THE COEFFICIENT OF DETERMINATION R^2

We want to derive an indicator that measures the strenght of the relation between x and y. In other terms, a coefficient that summarizes how informative x is to study y.

Setting: we observe two variables x and y on n units

A first descriptive statistic that we can compute is the correlation coefficient:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \in [-1,1]$$

where
$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$
$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

which is a measure of the strenght of the LINEAR RELATIONSHIP between the two variables.

In the context of linear regression, we can derive another (related) quantity to assess the strenght of the linear relationship between x and y: the coefficient R^2

This coefficient can be generalized to the case of p covariates $x_1, \dots, x_p$ (unlike ρ_{XY}).

Recall the sum of squares decomposition:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$SST = SSR + SSE$$

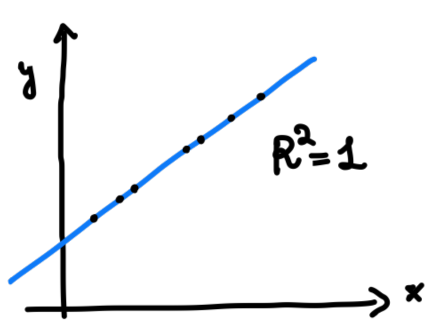
If the model fits the data well, I expect the SSR to be larger than the SSE (w.r.t. the SST, which does not depend on the model - is fixed given the data).

Hence I can study the ratio SSR/SST to understand how much variability is explained by the model.

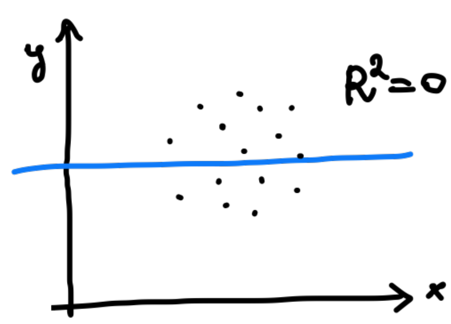
The COEFFICIENT OF DETERMINATION R^2 ("R-squared") is the PROPORTION of VARIABILITY of the DEPENDENT VARIABLE that is explained by the covariate.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \in [0,1]$$

The extremes are:



perfect alignment
$y = ax + b$



no relationship

The coefficient R^2 is a measure of the GOODNESS OF FIT of the model (how adequate it is to summarize the relationship between x and y with a straight line, i.e. the estimated model).

In the case of the SIMPLE linear model, $R^2 = r_{xy}^2$.