4.07 Multiple regression: Categorical response variable

In this video you'll learn how you can use quantitative predictors and indicator variables to predict a binary response variable, using **multiple logistic regression**.

Until now we looked at regression models that describe or predict a quantitative variable, but there are also regression models for ordinal and nominal response variables. In this course we will discuss only binary response variables, for which we use **logistic regression**.

Remember the example where we predicted popularity of cat videos with the predictors cat age, hairiness and presence of a cat in costume?
Suppose I wanted to send in some of my cat videos to the international cat video festival. Each year, out of approximately 10.000 videos only 100 videos are selected and shown. If I could get my hands on a random sample of the 10.000 videos from last year and determine cat age, hairiness and costume, I can see whether these variables help to predict what videos get selected.

So the response variable is one - getting selected - or zero - getting rejected. This is what my data look like in a scatterplot if I look at just the predictor cat age. We see that younger cats seem to get selected more often than older cats.
How can we model this relationship? A straight line would be inappropriate, since it results in estimated values other than zero and one. A discontinuous jump function like this looks good but it isn't ideal for mathematical reasons; we need a continuous, smooth function.

The logistic function, which has a sigmoid shape or s-curve, is a good choice. It produces estimated values that lie between zero and one. This means we have to redefine our predicted variable. We're not predicting getting selected in terms of zeroes and ones, but the *probability* to get selected. We'll change the continuous predicted probability back into a binary variable later on; all cases with a predicted probability of zero point five and higher are predicted festival selections, and all others are predicted festival rejections.

But first let's see what the logistic model looks like. Here's the equation at population level: $p(y) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$. If beta is positive, the curve goes up, if beta is negative the curve goes down. The size of beta determines the steepness of the curve. The horizontal location of the curve is determined by both alpha

and beta. The value $-\frac{\alpha}{\beta}$ determines at what value of x the inflection point lies - this is where the predicted value is exactly 0.5.

Why this logistic function and how do we find estimates for alpha and beta? Well the logistic function has some nice features. By transforming it you can change it into a straight line and use the formulas for linear regression to calculate the intercept and regression coefficient.
First we take the odds of getting selected, this is the probability of getting selected, divided by the probability of not getting selected:

$$odds(p(y)) = \frac{p(y)}{1-p(y)} = \frac{\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}}{1-\left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right)} = e^{\alpha+\beta x}\ .$$

This might seem to complicate matters unnecessarily, but if you go one step further and take the natural logarithm of the odds - also known as the log-odds or logit - the equation simplifies to a linear function: $ln(odds(p(y))) = \alpha + \beta x$.
You don't have to be able to follow along with this; you just have to understand why we use the log-odds - because it gives us a linear function. Also, when you use statistical software to calculate the intercept and regression coefficient they are reported in log-odds!

Let's look at the [1]output for our first example - predicting festival selection with just cat age. The log-odds regression coefficient estimate value is not very easy to interpret. It represents the linear change in log-odds with a one-unit change in the predictor.
The intercept and regression coefficient estimates are often also reported in terms of odds. Since the model in terms of odds corresponds to an exponential function, the regression coefficient can be interpreted as a ratio: it indicates by what multiplicative factor the odds will change with a one-unit increase in the predictor. It says by what percentage the probability of getting selected versus not getting selected will change if the cat is one year older. The regression coefficient expressed in terms of odds is therefore also called an odds-ratio. In this case the *odds* to get selected decrease with about 28 percent with every one-year increase in cat age.

If the odds are not reported we can calculate them from the log-odds by raising *e* to the power of the log-odds: $\text{odds} = e^{\text{log-odds}}$. We can manually calculate odds and probabilities at particular ages, for example at age two,

[1] `Coefficients:`

	Estimate	Std. Error	z value	P-value	exp(Est)
(Intercept)	1.7287	1.6405	1.054	0.292	5.6334
age	-1.2850	0.9876	-1.301	0.193	0.2767

by using the exponential or logistic function. If you find the logistic function hard to work with you can also calculate the probability by dividing the odds by one plus the odds: $p(y) = \frac{odds}{(1 + odds)}$.

The other output includes the standard error, z test statistic and the p-value. The test statistic is a z value, not a t value, since we're modeling proportions, not means of a continuous variable. The interpretation of the test statistic and p-value is the same as in linear regression. The relation between cat age and festival selection is not significant here.

We can add quantitative predictors and binary indicator or dummy variables, just like in linear multiple regression. As you can see here, adding the predictor hairiness and the indicator costume does not help; none of the predictors are significantly related with festival selection, while controlling for the other predictors.

One last bit of output that can be very useful is the **classification table**. This table shows the number of rejected videos that were correctly predicted to be rejected and were incorrectly predicted to be selected. The percentage of rejected videos that were correctly predicted to be rejected is called the **specificity** of our model.

The table also shows how many of the selected videos were incorrectly predicted to be rejected and were correctly predicted to be selected. The percentage of selected videos that was correctly predicted to be selected is called the **sensitivity**. Obviously we want both to be high, just like the overall percentage of correctly classified cases.