2 CORRELATION AND REGRESSION

2.05 Regression – How good is the line

In this video I will talk about the question how you can assess how well a regression line fits your data. The reason to look at how well a regression line fits, is that researchers want to know how accurately a regression analysis predicts the dependent variable in a study. The extent to which a regression line fits the data is expressed by means of the so-called **r-squared**.

Imagine you're in a class together with 99 other students and you have just finished a statistics exam. Your professor already has the results for 20 randomly selected students. The professor wants to share the grades of these 20 students, but she doesn't want you to know who these 20 students are. For that reason she anonymizes the grades. This is what you get. Because the results are anonymous, you don't know which student got which grade. Note that the worst grade you could get is a 0 and the best grade is a 10. Now, imagine you are asked to predict the grade of the student sitting next to you in the statics class. What would be your prediction? That would, of course, be the mean of these grades. That is 6.8.

Now imagine that the professor also gives you the grades these twenty students got in a *previous* statistics exam, again anonymized. *These* are the results. How would you now predict the grade of the student sitting next to you? Well, now you can make use of regression analysis. Here you see the scatterplot with the regression line and the regression equation. You see that those with a high grade for the previous exam also tend to have a high grade for the present exam. In fact, you can use the regression line and the corresponding equation to make a prediction. When you ask your neighbor what his previous grade was, you can use the regression line to predict what most likely would be his present grade. Imagine the previous grade was a 8.1. When you fill out the regression equation you get 2.80 plus 0.59 times X equals 7.6. Thus, the present grade would most likely be a 7.6.

What does this mean? When you have information about only one variable, the predictions you make are much less accurate than when you have information about two related variables. r-squared is nothing more than a number that tells you how much better a regression line predicts the value of a dependent variable than the mean of that variable.

Look again at our scatterplot. I have now added a horizontal line that represents the mean of the present exam. The line is horizontal because the mean is always 6.8; that never changes. You can see that the distances from the regression line to the individual cases (the residuals) are much smaller than the distances from the line of the mean to the individual cases. This shows at a glance that the regression line predicts much better than the mean. In our example r-squared is 0.69. This means that the prediction error is 69 percent smaller when you use the regression line than when you employ the mean.

R-squared can also be interpreted in another way. It is the amount of variance in your dependent variable Y that is explained by your independent variable X. The variance of a variable tells you how far the scores are spread out from the mean. So, in our case it means that 69 percent of the variance in the present grades can be predicted by the grades during the previous exam. This interpretation can be represented visually by *these* two circles. The left circle represents the variance of independent variable X and the right circle stands for the variance of dependent variable Y. The overlap represents r-squared or the explained variance. When there is only a little overlap r-squared is small, and when there's a lot of overlap r-squared is large.

One important thing you need to know about r-squared is that it is closely related to the Pearson's R. In fact, the r-squared is, and the name already suggests that, nothing more than Pearson's r ... squared! That is also the way to compute it. Once you have computed Pearson's r, you only need to square this number to obtain r-squared. This means that r-squared is always a positive number. After all, if you square two values, the outcome can never be negative. In our case Pearson's r equals 0.83. When you square that value you get an r-squared of 0.69. Notice that when we have a perfect linear relationship between two variables, both Pearson's r and r-squared are equal to 1. 1 squared is, after all, 1. If there is no linear relationship at all, both values equal 0. 0 squared is, after all, 0.

But you need to remember that the interpretation of r-squared is very different from the interpretation of Pearson's r. Pearson's r tells you whether the linear relationship between two variables is positive or negative, and it shows you how strong this relationship is. R-squared tells you nothing about the *direction* of a relationship between two variables because it's always a positive number. It *does* tell you two other things though. (1) it tells you how much better a regression line predicts your dependent variable than the mean of that variable. And (2) it shows you how much of the variance in your dependent variable is explained by your independent variable.