4.06 Multiple regression: Categorical predictors

In this video you'll learn how to include categorical predictors, called **indicators**, and how to interpret them in multiple linear regression.

In linear regression predictors have to be quantitative; categorical predictors are not allowed. Why is this? Well, the regression coefficients indicate by what amount the response variable changes with a one-unit change in the predictor.
Using linear regression doesn't make sense if the predictor is categorical - if a unit can't be defined for the predictor and distances between scale points can't be determined to be equal.

There's one exception though: If the categorical variable is binary, then there's just one distance between these two scale points, so no need to worry whether the distance between these points is the same as between other points, because there are none.
Such a binary predictor is called an **indicator**. An indicator represents - not the *quantity* of a measured property, but its *quality*. For example an indicator can be used to represent absence or presence, smoking or non-smoking, male or female, control group or experimental group, etcetera.

Consider the simple linear regression example where I predicted popularity of cat videos – measured as number of page views – with the predictor cat age - rated on a scale from zero to ten. I've noticed a recent increase in cat videos where people dress up their cat in a costume, with little hats and clothes.
To see whether wearing a costume is related to popularity I could add the indicator 'costume', which takes on the value zero if the cat is not costumed and the value one if the cat does wear a costume.

Let's look at the regression equation for this model. It's: μ_y - mean popularity in the population - equals $\alpha + \beta_{age} \cdot x_{age} + \beta_{costume} \cdot x_{costume}$. If we look at the equation when the value of costume is zero, the equation changes to $\mu_y = \alpha + \beta_{age} \cdot x_{age}$. When the value of costume is one, the equation becomes $\mu_y = \alpha + \beta_{age} \cdot x_{age} + \beta_{costume}$ or $\mu_y = (\alpha + \beta_{costume}) + \beta_{age} \cdot x_{age}$.

As you can see, this results in just two parallel regression lines. The distance between the lines - for any cat age - equals the size of the regression coefficient for the indicator costume.

[1]Suppose we estimate this model using the example data and find a value of 4.1 for $b_{costume}$. This means that the predicted popularity value for any given cat age is always 4.1 thousand page views higher for costumed cats compared to uncostumed cats. A test of the regression coefficient $b_{costume}$ is essentially a test to see whether the mean popularity score differs for costumed and uncostumed cats, while controlling for the influence of age on popularity.

What if we want to include a categorical predictor with more than two categories? Suppose I want to distinguish between cats in their natural state, cats dressed up in costumes and cats posed with props, for example posed with a book as if they're reading, or with a beer can as if they're drinking. We can't add the value two to the indicator 'costume' to represent cats with props, because then we would be saying the difference in 'dress-up' between the adjacent categories is the same, which makes no sense.

So how do we distinguish between natural cats, costumed cats *and* cats with props? Well, we use a trick; we add *another* indicator. The two indicators used to represent the three categories are referred to as **dummy variables**. We use the indicator 'costume' to indicate the difference between cats with and without costumes. This is our first dummy variable. As before, cats in a costume receive the value one; the rest receive the value zero.

We now create a new dummy variable 'props'. All cats that are posed with props receive the value one and all others receive the value zero.
We don't need another dummy variable to identify cats that appear in their natural state, without costume or without props, because this information is already reflected by a double score of zero on the dummy variables. Remember, you always need one less dummy variable than there are categories to represent. Actually, if you include an extra dummy variable you'll violate one of the technical requirements of regression that there should be no redundancy in the predictors. If you add a redundant dummy variable you'll get invalid results!

An individual test of the regression coefficient for the dummy variable 'costume' indicates whether there's a significant difference between the mean popularity of the costumed and natural cats, controlling for age.
A test of the regression coefficient for the dummy variable props indicates whether there's a significant difference between the mean popularity of cats with props and natural cats, controlling for age.

---

[1] $\hat{y} = 43.0 - 2.7 \cdot x_{age} + 4.1 \cdot x_{costume}$

A final remark: Earlier we saw that the difference in popularity between costumed and uncostumed cats was the same for each age, so costume was unrelated to age. Multiple regression forces the lines to be parallel, in other words, it forces the predictors to be independent.
Now this might represent the actual relation, but it's also possible that *in the population* the effect of costume on popularity depends on, or is related to cat age.

For example, a costume might be very effective for young cats, but could become less effective when cats are older. There are ways to model such a dependent relation - called an **interaction** - by including interaction terms, but we won't go into these methods right here. For now you should concentrate on familiarizing yourself with the concepts of indicators and dummy variables.