Correlation vs Regression

Oliver Snellman
oliver.snellman@gmail.com

May 2022

Abstract

Correlation and regression coefficients both describe a certain type of relation between two variables. Here I provide a visual intuition on how they differ.

Pearson's correlation coefficient $\boldsymbol{\rho}$ expresses how much variables X and Y co-vary *linearly.*

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

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

Correlation coefficient $\boldsymbol{\rho}$ **describes** how tightly the points (X,Y) align on a plane.

$\boldsymbol{\rho}$ **does not** tell us about the relative sizes of X and Y, that is what β_1 does.

Linear Regression coefficient $\boldsymbol{\beta_1}$ expresses how one variable Y depends on another variable X *linearly.*

$$Y = \beta_0 + \boldsymbol{\beta_1} X + \varepsilon \tag{1}$$

where $\beta_0, \boldsymbol{\beta_1} \in \mathbb{R}$, β_0 is an intercept and ε is an iid error term.

Regression coefficient $\boldsymbol{\beta_1}$ **predicts** the increase in Y, when X increases by one.

$\boldsymbol{\beta_1}$ **does not** tell us how close the values of Y are from the prediction line, that is expressed by $\boldsymbol{\rho}$ and ε.

Next I visualize how variables can perfectly correlate $|\boldsymbol{\rho}| = 1$, while the regression coefficient $\boldsymbol{\beta_1}$ differs. Likewise, the same value of $\boldsymbol{\beta_1}$ can occur for differing $\boldsymbol{\rho}$.

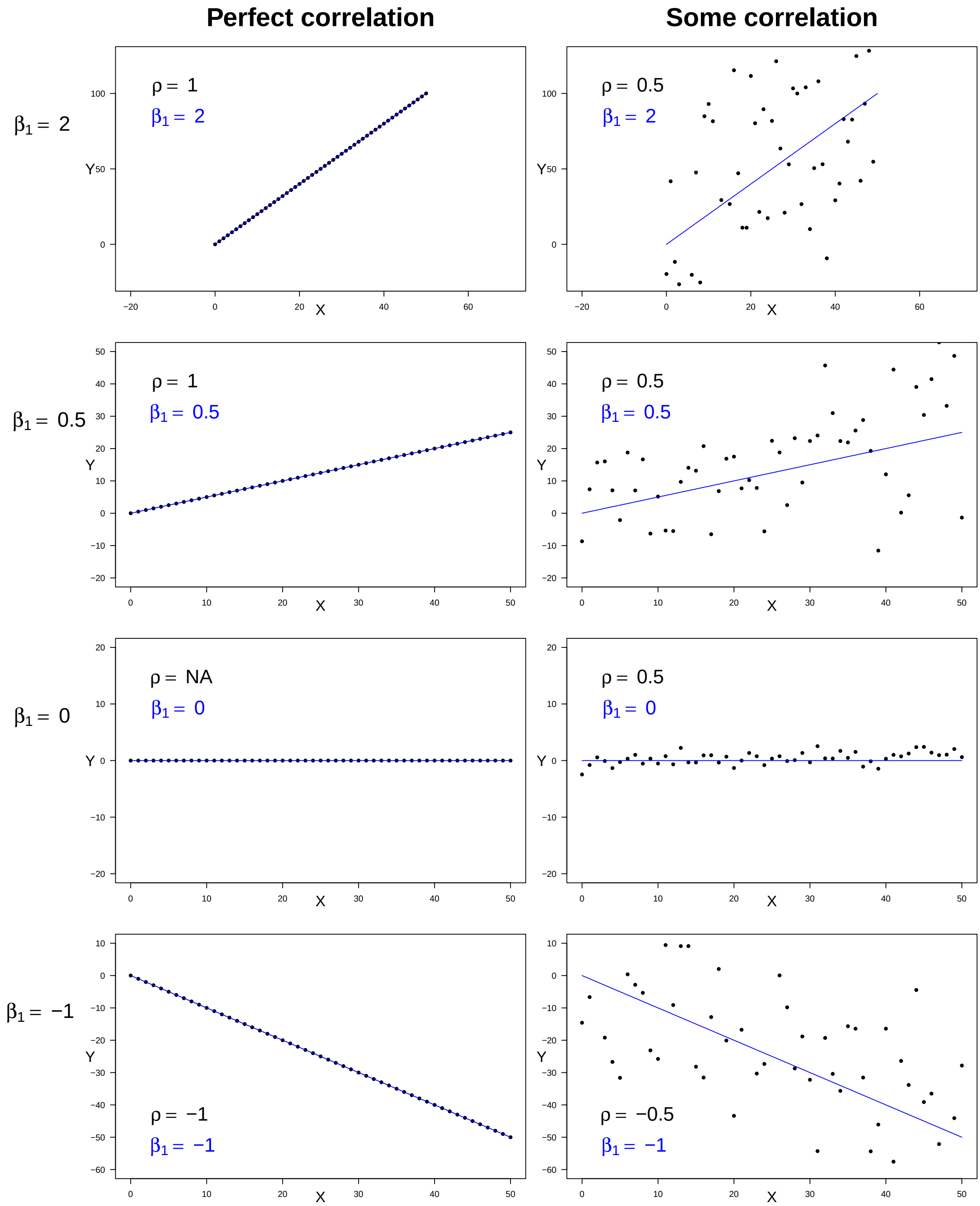

Figure 1: Each scatterplot has 51 observations with X and Y values, simulated from a linear model in Eq 1 without and with an error term. In the **left column** (no error terms) we can see that the correlation $\boldsymbol{\rho}$ equals one for points in all straight ascending lines, it is undefined for the horizontal line because of a zero in the denominator, and it equals minus one for all descending straight lines. **The rows** illustrate that the regression coefficient $\boldsymbol{\beta_1}$ (the slope of the blue line) can have the same value for differently correlated variables. I left out the case of a straight vertical line, because it would require more explaining.