3.01 Simple Regression: The regression line

In **simple linear regression** we analyze the *linear* relation between **two quantitative variables,** one **independent** and the other **dependent**. Regression forms the basis for a lot of advanced statistical techniques, so it's really important to understand what regression is. In this video we'll start with **correlation**, the **regression equation**, the **intercept** and the **slope** of the **regression line**.

Ok first, let's see *why we need regression analysis*. Suppose I want to become rich and famous by posting cat videos on the Internet. To be successful it would help to know what characteristics are associated with popular cat videos.
One thing I've noticed is that videos of kittens and young cats seem very popular. To confirm this, I could collect information on some existing cat videos and analyze the relation between popularity - measured by number of video views - and age of the cat - as mentioned in the video description. Suppose a scatterplot of the data looks like this.

To determine how strongly these two **quantitative variables** - popularity and cat age - are **linearly** related we can look at the **correlation coefficient**. This is a number between -1 and +1 that expresses how tightly the data fit around an imaginary *straight line* through the scatterplot.

From the correlation coefficient Pearson's r, we can learn whether the relation is positive or negative. In this case videos become less popular as cat age increases; Pearson's r is negative. We can also see that the correlation is relatively strong.
Of course this means it might be a better idea to record my new kitten instead of my older cat, since a younger age is strongly associated with a higher popularity score.

It would be useful to **describe** the relation more specifically and be able to **predict** an *exact* popularity score based on a cat's age. But the correlation doesn't give this information. This is *why* we use linear regression:
It describes the relation mathematically through a **regression equation**, giving popularity **predictions** for each cat age. This allows us to do a couple of interesting things.

We can use **inferential statistics** to test if the equation is likely to be an accurate description of the relation *in the population*.

We can also see how closely the **predictions** approximate the observed data points, in other words, how good our predictions are. We can use the regression equation to identify **outliers**, data points that deviate strongly from the rest. And finally, we can generate **predictions for new cases**, for example to estimate how popular videos of my new kitten will be.

So how does regression work? Well, in regression analysis we distinguish between an **independent** and a **dependent variable**. The **dependent**, or **outcome**, or **response variable** is the variable we want to predict, in this case video popularity.
The **independent variable** or **explanatory variable**, or **predictor** is the variable that can be used to predict the response variable; in this case we think cat age can predict popularity.
I'll use the terms response variable and predictor from now, because they're shorter and less easily confused. So, the **predictor** always goes on the x-axis. The **response variable**, in this case video views per thousand, always goes on the y-axis.

In many cases it's clear what variable we want to predict and is therefore the response variable, like popularity in our example. In some cases - when the causal direction is unclear - it's arbitrary which variable we consider the predictor and which the response variable. In such cases the choice is simply determined by how we choose to frame the research question.

Remember the imaginary line we just drew when we looked at the correlation? This actually is the **best-fitting straight line** through the scatterplot. We call this the **regression line.** It's described by the **regression equation.** It gives predicted **response variable** scores for each value of the **predictor**.

The equation is $\hat{y}_i = a + b \cdot x_i$. $\hat{y}_i$ is the **predicted score** on the **response variable y** - popularity - for case i given their value x_i on the **predictor x** - cat age. The predicted score is determined by the **intercept** - a - and the **regression coefficient**, or **slope** - b.
The **intercept** a is equal to the value of y when x equals zero, where it crosses the y-axis. It determines where the line is placed.
The **regression coefficient** b determines the slope of the line; it determines whether it goes up or down and how steeply it climbs or falls. It tells us by how much video popularity will decrease if the cat's age increases by one unit, so in this case one year.

Suppose that in our example, a is 44.95 and b is -3. Then the predicted popularity score for a video of a half-year-old cat is: $44.95 - 3 \cdot 0.5$. This equals 43.45 times a thousand, or 43,450 video views.
Similarly, the predicted score for a two-year-old cat is: $44.95 - 3 \cdot 2.0$. This equals 38.95 times a thousand, or 38,950 video views.

As you can see these are the y-values for x is 0.5 and x is 2.0 *on the regression line*; these are the **predicted scores**, which differ from the observed scores.