Medical Statistics, Informatics, and Telemedicine

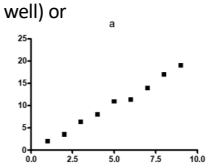
Lecture #7
Correlation. Simple Linear Regression
22nd October 2021

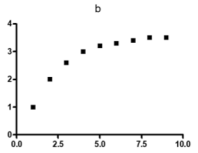
Gergely AGÓCS

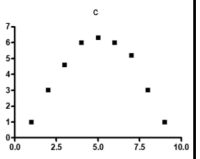
Relationship Between Two Variables

The type of **relationship** may be:

- **monotonous**:
  - **positive** (if *x* increases, *y* increases as well) or **negative** (if *x* increases, *y* decreases)
  - **linear** or **non-linear**, e.g.:
    - exponential: $y \sim e^x$,
    - logarithmic: $y \sim \log(x)$,
    - power: $y \sim x^a$
- **non-monotonous**:
  - **parabolic**
  - **periodic** (sinusoidal)





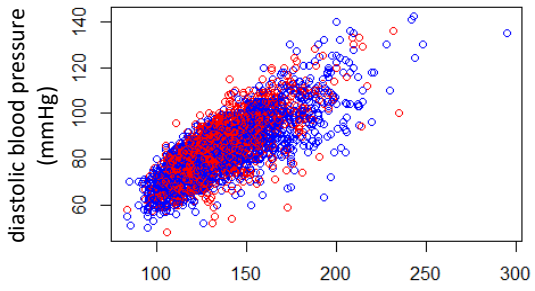


no relationship

(how would the plot look like?)

Correlation

– monotonous,
– symmetrical (cannot be told, which depends on which) relationship between **two random** (containing random error, not „set") variables.



(which depends on which?)

Correlation

How to express the **strength** of a correlation?

Correlation coefficients (c. c.):
supposing **linear** correlation: **Pearson's c. c.** (*r*)
supposing **monotonous** correlation: **Spearman's rank c. c.** (*ρ*),

The **value** of the correlation coefficient:

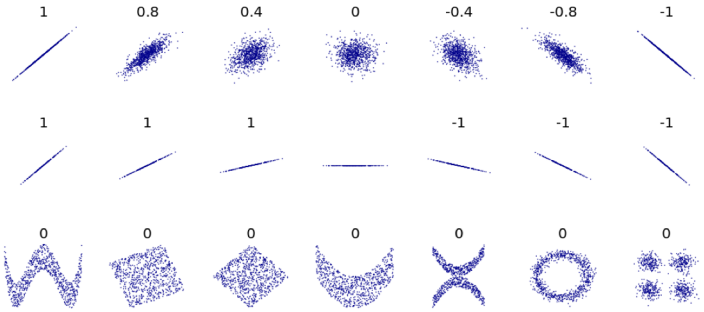
between –1 and +1
negative: negative correlation
positive: positive correlation
the further from 0, the stronger the correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

„distance from the middle" – both in the *y* and *x* directions

Remarks

- Always **plot** your data!
- Correlation does not imply causation
  e.g.: http://www.fastcodesign.com/3030529/infographic-of-the-day/hilarious-graphs-prove-that-correlation-isnt-causation



Pearson's correlation coefficient for different data sets.

Student's t-Test for the Significance of the Correlation Coefficient (or the Slope)

What am I interested in:
Are 2 variables (linearly) correlated (is r different from 0)?

Type of variables:
2 numerical variables (x and y)

Conditions:
– Independent observations (x and y pairs)
– We suppose a symmetrical, linear relationship
– Both x and y are random variables

Remarks:
We are testing the H_0: **correlation coefficent = 0**.

6

Regression

Function-like relationship (NON symmetric) between a dependent (outcome, eredmény, y) and an independent (explanatory, predictive, x) variable(s). [Y is a random variable, X not necessarily]

y* depends on *x – the direction of causality is **clinically** substantiated, but cannot be investigated stratistically.

Related:
- Is there a (given type of) function-like relationship? (statistical, NOT causal)
- What is the value of y for a given x? (estimation)
- What is the value of x for a given y? (estimation)
- Which function describes the dependence of y on x the best?

7

Linear Regression

We suppose a **linear relationship**.

In case of 2 variables the questions and calculations of regression and correlation may in most cases be made „equivalent" to each other.

8

Linear Regression

For the estimation of the linear: OLS (Ordinary Least Square method)

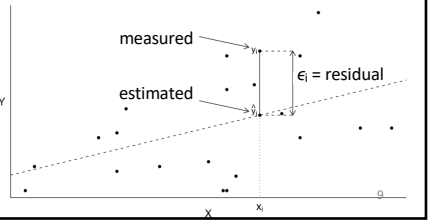
y axis intercept

slope

Linear function: $Y = \beta_0 + \beta_1 * X + \epsilon$

error term (residual): the **vertical** distance between the fitted line and the data point (i.e. the difference between the estimated and the measured value)

According to the OLS method the best linear is the one for which the sum of the **squared** line–datapoint vertical distances is the **least**.



9

Student's t-Test for the Significance of the Correlation Coefficient (or the Slope)

What am I interested in:

Does y depend on x linearly?

Type of variables:

2 numerical variables (*x* and *y*)

Conditions:

– Independent observations (*x* and *y* pairs)
– We suppose a linear relationship
– *x* values may be measured „ithout error" (they are set values)
– the distribution of residuals is
  – normal for every x
  – with the same variance

Remarks:

We are testing the H_0: **slope = 0**.

10

Slope and R^2

Slope

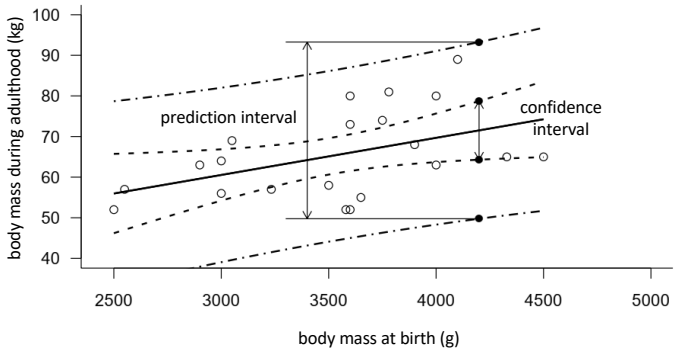
the average change of *y* corresponding to a unit change of *x*

R^2 – coefficient of determination

– the square of *r*
– what percentage of the variation (variance) of the *y* variable may be attributed to the variation (variance) of the *x* variable

11

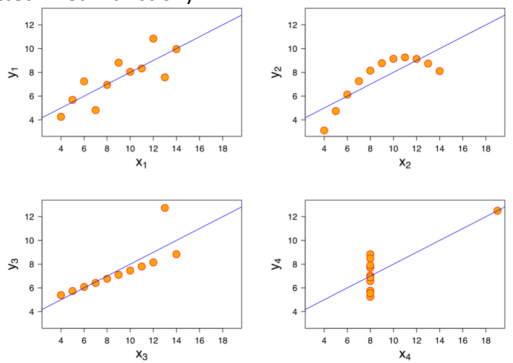
Confidence and Prediction Intervals



12

Anscombe's Quartet (1973)

– very different graphical appearance
– identical paramters (means, SDs, correlation coefficient, equation of the fitted linear function)



13