3.03 Simple Regression: The regression model

In this video we'll look at the **population regression equation** and see how it **models** the relation between predictor and response variable in the **population.** We'll see that it describes the linear relation between the **population means** of the **conditional distributions** of the response variable. Don't worry, you'll understand what that means at the end of the video!

Up until now I've been a bit too simplistic in my explanation of regression. Earlier we looked at an example where we predicted popularity of cat videos - measured as number of video views - using the cat's age as the predictor. In this very *small* sample we only considered the relation between these variables for *this particular* sample. But of course my hope is that the regression equation describes the relation *in general;* so not just for this sample, but for the entire population.
If it does, then I understand the world a little bit better. It also means I can generate useful predictions for *new cases;* I'll be able to predict the popularity of new videos of my kitten and my older cat. So the goal is to **model** the relation at the **population level**. Later on we'll see how we can draw *inferences* about this model of the population.

Ok that all sounds pretty abstract, so what exactly do we model in the population? Of course it's impossible, but suppose for a second that we could gather information about cat age and popularity for *all* cat videos that have *ever* been available online.
You can imagine that if we focus just on one-year-old cats we would find a huge number of videos, with varying popularity scores; they won't all be the same.

The distribution might look something like this, with a mean of 45.00. We could do the same for 1.5 year old cats and 2.74 year old cats, 5.38 year old cats and so on.
For each possible cat age we will find a distribution of varying popularity scores. For example, for the 1.5 year old cats with a mean of 43.75 and for the 2.74 year old cats with a mean of 40.65, and so on.

In simple linear regression we assume that ***conditional*** on the value of the predictor - in other words, for any given cat age - the shape of the distribution of popularity scores looks exactly the same.
Assuming that the relation is perfectly linear, the **population regression line** goes through the means of these distributions. More

formally: The line describes the **population means** of the **conditional response distributions**, which are assumed to have a uniform shape and standard deviation.

We can express this using the following equation: μ_y - the conditional population mean on the response variable y - equals $\alpha + \beta \cdot x$ - the predictor - with the same standard deviation σ at every x.

This looks very similar to the expression we used earlier for the ***sample* regression line**, but with three differences. One is that we use Greek symbols for the intercept and regression coefficient, to indicate we are talking about the ***population* regression equation**.
Another difference is that besides the parameters alpha and beta, we also specify the parameter sigma. Although this parameter is not in the equation it is an essential part of the model. It will become important later on when we use the model for inferential purposes.

The final difference is that we describe the population *means*, not predicted values for individual cases. Modeling the means of the conditional distributions per cat age is important because it allows for natural variation around the regression line in the population.
If we modeled the predicted response value for individual cases, for example by saying $y_i = \alpha + \beta \cdot x_i$, it would mean we expect all cat videos of all one-year-old cats in the population to have exactly the same popularity score. And of course that would be very unlikely.

There is a way to express the model at the individual level, by introducing an error term. The model looks like this: $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$. Epsilon indicates the variation around the conditional mean. It describes the conditional distributions we just saw. Conditional on the value of x, the errors are assumed to be distributed normally with uniform standard deviation sigma and are expected to have a mean of zero. Since inclusion of epsilon can be confusing, many textbooks don't present the model in this form, but I mention it so that if you see it presented like this you don't get confused and think it's an entirely different model.

Ok, back to our model of the conditional means. Ideally the population regression line fits perfectly and goes *exactly* through all these means.
Of course it's unlikely that in the population the means will line up perfectly, they probably won't, but the straight line is assumed to be a close enough approximation to result in a useful *model* for description or prediction.