Correlation

- describes to what extent two variables are related
- does not immediately mean causality
  - e.g. correlation between shark attacks and ice cream sales
    - shark attacks do not cause ice cream sales
    - ice cream sales do not cause ice cream sales
    - both are caused by third variable summer/heat

Pearson Correlation

- correlation between two quantitive variables

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

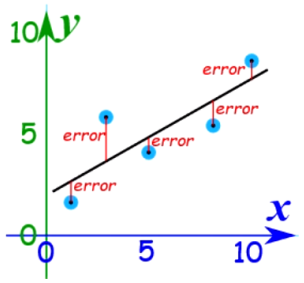
- $\overline{X}, \overline{Y}$ = average of X,Y respectively
- correlation coefficient r
  - describes strength of relation ship between X and Y
  - r=-1:
    - perfect descending linear relationship
    - high X <==> low Y
  - 0:
    - variables not systematically related
    - high X <==> high or low Y
  - 1:
    - perfect ascending linear relationship
    - high X <==> high Y
- correlation threshold
  - threshold depends on domain
  - different for each use case
- large population but small sample size
  - likelihood of correlation within subset
  - even though no correlation within whole population
  - null hypothesis
    - We want this likelihood to be small! Typical threshold values for “small enough” are p<0.05, p<0.01, p<0.005

Linear Regression

- approximates linear function between linearly correlated data
  - $y = a + bx$
- underlying assumption
  - never know all data
  - we just have training data
  - keep part of data for testing afterwards
- optimisation criterion
  - method of least squares
  - see [[NRLA]] script
  - Optimisation criterion: Minimal least squares error –minimal sum of distances (in whichever direction) of points to line.
  - Regression line: $y = a + bx$

    Slope $b = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$

    OR $b = r\left(\frac{s_y}{s_x}\right)$ with r the correlation coefficient

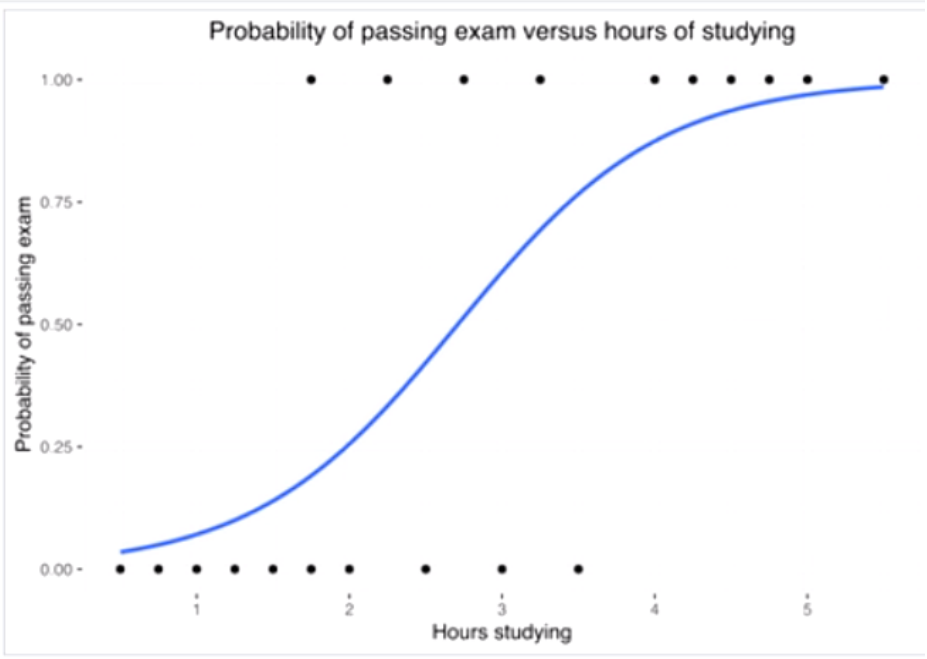
  - Intercept $a = \frac{\sum y - b\sum x}{N}$

    This method minimizes the sum of the squared errors.

    Error = difference between the estimated y-value (y=a+bx) for a given x value, and the real/measured y-value in the sample data for the same x value.

Other Types of Regression

- non-linear regression - curve fitting
  - fitting non-linear function to data



Example graph of a logistic regression curve fitted to data. The curve shows the probability of passing an exam (binary dependent variable) versus hours studying (scalar independent variable). See § Example for worked details.

- logistic regression
  - fitting log function to continous independent data
  - and dichotomous (zweigeteilt) out come data
  - classification method

Prediction with Correlation and Regression

- estimate value y_i, given x_i using regression line
  - y_i dependent outcome variable
  - x_i independent input variable

[[Machine Learning]]