4.03 Multiple regression: Overall test

In this video you'll learn how to test whether the predictors, together - as a set, are significantly related to the response variable. This test is referred to as the overall F-test of a multiple regression model.

Consider the example where we predicted popularity of cat videos–measured as number of page views–with the predictors cat age and hairiness–rated on a scale from zero to ten.
An overall test helps us decide whether cat age and hairiness, taken together, are related to video popularity in the population.

As always we start by specifying the null hypothesis. If there's no relation between the predictors and the response variable, this means that neither cat age nor hairiness helps to predict popularity. In other words the regression coefficients for both these predictors will be zero. We can visualize this as a flat plane in a three-dimensional graph.

The alternative hypothesis is that at least one of the predictors is related to the response variable. If cat age, hairiness or both are related to video popularity, the plane will no longer be flat. In other words at least one, several, or all of the regression coefficients will differ from zero.
This alternative hypothesis is very general; if there's a relation between the set of predictors and the response variable, we still don't now which predictors contribute. To find out which predictors contribute we will follow-up with individual tests of the regression coefficients later on. But for now we'll focus on the overall test.

The overall test, like always, is associated with a number of assumptions that need to be met in order for the test to give valid results. These assumptions are: **linearity** of each predictor and the response variable, for each value of the other predictors, and **normality**, **homoscedasticity** and **independence** of the residuals. Another, more technical requirement is that you need enough observations relative to the number of predictors.

I'll discuss the assumptions and how to check them later on. First let's see how to perform the test. To compute the **test statistic F**, we take the regression and error sums of squares we saw earlier when we calculated R-squared, we turn these sums of squares into variances and we divide them.

First we divide the regression sum of squares - the variation in the response variable captured by our model - by k minus one. k is the number of parameters in the model, which equals the number of predictors plus one for the intercept. The regression sum of squares divided by k minus one is called the **regression mean square,** Now don't be confused by this new term 'mean square'; it's just another word for variance. It's the variance in the response variable captured by our model.

To get the F value, the regression mean square is divided by the residual sum of squares, divided by the number of observations n minus k - the number of parameters. The residual sum of squares represents the variation in the response variable *not* accounted for by our model. If we divided it by n minus k we turn it into the residual *mean* square or **mean square error**, often abbreviated with MSE. Of course, this mean square error is no longer the variation but the variance of the residuals: The variance in the response variable that we failed to capture with our model.

So the F test statistic is the 'explained' variance divided by the 'error' variance. Here's an example of an F distribution. As you can see the lowest possible value is zero, which occurs when the regression mean square equals zero, when our model captures none of the variation in the response variable. As our model captures more of the variation the F value goes up.

The exact shape of the distribution is determined by two separate degrees of freedom. The first equals the number of parameters in the model minus one. The second equals the number of observations minus the number of model parameters. Again, the number of model parameters equals the number of predictors plus one for the intercept.
Notice that we used these values to turn sums of squares into mean squares earlier. This is why the first degree of freedom value is often referred to as the **numerator or regression degrees of freedom** and the second is often called the **denominator or error degrees of freedom**.

Once we've calculated the F statistic and the degrees of freedom, we can calculate or look up the associated p-value. We don't have to worry about choosing between the left or right tail here.
This is because the alternative hypothesis is non-directional; it only specifies that one or more predictors are related to the response variable but not which ones and not in which direction.
This means we always look in the right tail to obtain the probability of finding the calculated F value or a more extreme value.

Suppose in our example we find a regression sum of squares of 50.5 and a residual sum of squares of 18.3. The regression mean square then equals 25.25 since we divide by two. The mean square error equals 18.3 divided by 5 minus 3 is 2, which equals 9.15. This gives us an F value of 2.76. The p-value, calculate with statistical software, equals 0.266. If we use the table to look up the critical F value, we see that our calculated value 2.76 does *not* exceed the critical value of 19.000.

This means that we cannot reject the null hypothesis and cannot conclude that cat age or hairiness, or both, are related to video popularity.