MUNI
PHARM

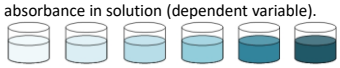
# Statistical methods

Biophysics

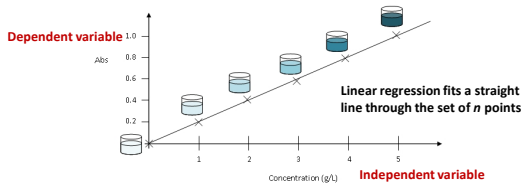
1

## Linear regression

- In experiments we are looking for dependences between two variables (measured and set variable).
- An example can be influence of concentration (independent variable) to absorbance in solution (dependent variable).



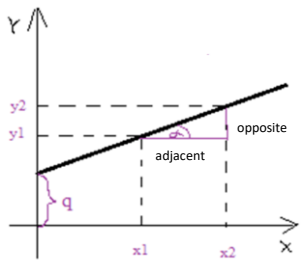
- Absorbance is a function of concentration, A= f (c)




2

## Data dependencies
*Equation of the line*

$$y = k \times x + q$$




$q$ =constant=shift on $y$ axis (intercept)

$k$ = slope= tg α=

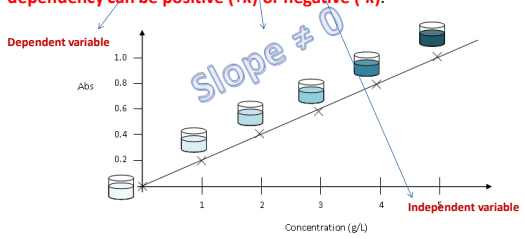
$(y_2-y_1)/(x_2-x_1)$


3

## Data dependencies

$$A = k \times c + 0$$

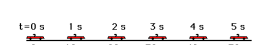
In the case of the absorbance, the line goes through the zero point, it means zero concentration = zero absorbance, therefore **intercept** $q$, or shift of the $y$ axis is 0. **Slope** $k$, describes the rate of change between the independent and dependent variables - **data dependency can be positive (+*k*) or negative (-*k*)**.

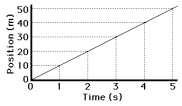



5

## Equation of the line

**Sample Problem**: Consider a car moving with a constant velocity of 10 m/s for 5 seconds.




**What is the meaning of the slope *k*?**

**$k$ = slope= $(y_2-y_1)/(x_2-x_1)$**

**The slope of the line is 10 meter/1 second = the slope of the line (10 m/s) is equal to the velocity of the car.**

The slope expresses the rate of the examined process (e.g., in chemical kinetics, the slope represents the chemical reaction rate).

7

## Method of Least Squares

The purpose of regression analysis is to analyze relationships among variables. We are trying to replace

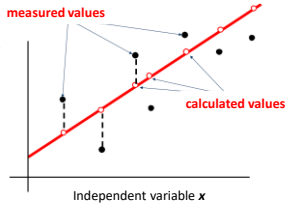
**each measured (experimental) value of the dependent variable $y_{exp}$**

**by value calculated (predicted) $y_{pred}$**

Dependent variable $y$

Linear regression makes the sum of vertical distances between the points of the data set and the fitted line as small as possible – **Method of Least Squares**

$\Sigma (y_{exp} - y_{pred})^2$ = min.

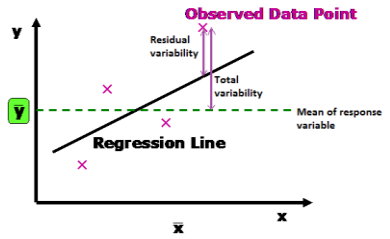



9

## Coefficient of determination $R^2$

Fitting data by regression line is expressed by coefficient of determination **$R^2$ (values from 0 to 1)**.

Is described by $R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

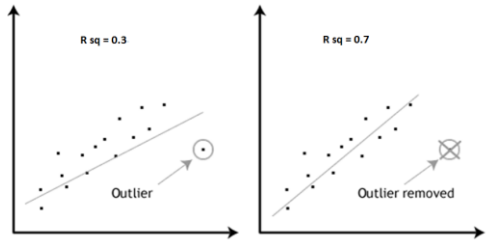



10

## Influence of outliers

Quality of the linear regression (i.e. dependencies between variables) can be significantly influenced by outliers, which is reflected by a decrease in the coefficient of determination.

**Outliers should be removed from data set**.

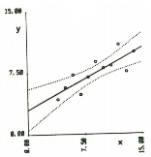



11

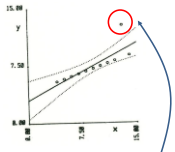
## Regression model choice

**Only $R^2$ can not be used to assess the quality of linear regression**.

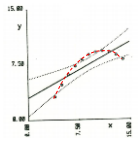
In these models is identical **$R^2$ = 0.66**



Linear regression is a suitable model



Linear regression is a suitable model but data contain outlier. We have to remove the outlier, then $R^2$ will be 0.99



Linear regression is not a suitable model. If a more suitable model is used, $R^2$ will increase (non-linear regression, $R^2$ = 0.99)

12

## Types of regression models

Examples of linear regression models:

| | |
|---|---|
| $y = q + k*x$ | - line |
| $y = q + k_1*x + k_2*x^2$ | - parabola |
| $y = q + (k/x)$ | - hyperbola |

Linear models can be models, whose graphical representation is not the line.

Examples of nonlinear regression models :

$y = q * x^k$

$y = q * e^{k*x}$

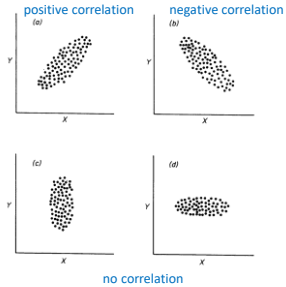
$y = q * e^{k/x}$

- They are able to model complex real processes, eg. kinetics of reactions, drug dissolution and absorption, drug elimination by the organism....
- More complicated calculation. Sometimes it is possible to transform (mathematically) nonlinear regression to linear regression, eg. first order kinetics (nonlinear r.):
  $C(t) = C(0) * e^{-kt}$ => linear r. **$\ln C(t) = \ln C(0) - k * t$**
  **$y = q - k * x$**

16

## Correlation

There are two variables (x and y) and we are asking whether they are independent, and if they are dependent (correlated), how much.




Output of correlation analysis is the correlation coefficient.

The correlation coefficient R value can range from -1 to +1.

18

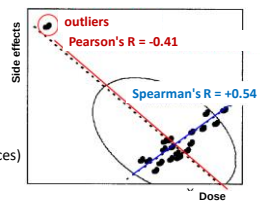
## Correlation coefficient

Pearson's correlation coefficient

- dimensionless measure of linear correlation
- is 0 – **1 for the positive correlation** or 0 – **(-1) for the negative correlation**
- the correlation coefficient is the same for the dependence of $x_1$ to $x_2$ or the dependence of $x_2$ to $x_1$

Spearman's correlation coefficient

- is based on the ordinal values for each variable
- reduce the influence of outliers
- limited use (usually in the natural sciences)




**Sample problem:** Drug dose (mg) versus side effects (%). The incidence of side effects increases with a dose, but allergic patients have strong side effects at a very small dose. From the data set, we can not exclude allergies (they are part of the population).

19

## Correlation coefficient

Critical values of the correlation coefficient (for different sample sizes) are tabulated. If the calculated correlation coefficient is greater than the tabulated values, the correlation is statistically significant.

| n | 5 | 6 | 7 | 8 | 9 | 10 | 15 | 20 | 30 | 50 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.05$ | 0.878 | 0.811 | 0.755 | 0.707 | 0.666 | 0.632 | 0.514 | 0.444 | 0.361 | 0.279 | 0.196 |
| $\alpha = 0.01$ | 0.959 | 0.917 | 0.875 | 0.834 | 0.798 | 0.765 | 0.641 | 0.561 | 0.463 | 0.363 | 0.254 |

$\alpha$ - significance level, n – number of points

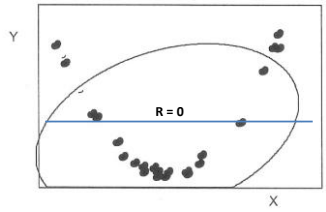
Typically $\alpha$ = 0.05; there is a 5% or less chance that our results is not true.

20

## Correlation

The correlation coefficient is used to express the „**straight-line fit**". Correlation analysis describes the linear relationships between variables.

LOW correlation coefficients DOES NOT MEAN INDEPENDENCE!

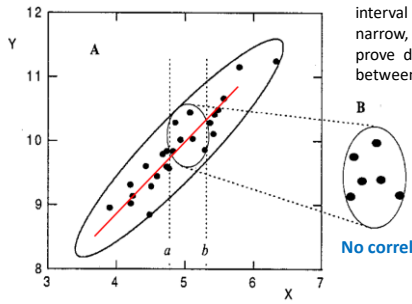



21

## The influence of range on R

The interval of the measured values ***x*** and ***y*** must be sufficiently wide. If the measured interval is too narrow, it may not prove dependency between data.

**Positive correlation**



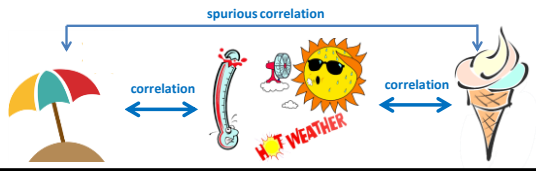

**No correlation**

22

## Spurious correlation

**Correlation don't expresses cause and effect!!!**

Sample problem:
The ice cream sales are mutually correlated with the number of beach umbrellas sales. In fact, the sales of ice cream don't affects the sales of beach umbrellas = spurious correlation.
**Spurious correlation** occurs when the other unobserved or not analyzed variable/s affects both variables that we study.



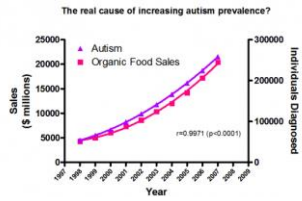

24

## Spurious correlation

**Correlation don't expresses cause and effect!!!**

**Spurious correlation** occurs when the other unobserved or not analyzed variable/s affects both variables that we study.

Sample Problem: Factors related to the increasing incidence of autism




Higher incidence of autism: better diagnosis, broader criteria for classification as autistic, changes in diagnosis due to better diagnosis (some autistic people were previously included in the group of mentally retarded).

25