4.02 Multiple regression: R and R-squared

In this video you'll learn about the **multiple correlation coefficient**, used to assess the strength of association between a response variable and a set of predictors. We'll also discuss the properties of **R-squared**, that tells us how much of the variation in the response variable is explained by our model.

In simple regression we used Pearson's r to determine how strongly the response variable and predictor were associated. In multiple regression we do something similar by determining how strongly the response variable is related to the set of predictors.

We use the **multiple correlation**, which is the correlation between response variable y and predicted scores $\hat{y}$. Since $\hat{y}$ is the linear combination of the predictors it can be used to represent them all in one single variable.

Mathematically the **multiple correlation**, denoted by a capital R, is expressed as $R = \frac{\Sigma(y_i - \bar{y})(\hat{y}_i - \bar{y})}{s_{y_i} s_{\hat{y}_i}}$.

In the best case, if the set of predictors predicts the response variable perfectly, then the observed scores and predicted scores will be the same and the **multiple correlation** will be perfect and equal to one.

In the worst case, if the set of predictors is entirely unrelated to the response variable, the predicted score for all cases will be the mean of the response variable. In that case the second term in the numerator will always be zero, resulting in a **multiple correlation** of zero. Since our predictions can't be worse than the mean, the **multiple correlation coefficient** can't be negative; its lowest possible value is zero.

In this regard - taking only positive values - the **multiple correlation R** is a bit different from Pearson's r. Apart from this, the interpretation of the size is the same as for any correlation coefficient.

Suppose in our example - where we predict the popularity of cat videos using cat age and hairiness as predictors - we find a multiple R of 0.734. This indicates a fairly strong relation between video popularity and the predictors taken together.

R-squared is calculated in the exact same manner we already saw in simple linear regression, we just use a capital R to denote that we are talking about **R-squared** in a multiple regression context. Remember that **R-squared** is the proportion of variation in the response variable, accounted for by the regression model.

We can express it as the total sum of squares minus the residual sum of squares, divided by the total sum squares: $R^2 = \frac{\Sigma(y_i - \bar{y})^2 - \Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}$. All the variation in y - the total sum of squares - minus the error in our model - the residual sum of squares - represents all of the variation in y captured by our model. This captured variation can also be expressed as the regression sum of squares: $\sum(\hat{y}_i - \bar{y})^2$.

In the worst case, when none of the predictors are related to the response variable, the best prediction we have is the mean of y. In that case the regression sum of squares, and thereby **R-squared**, equals zero, since the mean of y - the prediction - minus the mean of y is zero for every element in our sample, we capture none of the variation.
In the best case the predictions - $\hat{y}$ - are identical to the observations - y, in which case the regression sum of squares is equal to the total sum of squares, resulting in an **R-squared** of one; we capture all the variation.

Another interesting property of **R-squared** is that as we add predictors to our model, recalculating R-squared each time, R-squared either stays the same or gets larger. It can never get smaller. To understand this it helps to visualize the variation in the response variable and predictors using a Venn diagram.
This circle represents the variation in the response variable. In simple regression we try to capture this variation with one predictor represented by the second circle. The overlap represents this captured variation: The regression sum of squares. **R-squared** is the overlapping area divided by the total area of the circle representing the response variable.

You can see that if we add another predictor the overlapping area will never become smaller. At worst, the new predictor will be unrelated to the response variable, showing no overlap, or, showing overlap that was already captured by the first predictor.
In the behavioral and social sciences the predictors we use are generally related to each other. This is why after the first couple of predictors **R-squared** usually doesn't increase much anymore.

Both **multiple R** and **R-squared** do not depend on the scale of the response variable. Multiple R is a unitless, positive measure of association. **R-squared** represents a proportion. Just like in simple regression you can find **R-squared** by squaring the **multiple correlation R**. Reversely, you can safely calculate multiple R by taking the square root of **R-squared**, because **multiple R** cannot be negative.