Correlation and Linear Regression

OpenIntro Statistics Chapter 8

Correlation (*r* or *R*)

A measure of the strength of the linear relationship between two variables.

Takes values between -1 and 1.

<u>*General*</u> Rules of Thumb:

$r \leq \lvert .20 \rvert$	**Weak** relationship
$\lvert .20 \rvert < r \leq \lvert .50 \rvert$	**Moderate** relationship
$r > \lvert .50 \rvert$	**Strong** relationship

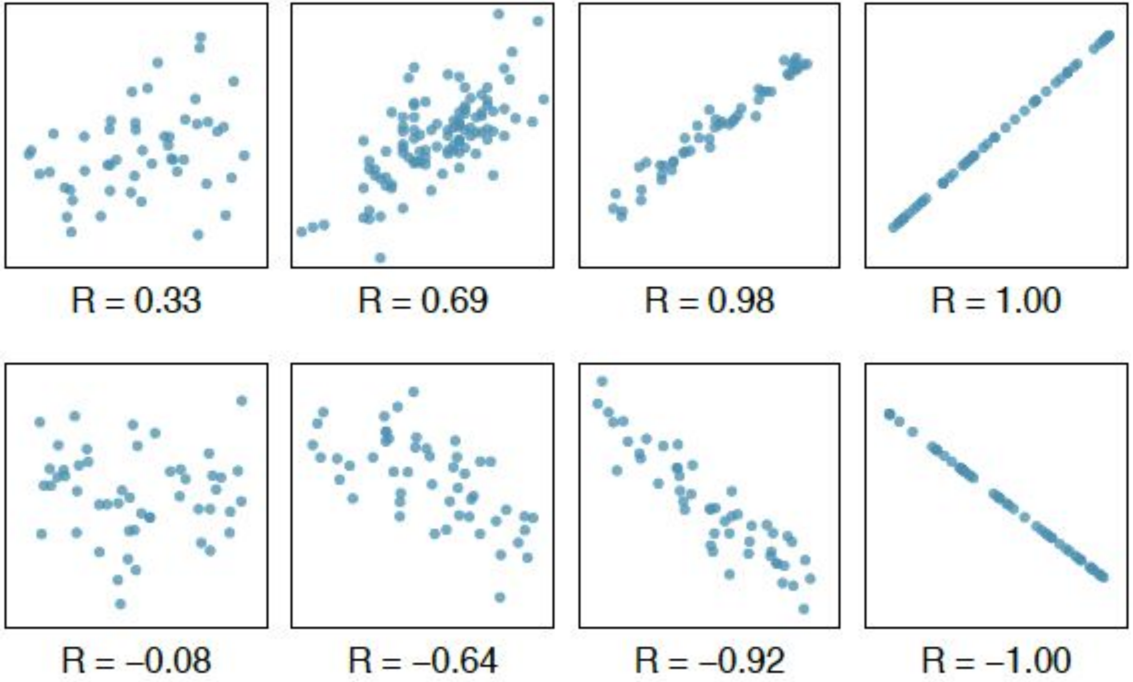
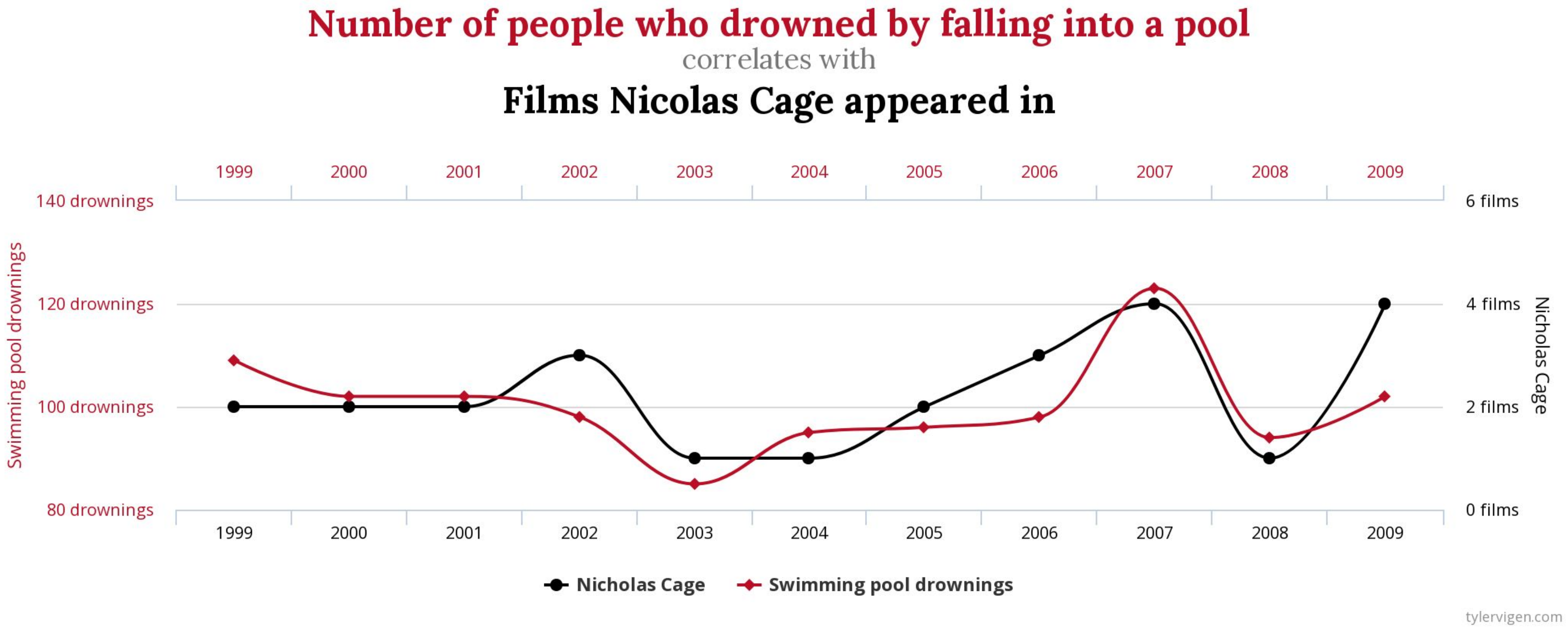


Figure 7.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

https://openintro.org OpenIntro Statistics, p. 339

Cautions about Correlation

Beware of spurious correlations! (especially when you have a lot of variables and not a lot of observations)

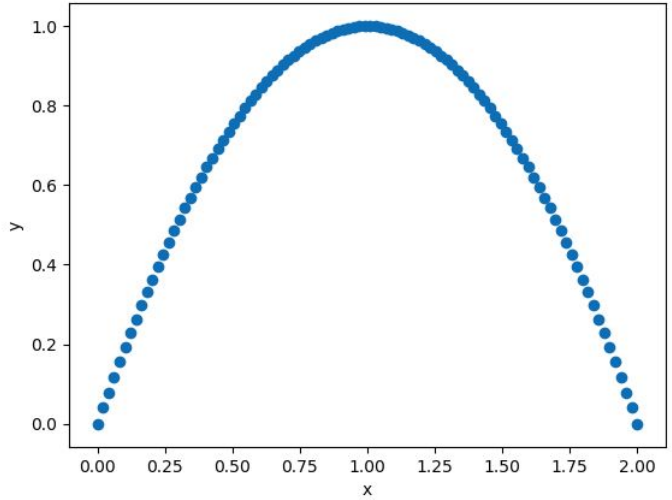


Cautions about Correlation

Correlation does not imply causation!

Independence implies zero correlation, ***but*** zero correlation <u>does not</u> imply independence.

These variables have 0 correlation, but there is a clear relationship between the two.



Ordinary Least Squares Regression

$$y = \beta_0 + \beta_1 x + \epsilon$$

Response Variable

Predictor Variable

Normally Distributed
Mean = 0, Constant Variance

Ordinary Least Squares Regression

$$y = \beta_0 + \boxed{\beta_1} x + \epsilon$$

A one unit change in the predictor variable will result, on average, in this big a change in the response variable.

Assessing Fit of an OLS Model

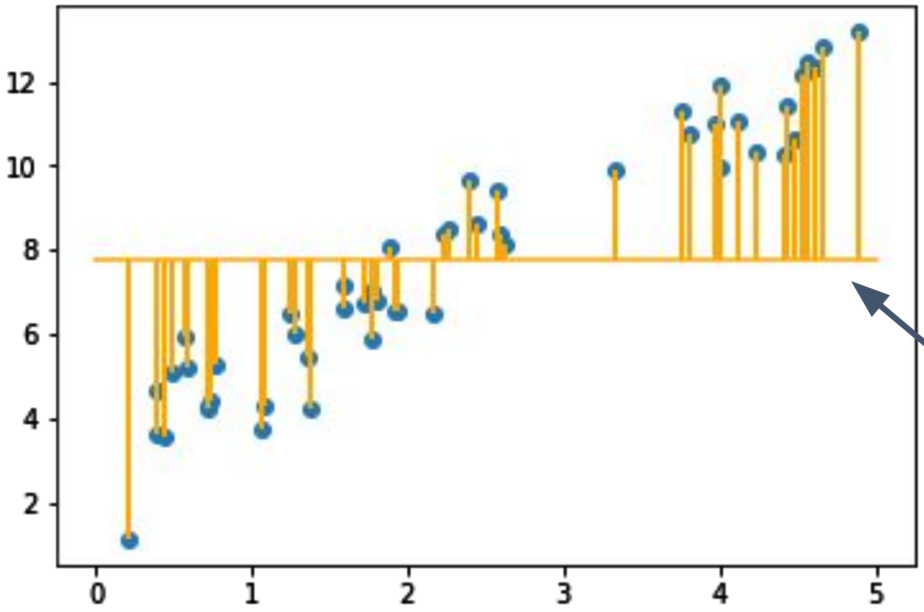
R^2: Measures the amount of variation in the response variable that is explained by the least squares line.

Takes values between 0 and 1, and larger is better.

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

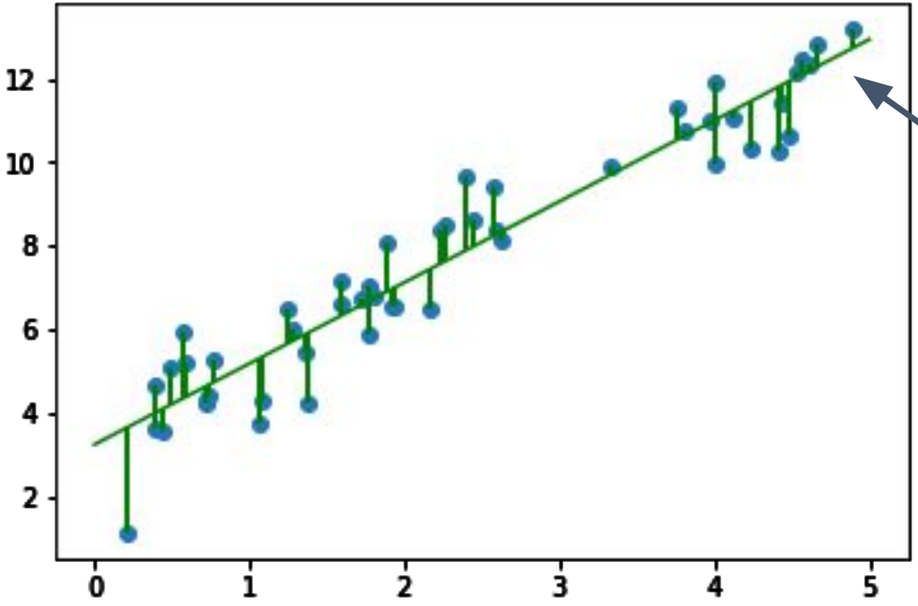
TSS = Total Sum of Squares

RSS = Residual Sum of Squares



Average Observed y

TSS = Total Sum of Squares
The total squared distance between the response values and the average response value.



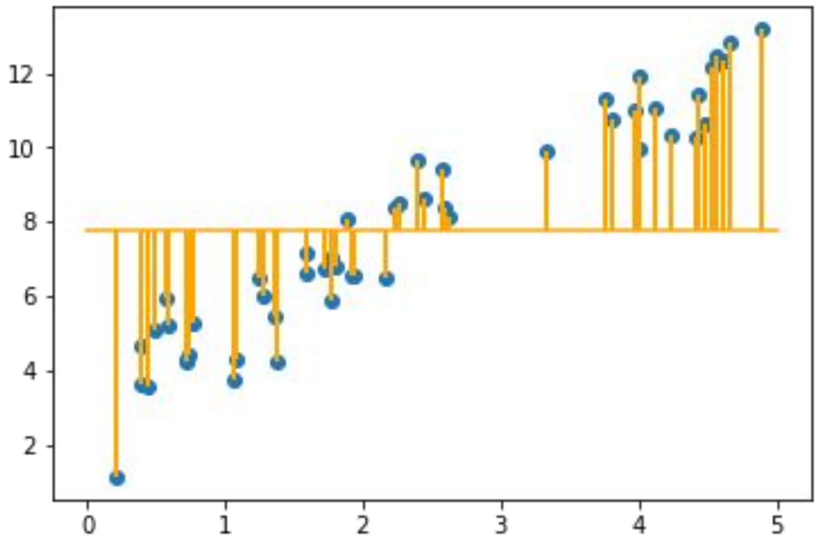
Fitted Regression Line

RSS = Residual Sum of Squares

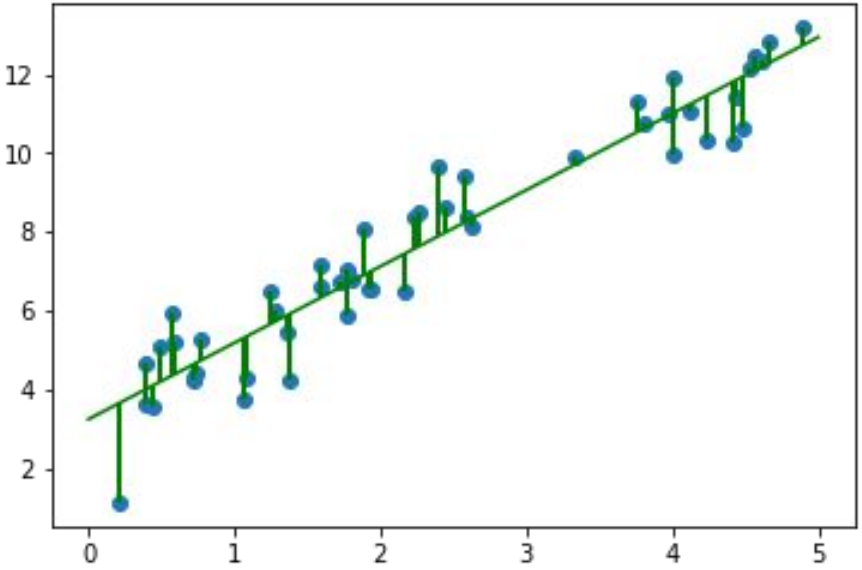
The total squared distance between observed y-values and “predicted” y-values.

$$R^2 = 1 - \boxed{\frac{\text{RSS}}{\text{TSS}}}$$

If RSS is small compared to TSS, this ratio is closer to 0, and we get a value of R^2 closer to 1. This corresponds to a “better” fit line.

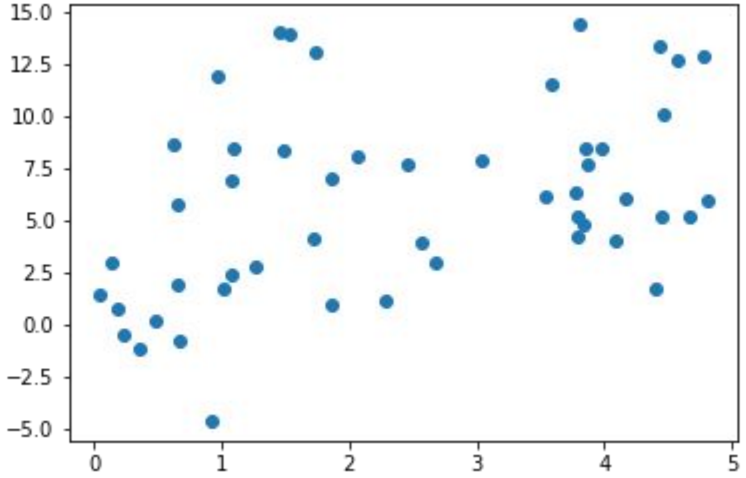


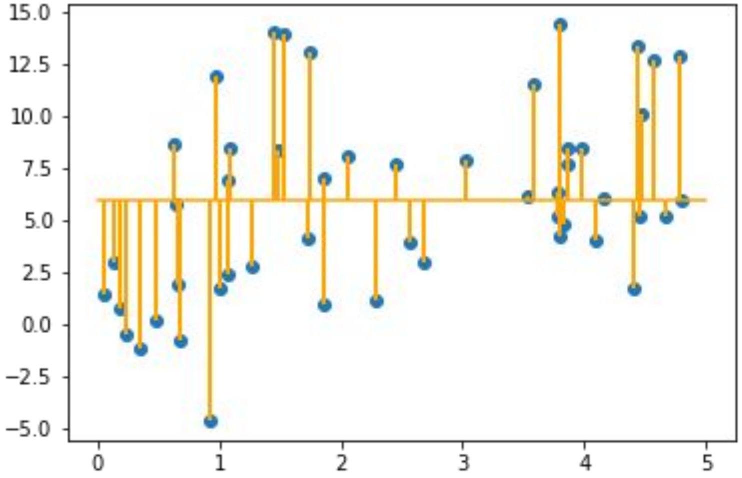
TSS



RSS

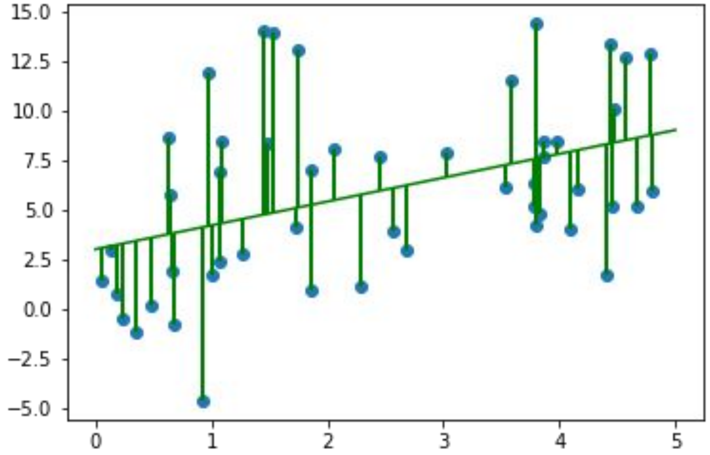
$$R^2 = 0.912$$





TSS

Data



$$R^2 = 0.171$$

RSS