4.05 Multiple regression: Checking assumptions

In this video you'll learn how to check whether the **assumptions** of multiple linear regression hold. We'll discuss the assumption of **linearity** of each predictor and the response variable at given values of the other predictors; and we'll discuss the assumptions of **normality**, **homoscedasticity** and **independence** of errors. We'll also look at the technical requirements of **sufficient observations** compared to the number of predictors and the **absence of regression outliers**.

We need to meet these assumptions to ensure that our analysis results in *valid* decisions about our regression model. If they're not met, we might over- or underestimate the p-values of our overall and individual tests, and draw the wrong conclusions.

To explain the assumptions I'll use the example where we predicted popularity of cat videos – measured as number of page views – with the predictors cat age and hairiness - both rated on a scale from zero to ten.

The **linearity** assumption requires that for every predictor the relation between that predictor and response variable is linear, for any given combination of values of the *other* predictors. In simple regression we could check the scatterplot of the response variable and predictor for a linear pattern. In multiple regression - at least with more than two predictors - it's impossible to visualize the relation between the predictors and response variable directly in one graph. Instead, for each predictor, we look at the residuals plotted against the predictor.

The residuals should be scattered around the value zero, with more extreme residuals becoming less frequent, and the variability should be the same for all values of the predictor.
If you see a curved pattern, then the relation between the predictor and response variable is not linear, given the other predictors. Just like in linear regression, determining whether the assumption holds is subjective; it comes down to eyeballing the plot.

We can use the graphs of the residuals plotted against the predictors not just to check linearity, but also to check the assumption of **homoscedasticity**. This assumption requires that for each predictor the variability of the residuals is the same over the entire range of values for that predictor. So the variation in prediction error should be the same for young cats and old cats; the same goes for hairless and very hairy cats.

If the residuals fan out at some point, then the assumption of homoscedasticity is violated and regression analysis shouldn't be performed at all, unless a transformation of the response variable or predictor can fix the problem; but we won't go into using transformations here.

The assumption of **independence of errors** means that the residuals can't be related to each other. Random sampling - or random assignment in experiments - usually ensures this assumption is met. Related residuals can occur more easily in time-series data, but we won't go into this type of analysis here.

The assumption of **normality** requires the residuals to be distributed normally. To check this assumption we look at a histogram of the residuals. Like with linearity, assessing normality is subjective because we rely on visual inspection of a graph.
However, just like simple linear regression, multiple linear regression is robust against violation of the normality assumption. This means deviation from normality is not a problem, as long as the sample is large enough and we use two-sided tests when we conduct individual t-tests. Of course you shouldn't perform multiple linear regression if your sample is very small and the distribution of the residuals is highly skewed.

So how large should your sample be? Well in regression, bigger is generally better. But what's big enough? That's hard to say in general, but a technical requirement is that you need enough observations relative to the number of predictors. If you don't meet this requirement, in some cases it can become technically impossible to even calculate estimated values for the intercept and regression coefficients. As a rule of thumb you need at least ten observations for each predictor in your model.

Another requirement is the absence of influential regression outliers. Regression outliers can substantially alter the value of regression coefficients. To check for outliers you can look for extremely large negative and positive standardized residuals. As a rule of thumb you should inspect standardized residuals more extreme than minus and plus three, since you would only expect to find such extreme values in about one percent of the cases. Remember, only remove an extreme case if there's a clear reason why the data are invalid and should not be in the data set. Data should not be removed just because they don't fit your model!