Correlation

images/corr-globalwarming.png

Statistical Reasoning
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 8

Outline

Today is almost entirely scatterplots and correlation coefficients. We'll make fun of "correlation looks like causation" issues in due time.

images/hamster-contraception.png

Getting there

- **Description**, through univariate statistics, establishes the distributional characteristics of your data.
  - Measures of central tendency and spread
  - Normality and variable transformations
- **Association**, through significance tests, establishes the basic framework of comparison between variables.
  - Comparison of groups (Chi-squared test)
  - Comparison of means (t-test) or proportions

Correlation is different in that it provides the strength and directionality of bivariate relationships:

- Correlation comes with a **correlation coefficient** that indicates how strong the correlation is.
- Correlation also comes with a **significance test** to indicate whether H_0 (no relationship) can be rejected.

Getting through

- **Visualize** bivariate relationships with scatterplots:
  - sc generates individual scatterplots
  - gr mat generates a scatterplot matrix
- **Assess** the the strength of visual relationships:
  - pwcorr provides a correlation matrix of 2+ variables
  - corr also computes a correlation coefficient but pwcorr also adjusts for missing data through pairwise case deletion (see h corr for help).
- **Interpret**
  - **Use a standard vocabulary:** "strong—weak" for strength, "positive/negative" for direction.
  - **Non-linear relationships can produce significant linear correlations.** If you reduce a non-linear pattern to a linear one (due to methodological constraints), mention it in your analysis.
  - **Do not assume causation.** You need theoretical grounds to support a correlation, however significant it is.

Getting the math

- **Pearson's r** determines the quality of a correlation:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \times \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- **Reading guide:**
  - Pearson's r ranges from -1 to $+1$ and comes with a p-value.
  - -1 and $+1$ denote perfect negative and positive correlation.
  - The p-value for Pearson's r tests $H_0 : r = 0$ (no correlation).
- **Computation:**
  - $X_i - \bar{X}$ is the distance (or **residual**) between all observations $i_1, i_2, ..., i_n$ for variable X and the mean $\bar{X}$ of its distribution.
  - Pearson's r computes the **residual sum of squares** (RSS) and its product for variables X and Y.

Perfect positive/negative correlations

images/corr1.pdf

Significant (moderate–strong) correlations

images/corr2.pdf

Insignificant (weak/non-linear) correlations

images/corr3.pdf

Issues

The main concern with correlation is whether you set it right in the first place: **what are your correlating, and why?**

The next slides are from Richard Florida's "The Geography of Hate", *The Atlantic*, May 2011.

What do you **observe**?
What do you **infer**?
What do you **posit**?

images/hate-map.png

images/hate-sc1.png images/hate-sc2.png images/hate-sc3.png images/hate-sc4.png images/hate-sc5.png

Inference

The main concern with correlation is **inference**: not everything that correlates is causally related at the right level of observation.

Your interpretive skills are put once more to the test.

What do you **observe**?
What do you **infer**?
What do you **posit**?

images/ecological-fallacy.jpg