4.01 Multiple regression: The regression model

In **multiple regression** we use not one, but several quantitative predictors to predict a quantitative response variable. In this video you'll learn why multiple regression is useful and how we express a **multiple regression model** at the **sample** and **population** level and how to interpret the **regression coefficients** and **intercept**.

Earlier I tried to predict the popularity of cat videos - measured by number of page views - using cat age. If I were to collect actual data I would probably find that cat age doesn't predict popularity very well. This is no wonder: The popularity of cat videos is undoubtedly influenced by a lot of other things, besides the cat's age.

Possible other relevant characteristics are the cat's fluffiness or hairiness - in terms of hair length, its attractiveness, how funny its behavior is or to what extent it mimics human emotions. With multiple regression, I can add such variables as additional predictors, which will - hopefully - result in a **more appropriate**, **better fitting model** and **better predictions**.

I can also add variables to **control** for their possibly **confounding influence**. For example, the time a video has been available online will influence its popularity; it has nothing to do with the attractiveness of the video, but it might explain why some videos of very cute and funny kittens aren't as popular as expected and some videos of older cats are more popular than expected.

Ok so what does the model look like? Well it's an extension of the simple linear model. At the **sample level** we express it as $\hat{y}_i = a + b_1 \cdot x_{i1} + b_2 \cdot x_{i2}$ on until we reach the last predictor, denoted m, so we end with $\ldots + b_m \cdot x_{im}$. Note that the sub-i's indicate that y and the x's stand for individual values.

At the **population level** we express the model as $\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$. To understand how to interpret this model, let's consider a simple example with only two predictors. Suppose we add hairiness as a predictor to model video popularity. Hairiness is rated on a scale between zero and ten, with zero meaning hairless and a ten meaning longhaired, like a Persian cat. Say we find this regression equation: $\hat{y}_i = 34.372 - 1.775 \cdot age_i + 1.414 \cdot hair_i$.

We can visualize this model by considering the relation between cat age and video popularity at particular values of hairiness. Say we take

hairless cats, with a hairiness score of zero. Given this hairiness-score, what is the relation between age and popularity? Well if we fill in zero in the equation we simply get $\hat{y}_i = 34.372 - 1.775 \cdot age_i$.
This can be drawn as a simple regression line. Now consider the relation for a hairiness-score of one. If we enter a hairiness score of one in the equation we get $\hat{y}_i = 34.372 - 1.775 \cdot age_i + 1.414$, which equals $\hat{y}_i = 35.786 - 1.775 \cdot age_i$. If we enter a hairiness score of two we get $\hat{y}_i = 34.372 - 1.775 \cdot age_i + 2.828$, which equals $37.200 - 1.775 \cdot age_i$.

The regression lines, predicting popularity with cat age at given values of hairiness, all run parallel. From this we can see that in multiple regression, for a particular predictor, the regression coefficient gives you the change in the response variable per unit increase of that predictor, *given the values of the other predictors*.

It's important to note that the size of each regression coefficient depends on the scale of the predictor. So we can't say that the predictor age - which is larger - is more influential in predicting popularity than hairiness; age ranges from zero to about fifteen, while hairiness ranges between zero and ten.

Another thing to note is that the value of the regression coefficient for age in our multiple regression equation is different from the value in the simple regression equation, even though the observations are the same.
In the simple case with just age as predictor we consider the relation between cat age and popularity while ignoring all other variables. By adding hairiness as a predictor we *control for the effect of hairiness* when we consider the relation between cat age and popularity. We consider the relation for each level of hairiness, which might result in a stronger or weaker relation between cat age and popularity.

We can visualize the entire model by adding another axis - the z-axis - to represent hairiness. You can see that the parallel lines now form a plane in a three-dimensional graph. This plane represents the predicted values produced by the model. The intercept a is where the plane crosses the y-axis. So the intercept a represents the predicted value when cat age and hairiness are both zero.

Just like in simple linear regression I can calculate the residuals, the vertical distances between the observations and the predicted values, which in this case lie on the regression plane.
These residuals are used to find the intercept and regressions coefficients that provide the best-fitting plane through the data points.

Just like in simple regression, the residuals are minimized using the method of ordinary least squares. Because the resulting formulas for the intercept and regressions coefficients are more complicated we'll use statistical software to calculate them.

At the population level we model the means of the conditional distributions, μ_y. For every point on the plane, for every combination of cat age and hairiness, we assume there is a distribution of popularity scores. The mean of this distribution lies on the plane. The standard deviations of all these conditional distributions are assumed to be identical, so the spread of observations around the plane is assumed to be the same everywhere.

With more than two predictors it's no longer possible to represent the model visually in one graph but the logic and the interpretation of the intercept and regression coefficients is the same.