All these predicted y-values, or $\hat{y}$'s, are located on the regression line.

The fourth question we want to answer is what the value of r-squared is. That's easy. It's Pearson's r squared. So: 0.73 multiplied with 0.73 equals 0.53. But how should we interpret this number? Well, we can say that the prediction error is 53 per cent smaller when we use the regression line than when we employ the mean of the dependent variable. We can also say that 53 per cent of the variation in the dependent variable is explained by our independent variable.

So, what have we done? First, we determined the straight line that describes the relationship between our two variables best. Second, we have predicted values of our dependent variable based on this line and the corresponding regression equation. And, third, by means of Pearson's r and r-squared, we have investigated how well this line fits our data. And what have we learned substantively? Well, that tall leaders are more successful than short leaders. However… This conclusion is based on a sample of only 4 American presidents who don't differ much from each other when it comes to their physical height. It is up to you to decide if this warrants far-reaching inferences about the relationship between height and approval ratings.