

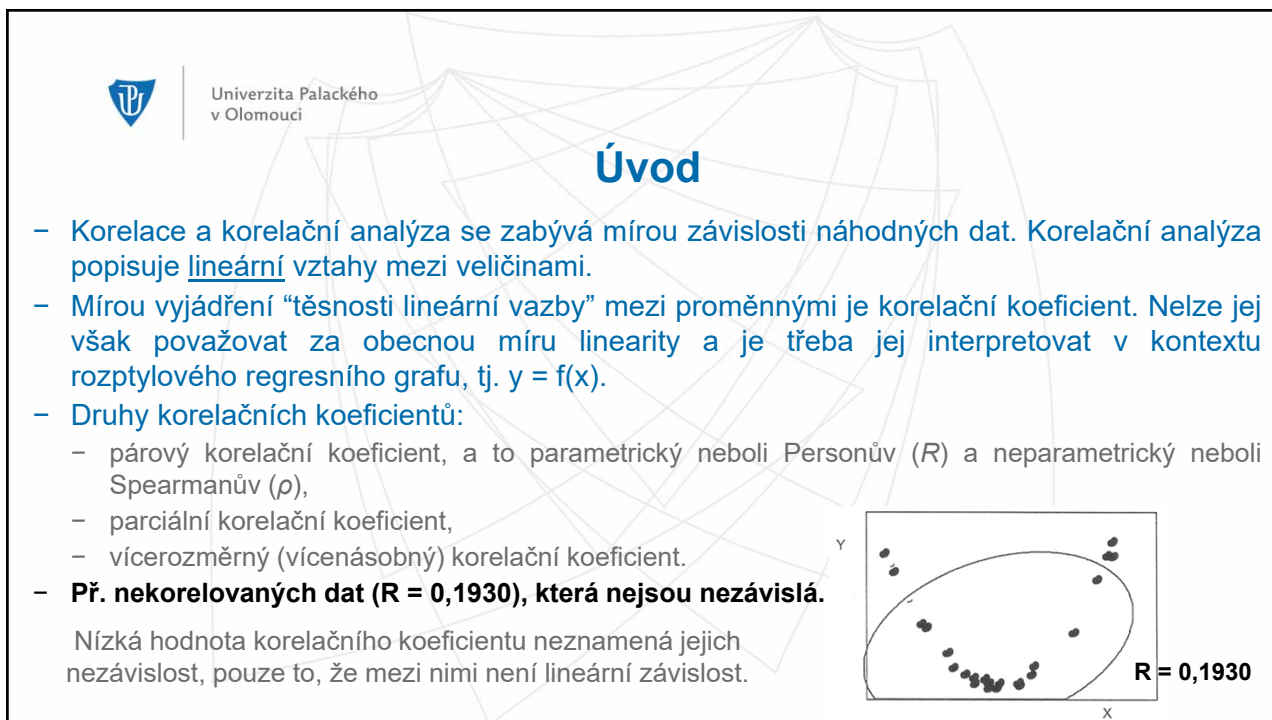
Univerzita Palackého
v Olomouci

Korelace

Chemometrie I (ACH/CHEX1)

(c) David MILDE, 2024

1

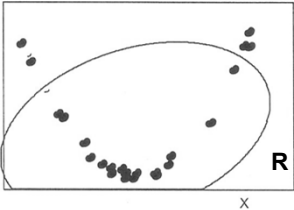


Univerzita Palackého
v Olomouci

Úvod

- Korelace a korelační analýza se zabývá mírou závislosti náhodných dat. Korelační analýza popisuje lineární vztahy mezi veličinami.
- Mírou vyjádření "těsnosti lineární vazby" mezi proměnnými je korelační koeficient. Nelze jej však považovat za obecnou míru linearitu a je třeba jej interpretovat v kontextu rozptylového regresního grafu, tj. $y = f(x)$.
- Druhy korelačních koeficientů:
 - párový korelační koeficient, a to parametrický neboli Personův (R) a neparametrický neboli Spearmanův (ρ),
 - parciální korelační koeficient,
 - vícerozměrný (vícenásobný) korelační koeficient.
- **Př. nekorelovaných dat ($R = 0,1930$), která nejsou nezávislá.**

Nízká hodnota korelačního koeficientu neznamená jejich nezávislost, pouze to, že mezi nimi není lineární závislost.



2



Univerzita Palackého
v Olomouci

Párový korelační koeficient

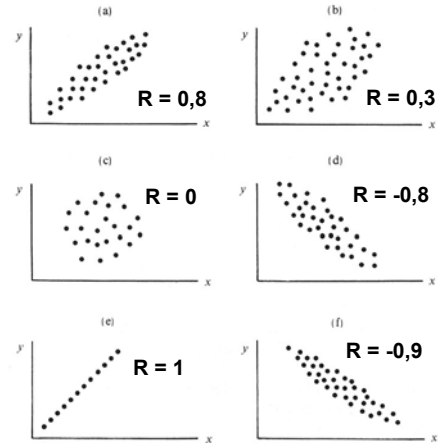
– Pearsonův párový korelační koeficient R :

$$R = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

- Může nabývat hodnot od -1 do +1.
- Je-li $R \approx 0$, proměnné jsou lineárně nekorelované.
- Je-li R blízko $\pm 0,5$, slabá korelace (lineární závislost).
- Je-li $R \approx +1$, vysoká přímá lineární závislost.
- Je-li $R \approx -1$, vysoká nepřímá lineární závislost.

– Párový korelační koeficient se používá v jednoduché LR pro vyjádření korelace mezi x a y . Dále se používají ve vícerozměrné LR a to pro vyjádření korelace mezi:

1. jednotlivými závisle proměnnými x_i a nezávisle proměnnou y , tedy např.: x_2 a y , či x_3 a y ,
2. vzájemně mezi závislými proměnnými, např.: x_1 a x_2 , či x_1 a x_3 .



3



Univerzita Palackého
v Olomouci

Párový korelační koeficient

– Koeficient determinace R^2 popisuje stupeň příčinné závislosti mezi x a y , nabývá hodnot od 0 do +1.

- Např. pro $R = 0,7$ je $R^2 = 0,49$ a to znamená, že pouze 49 % variability mezi proměnnými x a y se dá vysvětlit jejich lineárním vztahem.

– Test významnosti korelačního koeficientu:

$$H_0: R = 0$$

$$(1.) H_1: R \neq 0$$

$$(2.) H_1: R > 0$$

$$(3.) H_1: R < 0$$

$$t = \frac{R \cdot \sqrt{n-2}}{\sqrt{1-R^2}}$$

- testační statistiku t porovnáváme s t_{krit} . Oblast zamítnutí H_0 je:

$$(1.) |t| > t_{krit(1-\alpha/2, n-2)}$$

$$(2.) |t| > t_{krit(1-\alpha, n-2)}$$

$$(3.) t < -t_{krit(1-\alpha, n-2)}$$

4



Univerzita Palackého
v Olomouci

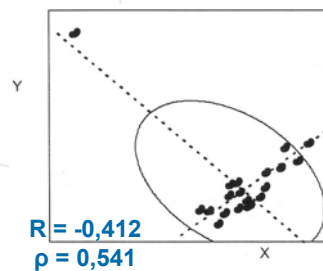
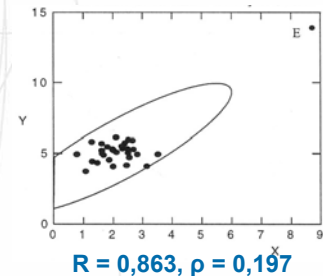
Párový korelační koeficient

- Spearmanův korelační koeficient ρ – pořadový korelační koeficient, je robustní. Jeho vztah k Pearsonovu korelačnímu koeficientu je analogií vztahu mediánu k aritmetickému průměru.

$$\rho = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^n (P_{1i} - P_{2i})^2$$

P_{1i}, P_{2i} jsou pořadová čísla jednotlivých hodnot

- Spearmanův korelační koeficient není ovlivněn výskytem extrémních bodů, které naopak ovlivňují hodnotu Pearsonova R . Při korelační analýze nestačí pouze vypočítat R a testovat jeho významnost. Je třeba posuzovat i rozptylový regresní graf.
- Je vhodné sledovat rozdíl mezi R a ρ . Velký rozdíl naznačuje přítomnost extrémního bodu či extrémních bodů.



5



Univerzita Palackého
v Olomouci

Další korelační koeficienty

- **Parciální korelační koeficient:**
 - Popisuje vztah mezi dvěma proměnnými (např. x_i a x_j) při zkonstantnění dalších proměnných.
 - Vyjadřuje tedy vzájemnou korelaci s vyloučením vlivu všech ostatních proměnných, tj. nezkrácenou vlivem závislostí mezi ostatními proměnnými.
 - Má smysl jen v případě více než 2 proměnných.
 - Vzorec pro parciálního korelačního koeficientu (1. řádu) v případě 3 proměnných:

$$R_{1,2(3)} = \frac{R_{12} - R_{13}R_{23}}{\sqrt{(1 - R_{13}^2)(1 - R_{23}^2)}}$$

$R_{1,2(3)}$ je parciální korelační koeficient mezi proměnnými 1 a 2, je-li proměnná 3 konstantní
 R_{ij} jsou párové korelační koeficient mezi proměnnými i a j

- **Vícenásobný korelační koeficient:**
 - Vyjadřuje míru lineární závislosti zvolené proměnné na všech ostatních dohromady.
 - Tyto koeficienty jsou vyšší než největší odpovídající párový koeficient a vždy rostou s počtem proměnných, i když jsou párové korelace nevýznamné a obvykle nadhodnocují skutečnost.
 - Používá se u vícerozměrné (vícenásobné) lineární regrese.

6