# Analýza dat v neurologii

## LXXVI. Korelační analýza vícerozměrných souborů kvantitativních a kvalitativních dat – představení vybraných ukazatelů

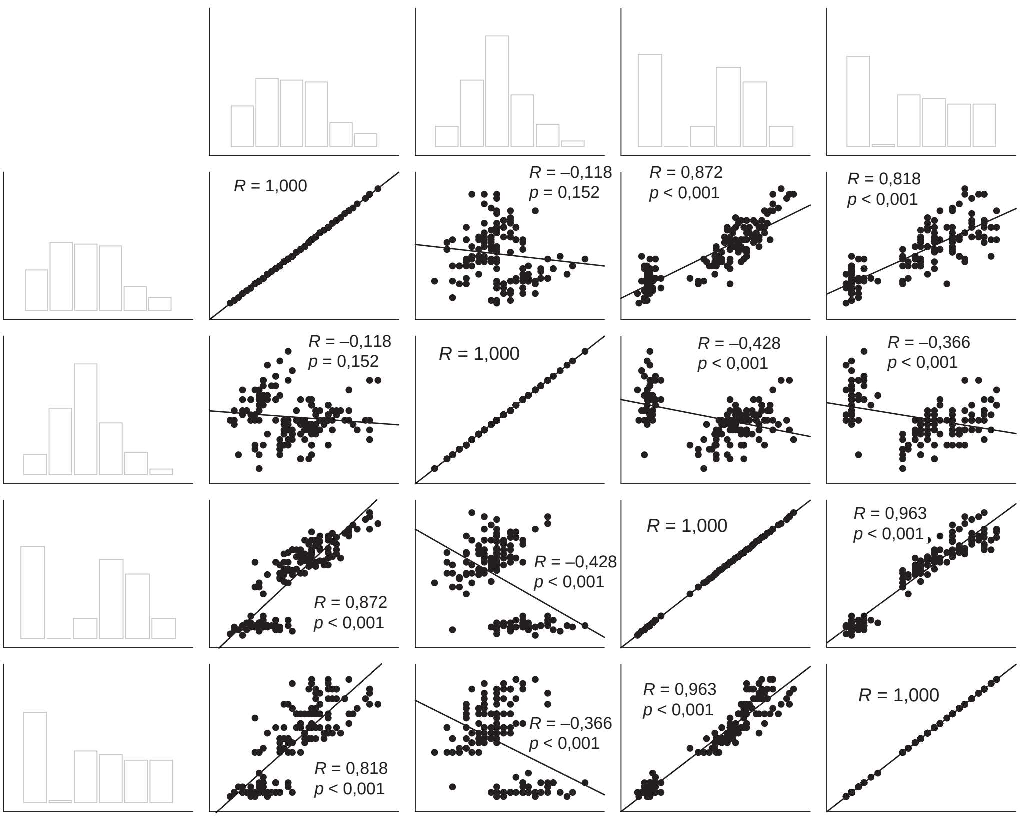
L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková
Institut biostatistiky a analýz,
LF MU, Brno

prof. RNDr. Ladislav Dušek, Ph.D.
Institut biostatistiky a analýz,
LF MU, Brno
e-mail: dusek@iba.muni.cz

Tímto dílem vstupujeme do závěrečné části výkladu různých aspektů korelační analýzy. V několika předchozích dílech jsme tuto analýzu představili z různých pohledů jako nástroj pro studium síly vztahu dvou kvantitativních proměnných, představili jsme parametrické i neparametrické korelační koeficienty a vysvětlili principy hodnocení jejich statistické významnosti. Avšak svět klinického a biomedicínského výzkumu většinou nepracuje pouze se dvěma charakteristikami zkoumaných subjektů. Typickým výstupem probíhajících měření jsou tzv. mnohorozměrné (vícerozměrné) soubory dat, kdy je $N$ jedinců popisováno $K$ proměnnými a zápis datového souboru vytváří matici $N \times K$. S rozšiřujícím se arzenálem různých vyšetřovacích metod a zejména s nástupem molekulárně biologických a genetických vyšetření se tento trend týká i klasického klinického výzkumu a výsledné datové matice zahrnují i mnoho desítek proměnných. Logicky vzniká potřeba vyhodnotit vzájemnou korelaci všech těchto proměnných.

Problémem korelační analýzy mnohorozměrných souborů může být již samotný vysoký počet vzájemných korelací sledovaných proměnných. Ty je nutné nějak přehledně znázornit a dále s nimi efektivně pracovat.

Pro hodnocení vzájemného vztahu více spojitých proměnných je využívána matice korelačních koeficientů. Jde o čtvercovou matici, jejíž buňky obsahují korelační koeficienty příslušných dvojic proměnných. Matici lze prezentovat i graficky v tzv. korelogramu, jak dokládá ukázka níže.




Příklad 1. Znázornění korelační matice v tzv. korelogramu.

Např. je třídit podle síly a významnosti nebo seskupovat proměnné do skupin podle toho, jaký mezi sebou mají vztah. Jistou pomocí zde jsou tzv. korelační matice. Při současném zpracování *K* proměnných hodnotíme korelaci pro K*(K – 1)/2 dvojic proměnných, které sestavujeme do tzv. korelační matice, jejíž řádky i sloupce jsou věnovány postupně první až *K*-té proměnné. Na průsečíku *i*-tého řádku a *j*-tého sloupce je uvedena korelace *i*-té a *j*-té proměnné. Korelační matice jsou tak logicky čtvercové (symetrické podle hlavní diagonály). Na hlavní diagonále korelační matice najdeme vždy hodnoty 1, neboť platí, že korelace proměnné *X* se sebou samou musí být absolutní, a tedy platí vztah *cor(X, X)* = 1.

Samotné vytvoření korelační matice sice částečně zpřehlední větší množství korelací, ale při jejich velkém počtu ani to není konečným řešením. Proto se korelační matice graficky znázorňují v tzv. korelogramu *(correlogram)*, což není nic jiného než vykreslené vzájemné korelace dvojic proměnných v celkovém grafu. Tento typ grafického znázornění jsme již popsali v díle 71 tohoto seriálu, pro přehlednost jej zde připomínáme v příkladu 1. Je zřejmé, že jde o poměrně funkční nástroj, který usnadní orientaci i ve velké korelační matici. Dostupnost tohoto typu grafu je velmi dobrá, zvládne jej automaticky vykreslit v podstatě každý software určený k statistickému zpracování dat.

Zde se jistě nabízí otázka, jakou přidanou hodnotu mají tyto mnohonásobné grafy proti „běžné" korelaci dvou proměnných? Odpověď je snadná a spočívá již v důvodu, proč byly tyto proměnné společně sledovány. Pokud u jednoho subjektu, pacienta, máme důvod sledovat současně *K* proměnných, pak nás jistě nezajímají jen jejich separované vzájemné vztahy, ale i odpovědi na následující otázky:

- Jaká je vzájemná provázanost jednotlivých proměnných? Nebo jinými slovy, do jaké míry se sledované proměnné vzájemně nezávisle doplňují a do jaké míry spolu souvisí, například až tak, že by jejich současné sledování bylo redundantní? Pokud by totiž mezi sebou některé proměnné velmi silně korelovaly (korelace blízké hraničním hodnotám –1 nebo +1), pak se vzájemně nahrazují a nepřinášejí novou informaci o sledovaných subjektech.
- A naopak, existuje v sadě sledovaných proměnných nějaká proměnná, která vůbec nekoreluje s ostatními, tedy je na nich nezávislá?
- Lze proměnné ve sledované sadě nějak třídit dle jejich vzájemné korelace? Např. do skupin proměnných, které jsou uvnitř silně vzájemně korelované, avšak nezávislé na jiných skupinách proměnných?
- Existují nějaké významné dílčí korelace mezi proměnnými? Existují některé proměnné, jejichž změny lze vysvětlit korelacemi s jinými proměnnými? Takovou analýzou se dá odhalit například maskující vliv některých vzájemně korelovaných znaků apod.

Takto bychom mohli v otázkách pokračovat dále, neboť vícerozměrná analýza dat reprezentovaných mnoha proměnnými samozřejmě nabízí velké množství pohledů a dílčích analýz. Konkrétním přístupům se proto budeme věnovat v příkladech v dalším díle seriálu. Zde se pokusíme vysvětlit výpočetní základnu pro tyto analýzy. Nejčastěji používanými statistikami v těchto sofistikovaných analýzách jsou tzv. vícenásobné koeficienty korelace a dílčí (parciální) koeficienty korelace. Jejich výpočty doložíme formou příkladů, avšak nejprve se mu-

Determinant čtvercové matice řádu *K* (*K* je počet řádků i počet sloupců matice) je součet všech součinů *K* prvků této matice, přičemž v žádném z kalkulovaných součinů se nesmí vyskytovat dva prvky z téhož řádku ani z téhož sloupce. Každý součin je označen znaménkem dle pravidla, které ukazuje níže uvedený příklad. Pro matici ***A*** označujeme její determinant jako det ***A***.

**1a. Výpočet determinantu čtvercové matice 2 × 2.**

Vztah pro výpočet determinantu matice 2 × 2: $\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}*a_{22} - a_{12}*a_{21}$

Příklad výpočtu determinantu matice 2 × 2, pokud $a_{11} = 2$; $a_{22} = 5$; $a_{12} = 1$; $a_{21} = 9$: $\det \begin{pmatrix} 2 & 1 \\ 9 & 5 \end{pmatrix} = (2 * 5) - (9 * 1) = 1$

**1b. Výpočet determinantu čtvercové matice 3 × 3.**

Tento příklad rovněž znázorňuje pravidlo přidělování znamének jednotlivým dílčím součinům: plné čáry označují součiny kladné a čerchované čáry součiny záporné.

Vztah pro výpočet determinantu matice 3 × 3 pomocí tzv. Sarrusova pravidla:

$$\begin{matrix} a_{11} & a_{12} & a_{13} & a_{11} & a_{12} \\ a_{21} & a_{22} & a_{23} & a_{21} & a_{22} \\ a_{31} & a_{32} & a_{33} & a_{31} & a_{32} \end{matrix}$$

$$\det = \underline{(a_{11}*a_{22}*a_{33} + a_{12}*a_{23}*a_{31} + a_{13}*a_{21}*a_{32})} - (a_{31}*a_{22}*a_{13} + a_{32}*a_{23}*a_{11} + a_{33}*a_{21}*a_{12})$$

Příklad výpočtu determinantu matice 3 × 3, pokud $a_{11} = 1$; $a_{22} = 4$; $a_{33} = 9$; $a_{12} = 2$; $a_{13} = 5$; $a_{23} = 7$; $a_{21} = 3$; $a_{31} = 6$; $a_{32} = 8$

$$\det \begin{pmatrix} 1 & 2 & 5 \\ 3 & 4 & 7 \\ 6 & 8 & 9 \end{pmatrix} = (1*4*9+2*7*6+5*3*8) - (6*4*5+8*7*1+9*3*2) = 120 - 110 = 10$$

Příklad 2. Ukázka výpočtu determinantu matice a jeho výpočet pro matici 2 ×2 a 3 ×3.

Korelační matice obsahuje vzájemné korelační koeficienty dvou nebo více proměnných. Na její hlavní diagonále jsou tak hodnoty rovny 1 (jde o korelaci dané proměnné se sebou samou, což musí být logicky korelace absolutní). Příklad dokumentuje význam hodnoty determinantu korelační matice ve vazbě na hodnoty korelačních koeficientů v ní obsažené. Je zřejmé, že hodnota determinantu reaguje na znaménka korelačních koeficientů a dále, že absolutní hodnota determinantu klesá s rostoucí velikostí korelací v matici.

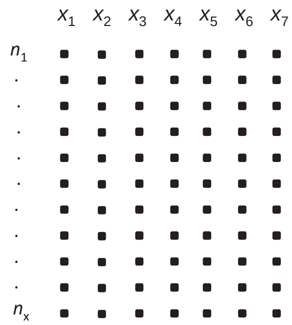
$$\det \begin{pmatrix} 1{,}0 & 0{,}5 & -0{,}7 \\ 0{,}5 & 1{,}0 & 0{,}8 \\ -0{,}7 & 0{,}8 & 1{,}0 \end{pmatrix} = -0{,}94 \qquad \det \begin{pmatrix} 1{,}0 & 0{,}2 & 0{,}1 \\ 0{,}2 & 1{,}0 & 0{,}1 \\ 0{,}1 & 0{,}1 & 1{,}0 \end{pmatrix} = 0{,}944$$

$$\det \begin{pmatrix} 1{,}0 & 0{,}6 & 0{,}3 \\ 0{,}6 & 1{,}0 & 0{,}8 \\ 0{,}3 & 0{,}8 & 1{,}0 \end{pmatrix} = 0{,}198 \qquad \det \begin{pmatrix} 1{,}0 & -0{,}3 & -0{,}6 \\ -0{,}3 & 1{,}0 & -0{,}9 \\ -0{,}6 & -0{,}9 & 1{,}0 \end{pmatrix} = -0{,}584$$

Příklad 3. Příklady determinantů různých korelačních matic.

Výpočet matice korelačních koeficientů a jejího determinantu je prvním krokem k statistické analýze takových vícerozměrných souborů.

**Matice *x* pacientů popsaná 7 proměnnými**




**Výpočet Pearsonova korelačního koeficientu**

**Matice korelačních koeficientů 7 × 7**

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 1{,}0 & -0{,}4 & 0{,}3 & 0{,}0 & 0{,}5 & 0{,}3 & 0{,}0 \\ -0{,}4 & 1{,}0 & 0{,}0 & 0{,}1 & -0{,}4 & -0{,}3 & -0{,}5 \\ 0{,}3 & 0{,}0 & 1{,}0 & -0{,}8 & -0{,}3 & 0{,}3 & -0{,}2 \\ 0{,}0 & 0{,}1 & -0{,}8 & 1{,}0 & 0{,}3 & -0{,}4 & -0{,}3 \\ 0{,}5 & -0{,}4 & -0{,}3 & 0{,}3 & 1{,}0 & 0{,}6 & 0{,}6 \\ 0{,}3 & -0{,}3 & 0{,}3 & -0{,}4 & 0{,}6 & 1{,}0 & 0{,}8 \\ 0{,}0 & -0{,}5 & -0{,}2 & -0{,}3 & 0{,}6 & 0{,}8 & 1{,}0 \end{pmatrix} \Rightarrow \det = -0{,}003723$$

Příklad 4. Korelační matice většího množství proměnných a její determinant.

síme zastavit u pojmu determinant matice, v našem případě půjde o determinant korelační matice.

Determinant matice zjednodušeně definujme jako číslo, které lze spočítat pouze u čtvercové matice. Determinant matice ***A*** se označuje det***A***. Výpočet determinantu se liší podle velikosti matice, nejjednodušší je postup u matic druhého řádu 2 × 2 nebo třetího řádu 3 × 3, s rostoucím řádem složitost výpočtu narůstá. To ale nemusí čtenáře trápit, výpočet determinantu matic je dostupný i v běžných tabulkových procesorech (např. MS Excel [Microsoft, Redmond, WA, USA]) anebo lze využít řady on-line dostupných webových kalkulátorů. Příklady výpočtu pro nejjednodušší matice přibližuje příklad 2.

Mnohem důležitější než samotný výpočet je význam a interpretace hodnoty determinantu korelační matice. Platí totiž, že s rostoucí mírou vzájemné závislosti (korelace) proměnných v matici hodnota determinantu klesá. Při silné vzájemné lineární závislosti analyzovaných proměnných se determinant korelační matice málo liší od nuly (viz ukázky uvedené v příkladu 3). Výpočet korelační matice vyššího řádu a jejího determinantu dokládá příklad 4. Determinant tedy můžeme vnímat jako číselnou prezentaci korelační matice, která ukazuje na míru vzájemné lineární závislosti proměnných, tzv. multikolinearity. Z těchto důvodů je determinant silně využíván v statistické analýze vícerozměrných dat.

Determinant korelační matice využijeme k technickému výkladu výpočtu výše zmíněného vícenásobného koeficientu korelace a parciálního koeficientu korelace. Vztahy a postup výpočtu těchto statistik přibližují příklady 5 a 6.

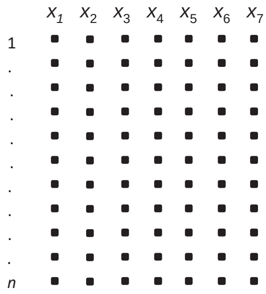
- Mnohonásobný korelační koeficient (příklad 5) vyjadřuje míru závislosti jedné proměnné na dalších proměnných v souboru. Taková analýza je velmi užitečná například při ověřování, zda je či není sledování této proměnné v daném souboru redundantní. Rovněž takto lze posuzovat vysvětlující vliv některých proměnných pro změny hodnot vybrané proměnné apod.
- Parciální korelační koeficient (příklad 6) sleduje v podstatě opačný cíl než koeficient mnohonásobný. Touto korelací hodnotíme vztah dvou spojitých proměnných při vyloučení vlivu ostatních proměnných v souboru. Tato analýza je velmi užitečná, chceme-li odhalit či vyloučit vliv jiných proměnných na míru vztahu dvou separátně sledovaných proměnných v souboru.

Jsme si vědomi, že čtenářům touto snad ještě srozumitelnou formou předkládáme relativně složité statistiky kalkulované na

Mnohonásobný korelační koeficient vyjadřuje míru závislosti jedné proměnné na dalších proměnných v souboru. Jeho rozsah je 0 až 1. Hodnota 1 znamená, že hodnocená proměnná je lineární kombinací ostatních proměnných, hodnota 0 znamená, že hodnocená proměnná není nijak korelována s žádnou z ostatních proměnných.

**Postup výpočtu:**

**Matice *n* pacientů popsaných 7 proměnnými ($X_1 – X_7$)**



**Výpočet Pearsonových korelačních koeficientů**

**Matice korelačních koeficientů 7 × 7. Cílem následné analýzy je kvantifikovat mnohonásobný korelační koeficient proměnné $X_1$ vzhledem k ostatním proměnným**

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|
| 1,0 | –0,4 | 0,3 | 0,0 | 0,5 | 0,3 | 0,0 |
| –0,4 | 1,0 | 0,0 | 0,1 | –0,4 | –0,3 | –0,5 |
| 0,3 | 0,0 | 1,0 | –0,8 | –0,3 | 0,3 | –0,2 |
| 0,0 | 0,1 | –0,8 | 1,0 | 0,3 | –0,4 | –0,3 |
| 0,5 | –0,4 | –0,3 | 0,3 | 1,0 | 0,6 | 0,6 |
| 0,3 | –0,3 | 0,3 | –0,4 | 0,6 | 1,0 | 0,8 |
| 0,0 | –0,5 | –0,2 | –0,3 | 0,6 | 0,8 | 1,0 |

**Vztah pro výpočet mnohonásobného korelačního koeficientu:**

**Výsledek výpočtu:**

$$R_{1(2,..,7)} = \sqrt{1 - \frac{determinant\ matice}{determinant\ matice\ bez\ x_1}} = \sqrt{1 - \frac{-0{,}00372}{-0{,}0115}} = 0{,}822419$$

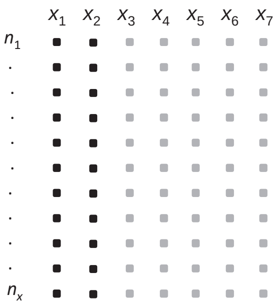
Mnohonásobný korelační koeficient proměnné $X_1$ vzhledem k ostatním zapojeným proměnným v souboru je 0,822. Tato hodnota prokazuje silnou závislost $X_1$ s ostatními proměnnými v souboru.

Příklad 5. Mnohonásobný korelační koeficient.

Parciální korelační koeficient hodnotí vztah dvou spojitých proměnných při vyloučení vlivu ostatních proměnných v souboru. Výpočet je možné provést jednak pomocí determinantu matice dle zde prezentovaného vzorce, jednak prostřednictvím analýzy reziduí odvozených z regresní analýzy, kdy jedna z hodnocených proměnných je nahrazena rezidui regresního modelu hodnotícího vztah ostatních proměnných, jejichž vliv chceme vyloučit, k této proměnné.

**Postup výpočtu:**

Matice ***X*** pacientů popsaná 7 proměnnými, hodnocen vztah $X_1$ a $X_2$ při vyloučení vlivu $X_3$ až $X_7$.



**Výpočet Pearsonova korelačního koeficientu**

Matice korelačních koeficientů 7 × 7 je vstupem do dalších výpočtů založených na výpočtu determinantu.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|
| 1,000 | –0,385 | 0,350 | –0,039 | 0,504 | 0,337 | 0,019 |
| –0,385 | 1,000 | 0,000 | 0,134 | –0,431 | –0,350 | –0,521 |
| 0,350 | 0,000 | 1,000 | –0,748 | –0,283 | 0,283 | –0,157 |
| –0,039 | 0,134 | –0,748 | 1,000 | 0,275 | –0,415 | –0,271 |
| 0,504 | –0,431 | –0,283 | 0,275 | 1,000 | 0,616 | 0,587 |
| 0,337 | –0,350 | 0,283 | –0,415 | 0,616 | 1,000 | 0,756 |
| 0,019 | –0,521 | –0,157 | –0,271 | 0,587 | 0,756 | 1,000 |

**Vzorec pro výpočet parciálního korelačního koeficientu proměnných $X_1$ a $X_2$:**

$$R_{12(3,..,7)} = \frac{(-1)^2 \det(R_{(12)})}{\sqrt{\det(R_{(11)}) \det(R_{(22)})}} = -0{,}392 \qquad \det(R_{(12)}) = determinant\ matice\ bez\ řádku\ 1\ a\ sloupce\ 2$$

Parciální korelační koeficient $x_1$ a $x_2$ má obdobnou hodnotu jako klasický Pearsonův korelační koeficient; mezi proměnnými tedy existuje vztah, který není zprostředkován vlivem jiných proměnných.

Příklad 6. Parciální korelační koeficient.

vnitřně komplikovaných mnohonásobných souborech dat. Avšak příklady 5 a 6 dokládají, že samotný výpočet mnohonásobných a dílčích koeficientů korelace není problém, a pokud uživatel zvládne na počítači výpočet determinantu matice, může tyto korelační koeficienty hodnotit velmi jednoduchými vztahy. Tím se mu otevírají možnosti velmi sofistikovaných analýz s významnou klinickou či biologickou interpretací. Těm bude formou příkladů věnován celý příští díl seriálu.