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 ρ expresses how much two variables, X and Y, co-vary linearly.

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

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

ρ **does not** tell us about the relative sizes of X and Y, that is what β_1 does.

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

$$Y = \beta_0 + \beta_1 X + \varepsilon$$
$$\beta_0, \beta_1 \in \mathbb{R}$$

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.

β_1 **does not** tell us about the accuracy of the prediction; how close the values of Y in a dataset are from the prediction, that is expressed by ρ and ε.

In the next page I visualize how variables can perfectly correlate $|\rho| = 1$, while the value of the regression coefficient β_1 differs. Also, the same value of β_1 can occur for different values of ρ.

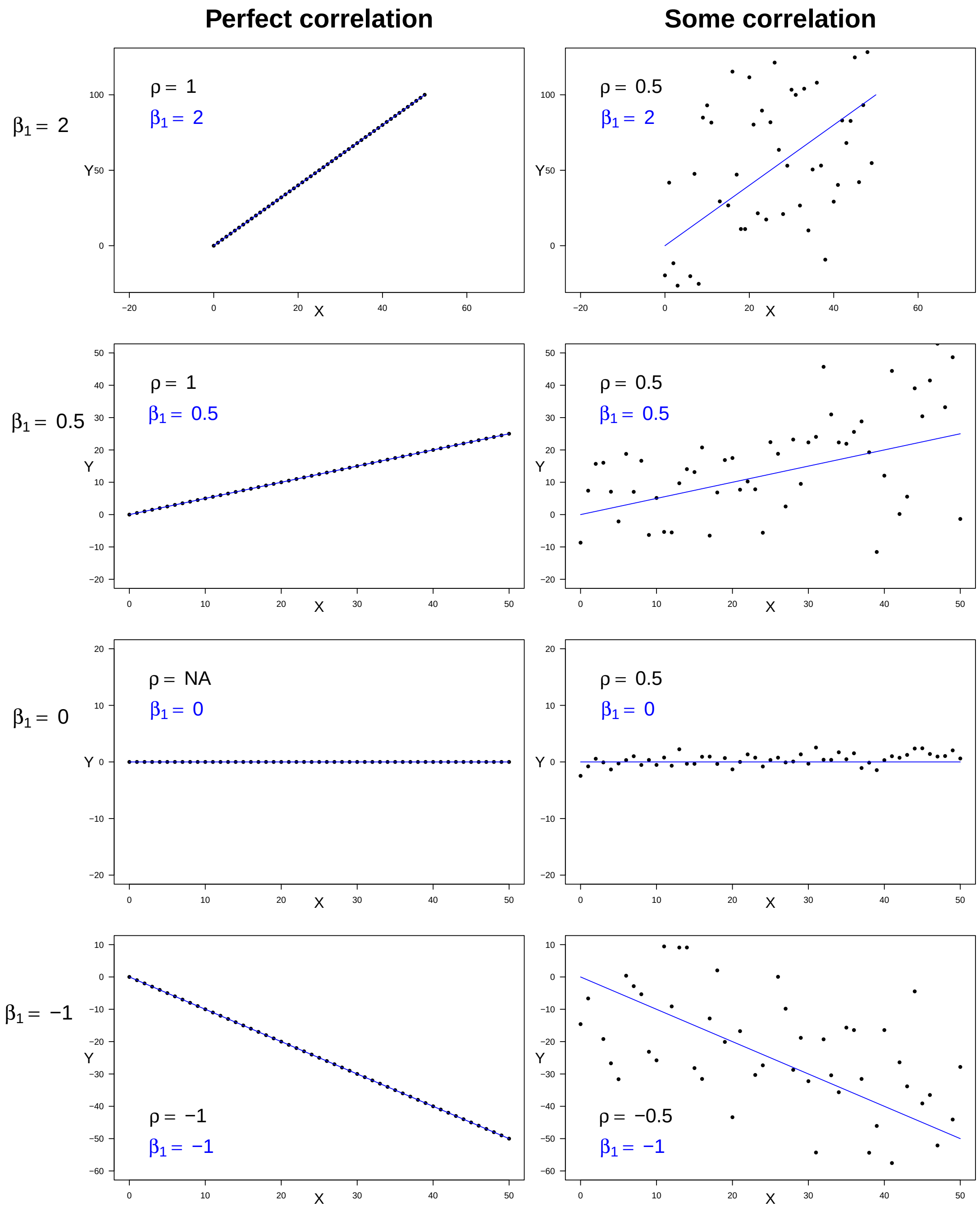

Figure 1: Each scatterplot has 51 observations with X and Y values, simulated with R. In the **left column** we can see that the correlation ρ equals one for points in all straight ascending lines, it is undefined for the horizontal line because of a zero in the denominator, and it equals minus one for all descending straight lines. **The rows** illustrate that the regression coefficient β_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.