Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 9. Association between variables. (31 January)
Alla Tambovtseva

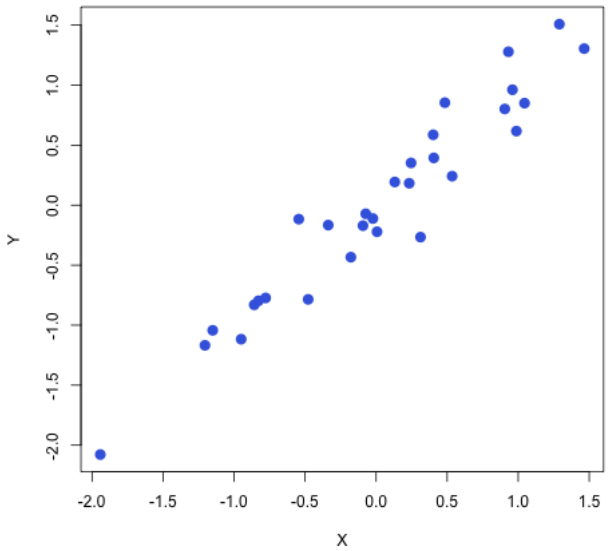
How to measure association between variables?

- Qualitative variables: chi-squared test.
- Quantitative variables (including ordinal scale): correlation coefficients.

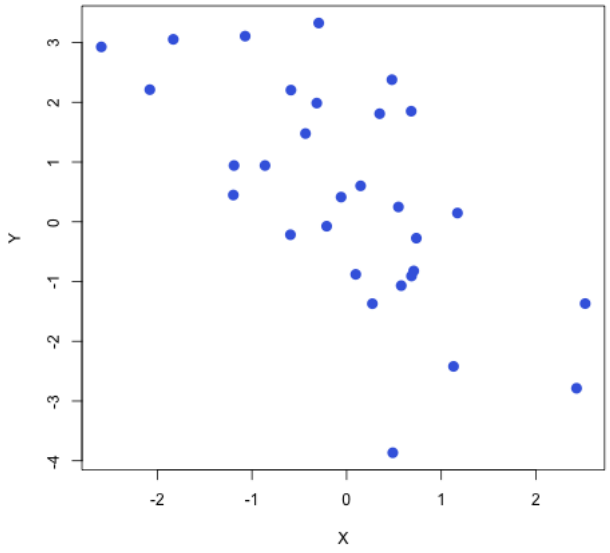
Scatterplots

A **scatterplot** is a graph that shows the relationship between two (sometimes three) quantitative variables. Scatterplots can be helpful for predicting the association between variables and detecting potential problems like the effect of outliers. Let us consider some examples.

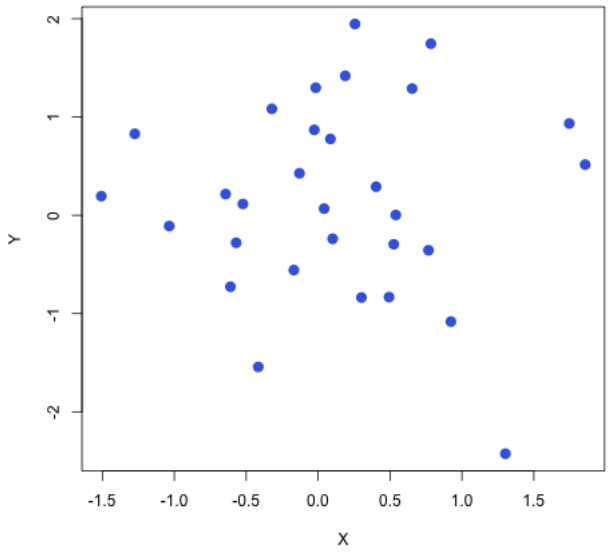
Example 1. Here variables X and Y are positively associated with each other (the higher is X, the higher Y is) and the association is strong (points are concentrated along an imaginary line with a positive slope).



Example 2. Here variables X and Y are negatively associated with each other (the higher is X, the lower Y is) and the association is moderate (points are dispersed around an imaginary line with a negative slope).



Example 3. Here variables X and Y are not associated (as X increases, Y does not change in a certain way, it fluctuates around an imaginary line with a zero slope). Or, at least, the association is extremely small.



Correlation coefficients

A Pearson's correlation coefficient (r)

- Detects a linear relationship between two variables.
- Usually used when both variables are measured in a ratio or in an interval scale.
- Non-robust, so sensitive to outliers in data.

Like any correlation coefficient, it takes values from -1 to 1.

How to see the direction of the association?

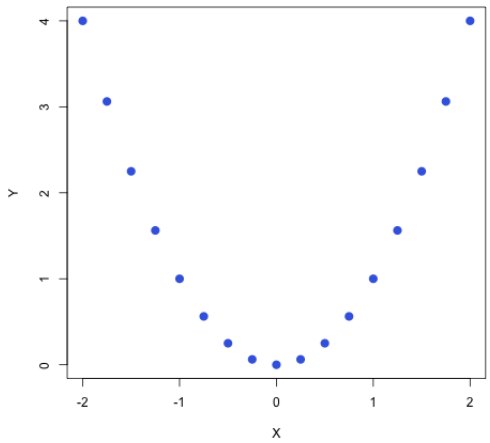
- if $r < 0$, association is negative;
- if $r > 0$, association is positive;
- if $r = 0$, no association.

How to find the strength of the association?

- $|r| < 0.3$, association is weak;
- $0.3 \leqslant |r| < 0.7$, association is moderate;
- $|r| \geqslant 0.7$, association is strong.

Note: the limits for evaluating the strength of the association above are approximate; there is no strict rule that is used for interpreting correlation coefficients.
A Pearson's correlation coefficient detects a linear relationship between variables. So, if it equals zero, it does not mean that there are no relationship at all, it might be non-linear, for example, quadratic. Consider an example below.

Example 4. A Pearson's correlation coefficient (a theoretical one, here X and Y are considered as populations or random variables) between X and Y is zero, but it is clear that X and Y are related, $Y = X^2$.



How to test whether the association between variables is statistically significant?

Test the corresponding null hypothesis about a correlation coefficient.

$$H_0 : R = 0 \text{ (there is no linear relationship between two variables)}^1$$

$$H_1 : R \neq 0 \text{ (there is a linear relationship between variables)}$$

So, if we reject the null hypothesis, it means that two variables are associated and this association is statistically significant.

A Spearman's correlation coefficient (ρ)

- A measure of a rank correlation (ranks of the observations are considered[2]).
- Usually used when at least one variable is measured in an ordinal scale.
- Robust, so less sensitive to outliers in data.

All things related to interpretation of values that were considered above are applicable to a Spearman's correlation coefficient as well. However, to be more precise, a null hypothesis is formulated in a slightly different way:

$$H_0 : \text{ there is no monotonic relationship between variables}$$

$$H_1 : \text{ there is a monotonic relationship between variables}$$

So, a Spearman's correlation coefficient detects any monotonic relationship between variables, not necessarily a linear one.

[1] Here R is capital since we want it to be different from r. R is a correlation coefficient between two populations that we estimate calculating a sample correlation coefficient r.

[2] A rank is a number of an observation in a sample sorted in a ascending order