3.05 Simple Regression: Pitfalls in regression

In this video you'll learn about some potential problems with regression that you should look out for when analyzing data. If these problems occur, correlation and regression results might be heavily distorted.
We'll discuss **nonlinearity**, problematic **outliers**, **erroneously inferring causation** from **correlation**, **inappropriate extrapolation,** the **ecological fallacy** and **restriction of range**.

The first thing to always keep in mind is that correlation and simple linear regression capture **linear** association. To show what can go wrong if you apply a linear model to nonlinear data, Anscombe created different data sets that show linear and nonlinear patterns, but that all result in the same means, standard deviations and correlation for x and y. In all cases Pearson's r is 0.8, and r-squared is 0.64.

The first scatterplot shows the linear pattern you would hope to find. In the *second* plot there is an obvious nonlinear, curved pattern. Remember, these data produce the same Pearson's r and r-squared as the first data set.
Without looking at the plot we might conclude the relation is strong and that the linear model fits well, but obviously a curve would describe these data much better.
There are ways to model this type of pattern - even using multiple linear regression, but we won't go into these methods here. For now you should realize that fitting a *simple* linear model to these data is not the optimal choice.

Two other data sets show how ***outliers*** can have an unwanted influence. In the third data set you can see that the outlier, the deviant data point, messes up a perfectly linear relation. It changes the line's location and slope and lowers our potential r-squared.
In the fourth data set you see the exact opposite. If we discount the outlier, then there's no variation in x, so the correlation is zero. But by adding the outlier, Pearson's r and r-squared suddenly become quite high. Obviously this regression line strongly misrepresents the actual relation between x and y.

Outliers can have an unwanted influence if they are extreme on one or both variables *and* deviate strongly from the regression line - if they are **regression outliers**. Outliers are more problematic in small samples, where the influence of each case is relatively strong.

Another problem is **erroneously inferring causation** from **correlation**. For example, just because regular intake of vitamin supplements is related to greater health, this doesn't mean we can infer use of supplements improves health.
Perhaps healthier people proactively pursue a healthy lifestyle and use supplements more often because they think it's healthy.
Just because we can use regression and use vitamin intake to predict health doesn't make the relation causal, we could also use health to predict vitamin intake. The only way to tell if the relation is causal is to perform a truly randomized experiment.

Inappropriate extrapolation is another problem in regression. Take the example where we predict popularity of cat videos using the cat's age. As the age of the cat increases, video popularity goes down. But our sample only covers cat ages from three months - 0.25 years, 2.5 years.
We can't just extrapolate - extend the regression line - endlessly beyond this range. For example, it wouldn't make sense to predict video popularity for a hundred-year-old cat, simply because cats don't reach that age.
It's also possible that between the ages of 5 and 25, for example, the drop in popularity actually decreases nonlinearly. So we should be careful when extrapolating beyond the sampled range.

The **ecological fallacy** refers to drawing inappropriate conclusions about individual cases when correlation or regression is based on aggregates of these cases. For example if we have a lot of data on the relation between vitamin intake and health from different countries, we could aggregate over countries.
Aggregation eliminates individual variability and tightens the data points around the best fitting line. As you can see, the ecological fallacy can heavily affect the correlation, and the same goes for regression.

As long as the relation is the same at the individual and country level, the location of the regression line will not change much. R-squared will be higher however, resulting in overestimation of the fit of the model.
Of course aggregation is ok as long as we don't use results based on the aggregate data to draw conclusions at the level of individuals.

The final potential problem I want to mention is **restriction of range**. Restriction of range means that our sample contains a limited range of predictor values. In our sample cat age varied between 3 months and 2.5 years. The average lifespan of an indoor cat is about 15 years, so our range is restricted; we are missing values between 2.5 to 15 years.

The correlation in our sample is 0.7. If we had collected data in the missing range - assuming the linear-relation continues to hold - the scatterplot would look a cloud of data points that is more ellipse-shaped than the limited sample. Restriction of range can seriously lower Pearson's r and r-squared; the location of the regression line is less affected, however.

As you can see from these examples it's important to:
- always look at a scatterplot of the data to check for non-linearity and outliers,
- to consider how representative the range of data is, and
- to consider what kind of inferences you're making.