

L03: Correlation

"If you are a User, then everything you've done so far has been according to a plan, right?"

Tron, Tron

Things we cover in this session

- Correlation between data samples
- Visualization of correlations between many variables

Things you need for this session

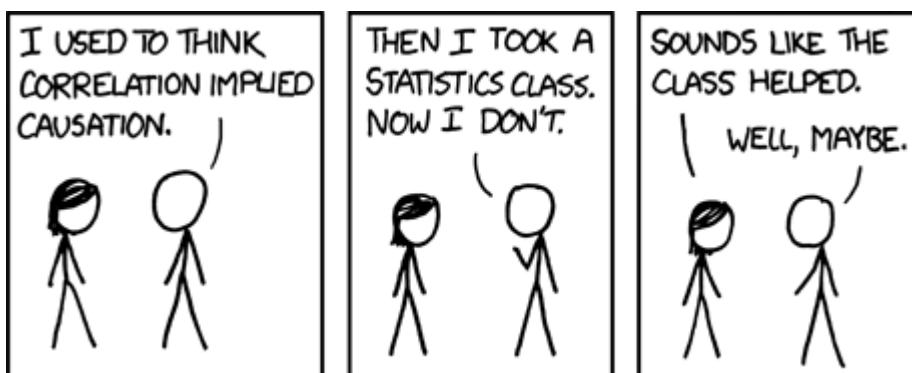
- [W04-1 Regressions](#)

Things to take home from this session

At the end of this session you should be able to

- correlate two variables
- make a correlation plot which shows the pairwise correlation of many variables

Correlation



Correlation measures the deviation of two (or more) variables from independence. Using correlation coefficients, this deviation can be estimated for random samples. While a variety of correlation coefficients exist, only two should be introduced in this context:

- Pearson's correlation coefficient - it is the mother of all correlation measurements and measures the degree of linear relationship of variable pairs.
- Spearman's correlation coefficient - it is a non-parametric measures of the degree of monotonically increasing variable pairs and does not assume any value distribution or linear dependency.
- Kendall's correlation coefficient - it is a non-parametric measure of the relationship between ranks of variable pairs.

One standard function used in R for computing the correlation is the `cor()` function which provides all three coefficient options. A minimalistic call of the function looks like

```
cor(x, y)
```

and returns pearson's correlation coefficient. If there are missing values in `x` and/or `y`, you have to handle them by using the attribute "use" of the `cor()` function (e.g. `use = "complete.obs"`). The `cor()` function can also be applied to entire data frames.

For more information, have a look on the help of the function by typing `?cor` inside your R environment or visit Quick-R's page on [correlation](#).

If you are interested in confidence intervals which give you an idea if the correlation coefficient is actually significant, try the `cor.test()` function which works analogously but can only be applied to two variables at a time (and not an entire data frame). See `?cor.test` for more information.

Time for practice

[W03-1: Correlation](#)

From: <http://bisfogo.environmentalinformatics-marburg.de/> - **BIS-Fogo**

Permanent link: <http://bisfogo.environmentalinformatics-marburg.de/doku.php?id=en:learning:schools:s01:lecture-notes:ba-ln-03>

Last update: **2017/10/30 09:20**

