Correlation vs Regression

Oliver Snellman
oliver.snellman@gmail.com

May 2022

Abstract

Correlation and regression are both tools to describe a certain type of relation between two variables. In this document I show visually how they differ.

Pearson's correlation coefficient ρ measures, how much two datasets, X and Y, co-vary linearly.

$$\rho = \frac{\mathbb{E}[(X - \bar{X})(Y - \bar{Y})]}{\sigma_X \sigma_Y} \in [-1, 1]$$

where $\bar{X} = \mathbb{E}[X]$ is the expected value and $\sigma_X = \mathbb{E}[(X - \bar{X})^2]^{\frac{1}{2}}$ is the standard deviation of X, and similarly for Y.

Correlation coefficient ρ describes, how tightly the points (X,Y) align to a single line, when plotted on a 2D plane.

Correlation coefficient does not tell anything about the relative sizes, that the values of X and Y might have (that is what regression does).

Linear Regression coefficient β_1 measures, how one variable Y linearly depends on another variable X.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\beta_1 \in (-\infty, \infty)$$

where β_0 is an intercept and ε is an error term.

Regression coefficient β_1 predicts, how much the value of Y is expected to increase, when X increases by one.

Regression coefficient does not tell anything about the accuracy of the prediction; how close the values of Y in the dataset are from the prediction (this is expressed by ε and the correlation coefficient).

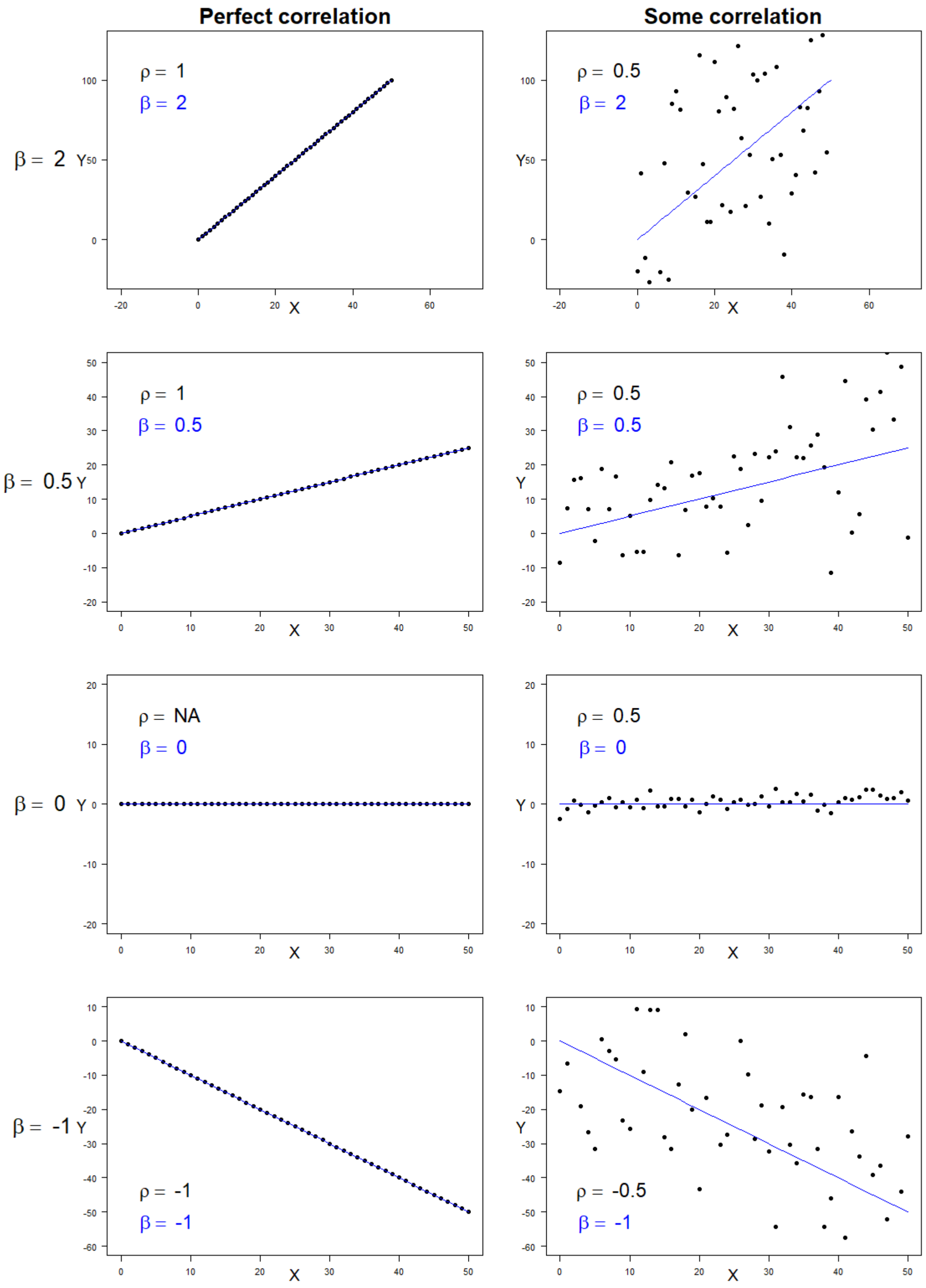

Figure 1: Correlation equals one for all points in a straight ascending line, it is undefined for the horizontal line because of a zero in the denominator, and it equals minus one for all descending straight lines. In the case of a straight vertical line, correlation is yet again not defined due to having zero in the denominator. But this time also linear regression (OLS) is undefined, as any linear curve could equally well represent the data. Simulations were conducted with R.