StatQuest: Linear Models

Part 1. Linear Regression

(keywords: residual, least-squares, sum of squares (SS), R^2, p-value)

Linear regression attempts to model the relationship between a variable y and the explanatory factors X using a linear function. For example, the least-squares method calculates a function that minimizes the sum of the squares of the residuals:

$$\operatorname*{argmin}_L \sum_{i=0}^{n} (y_i - L(X_i))^2 \tag{1}$$

Given the data and a function, we define R^2 as the reduction in the variance of y by taking X into account:

$$SS(mean) = \sum_{i=0}^{n} (y_i - \bar{y})^2 \tag{2}$$

$$SS(fit) = \sum_{i=0}^{n} (y_i - L(X_i))^2 \tag{3}$$

$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)} \tag{4}$$

R^2 is simple and intuitive, but it comes with a pitfall — it doesn't reflect the dimension of a function with respect to the number of observations that the function is fitted on. For example, if we fit a line on two data points, the R^2 on the line will always be 100%. Hence we introduce a *p-value* on R^2 to represent its statistical significance. The *p-value* is based off of something called F:

$$F = \frac{(SS(mean) - SS(fit))/(p_{fit} - p_{mean})}{SS(fit)/(n - p_{fit})} \tag{5}$$

The terms to the left are similar to those in R^2, except that the denominator becomes the residual sum of squares with the fitted function. The terms to the right are called "degrees of freedom", they turn the sums of squares into some kind of variances.