3.07 Simple Regression: Checking assumptions

In this video you'll learn how to check whether the **assumptions** of simple linear regression hold for a particular data set. We'll discuss the assumptions of **linearity** of predictor and response variable, and **normality**, **homoscedasticity** and **independence** of errors.

Consider the example where we predicted popularity of cat videos - measured as number of video views - using the cat's age as the predictor. Suppose we want to use regression analysis to infer whether the variables cat age and popularity are related or independent.
To ensure that our analysis results in *valid* decisions, we need to make sure the assumptions of linear regression are met. If they're not met, we might over- or underestimate the p-value, and draw the wrong conclusion.

The **linearity** assumption requires that the relation between predictor and response variable is linear in the population. You can use a scatterplot to check for **linearity**, using your sample as a proxy for the population.
Any systematic deviation from a linear, ellipse-shaped cloud of data points should give you reason to reconsider linear regression.
Determining whether the assumption holds is subjective; it comes down to eyeballing the plot. Unfortunately scatterplots can be ambiguous.
Sometimes it helps to plot the residuals against the predictor or predicted values to get a clearer picture. 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 pattern where the residuals are mostly positive for one range of x and mostly negative for another range, then the data are not linear. This could happen for example if video popularity doesn't decrease by much after a certain age, say for cats aged ten and older.

The assumption of **normality** requires the residuals to be distributed normally. One way to check this assumption is to look at a histogram of the residuals. Remember, using ordinary least squares, we found the line that *minimizes* the residuals; so most residuals should be close to zero, with more extreme values being less frequent.
The normality assumption is sometimes presented as pertaining to the response variable instead of the residuals, because if the response variable is distributed normally then so are the residuals. So another way to check this assumption is to check the histogram of the response variable for normality.

Again, assessing normality of the residuals is somewhat subjective because we rely on visual inspection of a graph. Fortunately minor deviation from normality is not a problem, as long as the sample is large enough, since according to the central limit theorem, the sampling distribution of the slope will be normally distributed in the long run, no matter how the response variable is distributed.
Our inference will still be valid, as long as our sample is large enough and we use a two-sided test so that any skew in the test-statistic distribution, resulting in a heavier tail on one side, is compensated by a smaller tail on the other side.
Only if our sample is very small - say less than thirty - and the distribution of the residuals is highly skewed or strangely bimodal do we need to worry.

Although not an 'official' assumption of linear regression, the presence of influential outliers can substantially alter the regression line and therefore the statistical test of the regression coefficient. To check for outliers you can check for extremely large negative and positive standardized residuals. Standardized residuals are like z-scores. The residuals - already having a mean of zero - are divided by their individual standard error. We won't go into how these are calculated. Suffice it to say the scale of the response variable is 'neutralized', making interpretation of their value easier.

Like with z-scores, we only expect standardized residuals with absolute values larger than two for five percent of the cases. Since we would expect several of these in a large data set the rule of thumb is to inspect cases that have an absolute standardized residual larger than three, which we would only expect to find in about one percent of the cases.

An extreme case can be removed if there's a valid reason why the case should not be in the data set. This should only be done if the case truly represents a erroneous observation or misrepresents the population - if a participant didn't take a test seriously, or if a cat video is extremely popular not because of the cat, but because it was posted by a celebrity.

The assumption of **homoscedasticity** means that the variability of the residuals has to be the same for all values of the predictor. So the prediction error should be just as large for old cats as for young cats. You can check this assumption by looking at the scatterplot or at the residuals plotted against the predictor or predicted values.
If the residuals fan out at some point, then the assumption of homoscedasticity is violated and regression analysis shouldn't be performed

at all, or perhaps a transformation of the response variable or predictor is in order.
Here the error in prediction is larger for old cats than for young cats. Maybe young cats are very similar in cuteness. Although cuteness fades with age, maybe some older cats just lose their appeal, while others might compensate their loss of cuteness with funny behavior?

Finally, 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. One situation where related residuals can occur is in time-series data, but we won't go into this type of analysis here.