3.02 Simple Regression: The regression equation

In this video you'll learn what **residuals** are and how they're used to find the **best-fitting straight line** in simple regression. You'll learn to use two formulas to calculate the **intercept** and the **regression coefficient** and how to interpret their values.

Suppose I want to predict the response variable popularity of cat videos - measured as number of views, using the cat's age as the predictor. I collect some data from online cat videos, resulting in this scatterplot. The question is: What line gives us the best predictions?

In simple linear regression we assume the relation between the predictor and response variable is **linear**, so first of all, the line must be straight. Second, we get the best predictions from the line that produces *predicted* scores that lie as close to the *observed* scores as possible.

So we need to find the straight line that *minimizes* the distance between **observed** and **predicted** scores for all the cases in our sample. The difference between the observed and predicted score for each case is called a **residual** or **prediction error.** It's expressed as $y_i - \hat{y}_i$for case i.

Earlier we saw that the general form of the equation is $\boldsymbol{\hat{y}_i = a + b \cdot x_i}$. The predicted popularity score equals the intercept plus the regression coefficient times the age of the cat. So what values for the **intercept** and **regression coefficient** produce the best-fitting line that minimizes the **residuals**?

In order to find these values, we use the **method of ordinary least squares**. We take the **residuals** - $y_i - \hat{y}_i$, we square these residuals and we add them up. We square because otherwise, the positive residuals - where the observations lie above the line - will cancel out the negative residuals - where the observations lie below the line.
Now for the tricky part: We can express the sum of squared residuals as a quadratic function that we can minimize. Mathematically, this involves taking the partial derivatives with respect to a and b, setting them to zero and solving for a and b.

Now don't worry, you can ignore this bit of math; you don't have to perform the ordinary least squares method yourself, because it results in two simple formulas that you can use to easily calculate the **intercept** and **regression coefficient.**

These formulas give you the **intercept** and **regression coefficient** values that result in the smallest possible residuals. These produce the best-fitting regression line, also called the **least squares line**.

The **regression coefficient** is equal to the correlation between x and y - cat age and video popularity - times the standard deviation of y, divided by the standard deviation of x.
Suppose in our sample the correlation between cat age and video popularity is -0.700, and the standard deviations are 0.968 and 4.147 respectively.
Then the regression coefficient is -0.700 times 4.147 divided by 0.968.
Rounded off this equals -3.

Looking at this formula you can see that the **regression coefficient** of x - used to predict y - is an unstandardized version of the **correlation** between x and y. If the correlation is positive, the regression coefficient is positive and the regression line will go up.
If the correlation is negative, so is the regression coefficient, and the regression line will go down. If the correlation is zero, then so is the regression coefficient; the regression line will be horizontal.

Because the correlation coefficient is standardized -it's independent of measurement scales and lies between minus one and plus one, it's useful as a measure of strength of association. Take the correlation of -0.7.
In the behavioral and social sciences we consider an absolute correlation value of 0.3 to be small, 0.5 to be medium and 0.8 to be large. So a correlation of -0.7 indicates a fairly strong relation between cat age and video popularity.

In contrast to the correlation, its unstandardized version - the regression coefficient, is less appropriate to assess strength of association. This is because the size of the regression coefficient depends on the - often arbitrary - choice of scale for the predictor and response variable.
In our example the regression coefficient tells you how much the predicted popularity score will go down if cat age increases with one unit, in this case one year. If we change the age scale from years to months, the regression coefficient will change accordingly, while the correlation stays the same.

Finally we have the **intercept**. The intercept is equal to the mean of y, minus the regression coefficient times the mean of x ($a = \bar{y} - b \cdot \bar{x}$). Suppose in our sample the mean cat age is 1.25 years and the mean video popularity score is 41.20.

We just determined that the regression coefficient is -3, so the intercept equals $41.20\ --3 \cdot 1.25$, which equals 44.95. You can see that this is in fact where the line crosses the y-axis, where x equals zero.

An interesting situation occurs if the regression coefficient equals zero. Then the regression equation reduces to the intercept, which reduces to the mean of y. So the regression line is flat and lies at the mean of y.
This makes sense: What if the predictor - cat age - is unrelated to popularity and doesn't provide any help in predicting it, then what is your best prediction? Well, with only the variable video popularity providing useful information, the best prediction would be the mean popularity score.