5.06 Analysis of Variance: ANOVA and regression

In this video we'll discuss the link between ANOVA and multiple regression. You may have noticed that the between and within-group variances are expressed in terms of sums of squares and mean squares, which we also use in multiple regression. It's also not a coincidence that the table of mean sums of squares and sums of squares in multiple regression is often referred to as the ANOVA table. In fact, multiple regression and ANOVA are technically the same.

To illustrate this we'll perform a one-way ANOVA using multiple regression.

Let's consider our one-way example where we compare healthiness of three groups of cats that consumed different diets: Raw meat, canned food and dry food. A vet rated cat health on a scale from zero to ten.

We have a quantitative response variable and a categorical independent variable or indicator. The indicator has three levels, and only binary indicators are allowed in regression. So we need to create two dummy variables to identify these levels. The first dummy variable x_{raw} distinguishes between cats fed on a raw meat diet and all other cats. The second dummy variable x_{canned} distinguishes between cats on canned food and all other cats. Cats fed on dry food are identified by a score of zero on both dummies.

The regression model is $\mu_y = \alpha + \beta_{raw} \cdot x_{raw} + \beta_{canned} \cdot x_{canned}$. How do we interpret the intercept and the regression coefficients? We'll consider the regression model for each diet group. For cats fed on raw meat, the population mean equals $\alpha + \beta_{raw}$. For cats fed on canned food the population mean equals $\alpha + \beta_{canned}$. For cats fed on dry food the population mean equals α.

So the intercept represents the population mean of the last group - fed on dry food, called the reference group. From this it follows that the regression coefficient β_{raw} represents the difference in the population mean of the raw group minus the population mean of the dry food group - the reference group. β_{canned} represents the difference in the population mean of the canned food group minus the population mean of the dry food group.
With an overall F-test we test the null hypothesis that all regression coefficients are zero ($\beta_r - \beta_c = 0$). So we're testing whether the difference between the raw and dry food group is zero and whether the difference between the canned and dry food group is zero ($\mu_r - \mu_d = \mu_c - \mu_d = 0$). If we rewrite this equation we can see that we also implicitly test

whether the raw and canned group differ from zero ($\mu_r - \mu_d - \mu_c + \mu_d = 0$ => $\mu_r - \mu_c = 0$).

Since we have only two indicators we can represent the data visually in a three-dimensional graph. As you can see the three groups are located at the corresponding values zero and one of the dummy variables. The plane goes through the means of these groups.
The observations are scattered around these means. In multiple regression the null hypothesis corresponds to a flat plane, where the means are all the same, resulting in regression coefficients of zero. As soon as one or more means differ from the rest the plane will be tilted.

If we recall the visual representation of the null-hypothesis in one-way ANOVA you can see that the flat plane corresponds with population distributions with the same mean and the alternative corresponds with population distributions with different means.

In regression, the predicted health value for cats that eat raw food is the mean health score in the raw meat group. The same goes for the other groups. In multiple regression the variation in the residuals or prediction errors is the variation in the observations in each group around the group mean. So the *residual* or *error* mean sum of squares in multiple regression *is* the *within-group variance* in ANOVA.

The *regression* mean sum of squares in multiple regression is the variation in the predictions around the mean of the response variable. This corresponds to the variation of the group means around the grand mean in ANOVA. So the *regression* mean sum of squares in multiple regression *is* the *between-group* variance in ANOVA. Of course the individual t-tests of the regression coefficients do not necessarily correspond to the post-hoc comparisons in ANOVA. But the overall F-test in regression and the F-test in ANOVA are not just similar; they are the same!