3.04 Simple Regression: Predictive power

In this video you'll learn how to assess whether the predictor in a regression model provides a good description of the response variable. You'll learn to assess predictive power of a regression model by using the **proportion of explained variation**, referred to as **r-squared (r^2)**.

Consider the example where we predicted popularity of cat videos - represented by the number of video views - using the cat's age as the predictor. We hypothesized that videos of younger cats will be more popular. Suppose we collected some data and calculated the intercept and regression coefficient. The obvious question to ask is: How well does our regression model describe the observations?

A very useful coefficient is r-squared, which is exactly what it looks like, the *square* of the correlation coefficient, which thereby varies between zero and one. Let's see how you should interpret r-squared.

Regression tells us whether variation in the predictor goes together - or *covaries* - with variation in the response variable; for example when lower cat age goes together with higher video popularity.
If cat age covaries *perfectly* with video popularity, then all the variation, each change in video popularity, is perfectly predicted or explained by a corresponding change in cat age.
r-squared tells us how closely our sample approximates this ideal situation; it tells us - out of all the variation in the response variable popularity - what proportion is explained by the predictor cat age.

Mathematically all the variation in the response variable is expressed as the total sum of squares. You get the total sum of squares by taking the differences between each observed popularity score - y_i and the mean - $\bar{y}$, squaring these differences and adding them: $SS_{total} = \sum(y_i - \bar{y})^2$.
Don't forget to square; otherwise the negative and positive differences will cancel each other out. Notice that this measure of variation is almost the same as the variance. We just don't divide by n minus one.

So what part of the total variation is explained by our predictor? Well we already know which part it *doesn't* explain. That's the error in our model, called the residuals - the variation that we failed to capture.
Remember, the residual sum of squares is calculated by adding the squared differences between the observations y_i and the predictions $\hat{y}_i$: $SS_{res} = \sum(y_i - \hat{y}_i)^2$.

If we take the total sum of squares and subtract the residual sum of squares we get the *regression* sum of squares: The variation in video popularity that is accurately captured by our model: $SS_{reg} = \sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2$.
If you divide the regression sum of squares by the total sum of squares, you get r-squared: the proportion of variation in the response variable explained by the predictor: $R^2 = \frac{\sum(y_i-\bar{y})^2-\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} = \frac{SS_{total}-SS_{res}}{SS_{total}}$. We can visualize this as the part of the variation around the mean explained by the regression line.
In simple linear regression you can find this proportion by manually calculating the total and residual sum of squares, or you can simply square the correlation; both methods give the same result.

What happens if our model predicts the observations perfectly? Well then the residuals are zero; there is no error. In that case r-squared equals the total sum of squares divided by the total sum of squares, in other words r-squared equals one.

What if the predictor is unrelated to the response variable and provides no help at all in predicting it? The scatterplot will look something like this, with random data points. What is the best prediction you can make in this situation?

Well if the predictor is useless, the response variable provides the only helpful information. The best guess is the mean popularity score of all the videos in our sample. This produces a horizontal line with an intercept equal to the mean of the response variable.
As a consequence the residuals - the differences between each *prediction and observation* - are the differences between each observation and the mean. The residual sum of squares is the same as the total sum of squares; subtracting them will result in zero, so r-squared will be zero. In this worst-case scenario, our model captures none of the variation in the response variable.

In our example the value of r-squared is 0.49, which is a pretty high value for these types of variables. In the behavioral and social sciences relationships between variables are often complicated and influenced by many other factors. This is why - with real data - we're generally already very happy with r-squared values of 0.25.
But you should remember that the value of r-squared really depends on the type of variables you are investigating. In some cognition, medical or biology research fields you might see much higher values.