Analýza dat v neurologii

LXXVII. Korelační analýza vícerozměrných souborů kvantitativních dat – příklady

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

V minulém díle jsme zahájili výklad statistické analýzy více korelačních koeficientů, které můžeme uspořádat do korelační matice. Připomeňme, že jde vždy o matici čtvercovou, která obsahuje vzájemné korelační koeficienty K společně měřených proměnných (X_1, X_2, …, X_K) a na hlavní diagonále obsahuje hodnoty 1. Již tím, že tyto proměnné sledujeme současně v jednom experimentu, dáváme najevo, že jejich vzájemné vztahy jsou podstatné. Čím více takových proměnných do experimentu či klinické studie zařadíme, tím více potenciálních dílčích vztahů můžeme zkoumat. Analýzy vysvětlující různé kombinace vzájemně korelovaných proměnných mají velký interpretační význam a mohou přispět i k objevu nových interakcí různých znaků.

Ambicí tohoto dílu seriálu je formou příkladů přiblížit čtenářům význam těchto analýz a přispět tak k jejich širšímu využívání. Ačkoli vše na první pohled vypadá relativně složitě, jde o výpočty, které jsou dostupné i běžnému uživateli počítačů a ke kterým není třeba exaktní matematické vzdělání. V předchozím výkladu jsme takto představili výpočty mnohonásobného koeficientu korelace a parciálních korelačních koeficientů, k jejichž vyčíslení je třeba pouze schopnost spočítat determinant korelační matice (výpočet byl v minulém díle dokumentován v příkladech 2 a 3). Oba tyto korelační koeficienty jsou typickými představiteli souhrnných koeficientů pracujících s korelacemi více proměnných současně. Obecně je charakterizujeme jako mnohorozměrné ukazatele vzájemné lineární závislosti náhodných veličin. V tomto díle budeme ve výkladu pokračovat a zahájíme jej shrnutím různých typů mnohonásobných korelací.

Zásadní pro kategorizaci koeficientů odvozovaných z korelační matice více proměnných je jejich smysl, resp. interpretační cíl. Typologie těchto koeficientů je užitečná i pro praxi, neboť definuje vlastní záměr vědeckého zpracování dat:

- **Koeficienty vícenásobné** – Kvantifikují lineární vztah mezi jednou vybranou proměnnou a všemi dalšími v experimentu nebo několika dalšími zařazenými v experimentu. Ve skutečnosti je hodnocen lineární

Příklad 1a. Korelační matice vykazující mnohonásobný korelační koeficient pro proměnnou X_1 blízký 1.

	X_1	X_2	X_3	X_4
X_1	1	0,7	–0,2	0,6
X_2	0,7	1	0,5	0,1
X_3	–0,2	0,5	1	–0,3
X_4	0,6	0,1	–0,3	1

$$R_{1(2,3,4)} = \sqrt{1 - \frac{determinant\ matice}{determinant\ matice\ bez\ X_1}} = \sqrt{1 - \frac{0{,}0101}{0{,}62}} = 0{,}9918$$

Příklad 1b. Korelační matice vykazující mnohonásobný korelační koeficient pro proměnnou X_1 blízký 0.

	X_1	X_2	X_3	X_4
X_1	1	–0,01	0	0,01
X_2	–0,01	1	–0,02	0
X_3	0	–0,02	1	–0,01
X_4	0,01	0	–0,01	1

$$R_{1(2,3,4)} = \sqrt{1 - \frac{determinant\ matice}{determinant\ matice\ bez\ X_1}} = \sqrt{1 - \frac{0{,}9993001}{0{,}9995}} = 0{,}0141$$

Příklad 1. Ukázky různých výsledků výpočtu koeficientu mnohonásobné korelace kalkulovaného na korelační matici 4 × 4.

Příklad 2a. Výpočet mnohonásobných korelačních koeficientů.

Výpočet mnohonásobných korelačních koeficientů na matici se shluky vzájemně korelovaných proměnných.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1	–0,7	0,8	0	0	0	0
X_2	–0,7	1	–0,3	0	0	0	0
X_3	0,8	–0,3	1	0	0	0	0
X_4	0	0	0	1	0,2	0,5	0,8
X_5	0	0	0	0,2	1	–0,3	0,3
X_6	0	0	0	0,5	–0,3	1	0,2
X_7	0	0	0	0,8	0,3	0,2	1

Příklad záměrně pracuje s velmi zřetelnými shluky vzájemně korelovaných určitých proměnných ve vstupní korelační matici. Vzájemně korelované proměnné zde vykazují vysoké korelační koeficienty a ostatní korelace neexistují, tedy mají hodnotu 0. Již prostým prohlédnutím korelační matice je tak dána cesta k výpočtu správných mnohonásobných korelací, které dané shluky proměnných odhalí. Data v reálné praxi jsou samozřejmě výrazně komplikovanější a hledání vzájemně korelovaných shluků proměnných vyžaduje výpočet všech možných kombinací proměnných v mnohonásobné korelaci.

Statisticky významné mnohonásobné korelační koeficienty

$R_{X1,X2,X3} = 0{,}9341$
$R_{X2,X1,X3} = 0{,}8233$
$R_{X3,X2,X1} = 0{,}8790$

$R_{X4,X5,X6,X7} = 0{,}8777$
$R_{X5,X4,X6,X7} = 0{,}5069$
$R_{X6,X4,X5,X7} = 0{,}6898$
$R_{X7,X4,X5,X6} = 0{,}8336$

Mnohonásobné korelace identifikující vzájemně korelované shluky proměnných X_1–X_3 a X_4–X_6

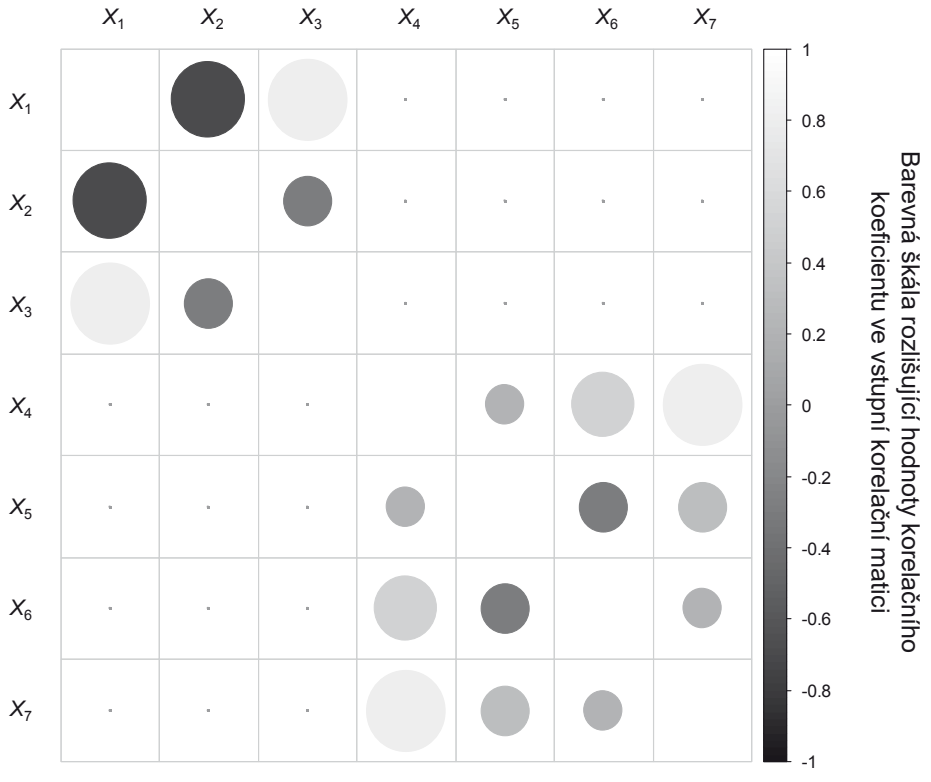
Statisticky nevýznamné mnohonásobné korelační koeficienty

$R_{X1,X4,X5,X6,X7} = 0$
$R_{X2,X4,X5,X6,X7} = 0$
$R_{X3,X4,X5,X6,X7} = 0$

$R_{X4,X1,X2,X3} = 0$
$R_{X5,X1,X2,X3} = 0$
$R_{X6,X1,X2,X3} = 0$
$R_{X7,X1,X2,X3} = 0$

Vzájemně vnitřně korelované shluky proměnných X_1–X_3 a X_4–X_6 mezi sebou nekorelují a jsou tedy nezávislé

Příklad 2b. Grafická dokumentace vstupních dat (vstupní korelační matice) pomocí korelogramu.

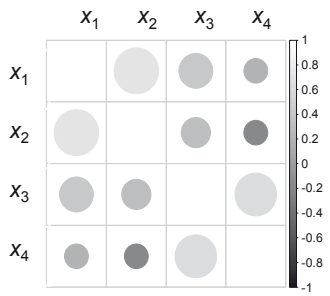


Příklad 2. Využití mnohonásobného koeficientu korelace pro hledání shluků vzájemně korelovaných proměnných.

Příklad 3a.
Parciální korelační analýza potvrzuje hodnotu vstupní korelace dvou proměnných

Vstupní matice korelačních koeficientů 4 × 4

	x_1	x_2	x_3	x_4
x_1	1,0	0,7	0,4	0,2
x_2	0,7	1,0	0,3	–0,2
x_3	0,4	0,3	1,0	0,6
x_4	0,2	–0,2	0,6	1,0



Vstupní korelační koeficient proměnných X_1 a X_2 $R_{X_1X_2} = 0{,}7$

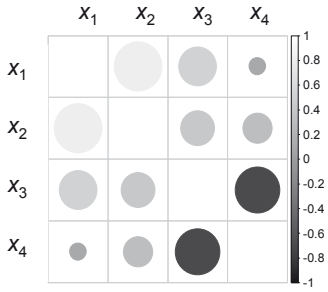
Parciální korelační koeficient proměnných X_1 a X_2 $R_{X_1X_2(X_3,X_4)} = 0{,}73$

V příkladu 3a parciální korelace proměnných X_1 a X_2 s vyloučením dalších proměnných v korelační matici potvrdila velikost vstupního korelačního koeficientu. Lze tedy konstatovat, že proměnné X_3 a X_4 významně neovlivňují korelaci proměnných X_1 a X_2.

Příklad 3b.
Parciální korelační analýza nepotvrzuje hodnotu vstupní korelace dvou proměnných

Vstupní matice korelačních koeficientů 4 × 4

	x_1	x_2	x_3	x_4
x_1	1,0	0,8	0,5	0,1
x_2	0,8	1,0	0,4	0,3
x_3	0,5	0,4	1,0	–0,7
x_4	0,1	0,3	–0,7	1,0



Vstupní korelační koeficient proměnných X_1 a X_2 $R_{X_1X_2} = 0{,}8$

Parciální korelační koeficient proměnných X_1 a X_2 $R_{X_1X_2(X_3,X_4)} = 0{,}35$

V příkladu 3b parciální korelace proměnných X_1 a X_2 s vyloučením dalších proměnných v korelační matici nepotvrdila velikost vstupního korelačního koeficientu. Proměnné X_3 a X_4 významně ovlivňují korelaci proměnných X_1 a X_2 a hodnota původní korelace 0,8 může být důsledkem vzájemné interakce těchto proměnných. Parciální korelační koeficient je významně nižší.

Příklad 3. Ukázky využití dílčích (parciálních) korelačních koeficientů.

vztah mezi vybranou proměnnou a lineárními kombinacemi těch dalších proměnných. Typickým zástupcem je v minulém díle představený mnohonásobný koeficient korelace.

- **Dílčí (parciální) koeficienty** – Cílem těchto ukazatelů je kvantifikovat „čistý" lineární vztah dvou proměnných při vyloučení vlivu všech ostatních nebo vybraných proměnných v experimentu. Vyloučení vlivu znamená, že zkoumáme vztah dvou proměnných při konstantní hodnotě třetí proměnné, nebo více dalších proměnných. Jde o ideální nástroj pro studium maskujících či zkreslujících vzájemných vlivů proměnných a také pro studium skutečně příčinných závislostí.
- **Podmíněné koeficienty** – Při výpočtu těchto koeficientů sledujeme kvantifikaci lineárního vztahu dvou proměnných pouze pro vybrané hodnoty jedné nebo několika dalších proměnných. Jde o velmi významné analýzy, které dokládají, zda je vztah sledovaných znaků nějak podmíněn konkrétními hodnotami znaků jiných. Na takto podmíněném vybraném intervalu hodnot některých proměnných lze rovněž hodnotit mnohonásobnou i parciální korelaci, jak je popsáno výše.

Význam výše uvedených analýz pro klinický výzkum jistě netřeba dále rozsáhle komentovat. Téměř si nelze představit studii či experiment, kde by nějaká forma vzájemného ovlivňování sledovaných proměnných neexistovala. Vzájemné ovlivňování proměnných může být jevem pro výsledky negativním až zavádějícím (např. sledujeme-li vztah mezi dávkou léku a jeho účinkem a tento je ovlivňován „zezadu" faktory jako doba trvání terapie, pravidelnost užívání pacientem, mírou spolupráce pacienta či jinými charakteristikami pacienta nebo nemoci), ale také pozitivním (např. pokud objevíme, že rostoucí účinnost zvyšujících se dávek léku je podmíněna hodnotami některých charakteristik nemoci). Strategický význam těchto analýz ještě zvyšuje prostý fakt, že prakticky nelze uspořádat experiment, který by vyloučil vliv všech proměnných již přímo při měření, zejména pak ne v reálné klinické praxi.

S výše uvedenými koeficienty pracujeme jako s normálními koeficienty mezi dvěma proměnnými. Hodnoty těchto koeficientů blízké 0 jsou nevýznamné. Parciální a podmíněné koeficienty mohou nabývat hodnot od –1 do +1, jako je tomu u běžné korelace. Mnohonásobné koeficienty jsou vždy kladné v rozsahu hodnot od 0 do +1. O interpretaci v podstatě rozhoduje již sám důvod, pro který byly koeficienty počítány. Různé situace přiblížíme v jednotlivých číselných příkladech, kde používáme postup výpočtu pomocí determinantu korelační matice (viz předchozí díl seriálu).

Příklad 1 dokládá různé možné varianty odhadu mnohonásobného koeficientu korelace kalkulovaného na korelační matici čtyř proměnných. Jde typickou situaci, kdy zjišťu-

Zásadní význam parciálních korelací je zejména v odhalování zprostředkovaných korelací, tedy vztahů mezi dvěma proměnnými, které jsou zprostředkovány vlivem jiné či jiných proměnných. V praxi mohou nastat i situace, kdy parciální korelace vede dokonce k opačnému znaménku korelačního koeficientu, než byl koeficient základní (vstupní). Tento příklad takovou situaci dokumentuje na souboru dětí a na korelaci výšky jejich postavy, hmotnosti a vzdáleností, kterou uběhnou za 5 min.

- Vstupní data vedou k základním korelacím, které potvrzují nepřekvapivý silný lineární vztah mezi hmotností dětí a výškou jejich postavy (r_1 **= 0,94**). Vyšší děti rovněž významně dále doběhnou, i když tato korelace mezi výškou a vzdáleností již není tak silná (r_2 **= 0,80**). Avšak pozitivně se vzdáleností uběhnutou za 5 min koreluje i hmotnost (r_3 **= 0,63**), což je jistě překvapivé.
- Pokud bychom v experimentu výšku postavy vůbec nesledovali, došli bychom k závěru, že rostoucí hmotnost dětí zvyšuje jejich výkonnost v běhu. Avšak tato korelace se projevila zejména proto, že soubor dětí byl relativně heterogenní ve výšce postavy a vyšší dětí jsou také zpravidla těžší. Výška postavy tak ovlivňuje, resp. zprostředkovává, korelaci hmotnosti a výkonnosti.
- Pokud vliv výšky postavy odfiltrujeme výpočtem parciální korelace mezi hmotností a vzdáleností, získáváme záporný parciální koeficient. Tedy zcela opačný výsledek, korelace zde dokonce změnila směr. Správný závěr tedy je, že při konstantní výšce postavy s rostoucí hmotností dětí jejich výkonnost v běhu klesá.

x_1 vzdálenost

x_2 hmotnost

x_3 výška

Matice korelačních koeficientů 3 × 3

	x_1	x_2	x_3
	1,0	0,63	0,80
	0,63	1,0	0,94
	0,80	0,94	1,0

Matice parciálních korelačních koeficientů 3 × 3

	x_1	x_2	x_3
	1,0	-0,60	0,78
	-0,60	1,0	0,94
	0,78	0,94	1,0

Příklad 4. Výpočet dílčí (parciální) korelace, která odhalí významně zavádějící vliv třetí proměnné na korelační analýzu.

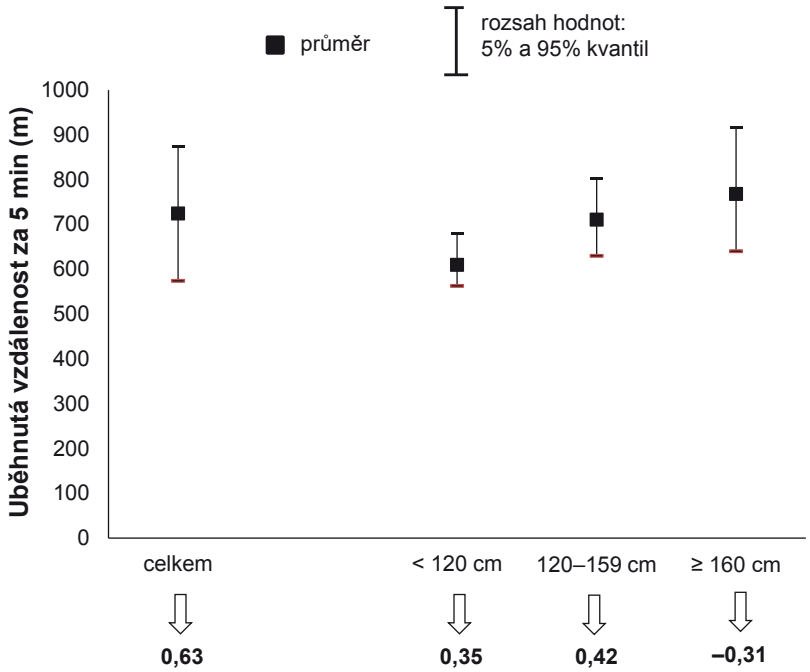
jeme sílu a významnost vztahu mezi zvolenou proměnnou (X_1) a několika dalšími (predikujícími) proměnnými, v našem případě X_2, X_3, X_4. Příklad 1a ukazuje na velmi silný vztah, kdy je proměnná X_1 téměř absolutně korelována s ostatními proměnnými, a jejich kumulativní vliv dovede její hodnoty téměř plně predikovat. V takovém případě lze diskutovat o tom, zda není proměnná X_1 v souboru nadbytečná. Příklad 1b dokládá jinou variantu možného výsledku; proměnná X_1 je zde nezávislá na dalších proměnných v experimentu.

U souborů obsahujících větší množství proměnných bývá častým úkolem prozkoumat vzájemné korelace všech proměnných a určit, zda tyto netvoří vzájemně nezávislé skupiny, které jsou ale uvnitř silně korelované (příklad 2). Nalezení takových skupin vzájemně korelovaných proměnných značně usnadňuje interpretaci mnohorozměrného měření. Pokud jsou takové skupiny znaků mezi sebou vzájemně nekorelované, pak tvoří komplexní dimenze, které mohou mít důležitou interpretaci. Příklad 2 je vymyšlen tak, aby nalezené dva shluky vzájemně korelovaných proměnných byly na první pohled patrné. V reálné praxi a u velkých korelačních matic tomu tak ale nebývá a uvedený kalkulační postup je potom nástrojem velkého významu, neboť může odhalit i skryté vztahy mezi proměnnými.

Z příkladů 1 a 2 je patrné, jak komplikované mohou být vztahy mezi více proměnnými měřenými v jedné studii. V reálném životě se jevy neprojevují izolovaně, vždy na námi sledované znaky působí další vlivy a ty buď v experimentu měříme, a můžeme je tedy podchytit, anebo neměříme, a jejich vliv nám uniká. Příklad 3 dokumentuje využití dílčích, parciálních korelací pro separování „čistého" vztahu mezi dvěma znaky, při vyloučení vlivu dalších proměnných. Výsledek v příkladu 3a ukazuje, že vztah dvou separovaných proměnných není další proměnnou ovlivněn, neboť parciální korelace je přibližně stejná jako základní korelace obou proměnných. Naopak v příkladu 3b bylo prokázáno, že pozorovaná základní korelace mezi dvěma separovanými proměnnými je zprostředkována vlivem dalších proměnných a po odfiltrování jejich vlivu je korelace významně nižší a nevýznamná.

Zásadní význam parciálních korelací spočívá zejména v odhalování zprostředkovaných korelací, tedy vztahů mezi dvěma proměnnými, které jsou zprostředkovány vlivem jiné či jiných proměnných. Učebnicovou ukázkou mohou být různá antropometrická měření, kdy jsou různé míry na postavě člověka ovlivňovány zejména výškou postavy. Představme si např. silnou korelaci mezi délkou dolních a horních končetin, kterou avšak eliminuje parciální analýza, při níž vyloučíme vliv výšky postavy. V praxi mohou nastat situace, kdy parciální korelace vede dokonce k opačnému znaménku korelačního koeficientu, než byl koeficient základní. Takovou situaci dokumentuje příklad 4, který pracuje se souborem dětí a s korelacemi výšky jejich postavy, hmotnosti a vzdáleností, kterou uběhnou za 5 min. Vstupní data příkladu vedou k základním korelacím, které potvrzují nepřekvapivý silný lineární vztah mezi hmotností dětí a výškou jejich postavy, vstupní korelace rovněž potvrzuje, že vyšší děti dále doběhnou. Avšak se vzdáleností uběhnutou za 5 minut zde pozitivně a významně koreluje i hmotnost dětí, což navozuje možný závěr, že rostoucí hmotnost dětí zvyšuje jejich výkonnost v běhu. Avšak tato korelace se projevila zejména proto, že soubor dětí byl velmi heterogenní ve výšce postavy a vyšší dětí jsou také zpravidla těžší. Výška postavy tak ovlivňuje korelaci hmotnosti a výkonnosti. Pokud vliv výšky postavy odfiltrujeme, získáváme záporný parciální koeficient. Tedy opačný výsledek, korelace zde dokonce změnila směr. Správný závěr tedy je, že při konstantní výšce postavy

Data z příkladu 4 využijeme i v příkladu 5, který ukáže výpočet podmíněné korelace. Tuto analýzu lze vnímat jako alternativu k výpočtu parciální korelace z příkladu 4, kde jsme hodnotili „čistou" korelaci mezi hmotností dítěte a jeho výkonností v běhu, a to při odfiltrování vlivu výšky. „Odfiltrováním" se zde myslí sledování korelace mezi hmotností a výkonností při konstantní výšce postavy. Alternativně můžeme hodnotit korelaci mezi dvěma proměnnými pro různé intervaly hodnot třetí (ovlivňující, podmiňující) proměnné. Tak se odstraní její zprostředkující vliv a my uvidíme sledovanou korelaci v rámci kategorií hodnot ovlivňující proměnné.



Korelační koeficient hmotnosti dítěte a jím uběhnuté vzdálenosti pro celý soubor a v rámci různých kategorií výšky dítěte

Korelační koeficient mezi uběhnutou vzdáleností a hmotností dítěte je na celém souboru 0,63, což naznačuje překvapivě se zvyšující výkonnost při rostoucí hmotnosti. Avšak při rozdělení dětí do skupin dle výšky se hodnota tohoto korelačního koeficientu významně snižuje. Ve skupině nejvyšších dětí (nad 160 cm) je korelace mezi hmotností a uběhlou vzdáleností dokonce záporná. Je zřejmé, že celková pozitivní korelace je důsledkem zprostředkovaného vztahu mezi hmotností a uběhlou vzdáleností prostřednictvím výšky dítěte. Celková korelace je z tohoto pohledu silně zavádějící, i když správně numericky kalkulovaná. Na vině je velká heterogenita výšky dětí ve zkoumaném souboru. Analýza korelace dvou proměnných v rámci různých tříd jedné či více jiných proměnných se nazývá podmíněnou korelací.

Příklad 5. Výpočet podmíněné korelace.

s rostoucí hmotností dětí jejich výkonnost v běhu klesá.

Velká rozdílnost zkoumaných dětí ve výšce jejich postavy byla v příkladu 4 příčinnou zavádějící korelace hmotnosti a výkonnosti. To je poměrně častý jev, neboť různorodost (nehomogenita) zkoumané kohorty jedinců, zvláště v takto podstatném parametru, je vždy zdrojem potíží. Předpokládejme, že šlo o přirozenou variabilitu ve výšce postavy u dětí určité relativně úzké věkové kategorie. Avšak pokud by takto byly zařazeny děti výrazně různého věku, pak by šlo o zcela nesprávně postavený experiment, a výpočet korelace mezi hmotností a výkonností by byl zcela zavádějící. Parciální korelace může stejným způsobem jako v příkladu 4 odhalit také vliv jedné nebo několika odlehlých hodnot, které generují klamný obraz korelovaných proměnných.

Data z příkladu 4 jsme dále využili v příkladu 5, který dokládá výpočet podmíněné korelace. Tuto analýzu lze vnímat jako alternativu k výpočtu parciální korelace z příkladu 4. Výpočtem parciální korelace v příkladu 4 jsme počítali „čistou" korelaci mezi hmotností dítěte a jeho výkonností v běhu, a to při odfiltrování vlivu výšky. Alternativně můžeme hodnotit korelaci mezi dvěma proměnnými, a to pro různé intervaly hodnot třetí, ovlivňující, proměnné. Tak se odstraní její zprostředkující vliv a my uvidíme sledovanou korelaci v rámci tříd hodnot ovlivňující proměnné.

V příkladu 4 i 5 jsme takto odhalili velmi zavádějící vliv třetí proměnné na studovanou korelaci, přičemž šlo o příklady jednoduché, jejichž výsledek bylo možné uhodnout předem. Pokud si ale představíme ve skutečné studii pole několika desítek proměnných, z nichž některé mají i neznámou interpretaci (např. nové molekulárně biologické markery apod.), pak je takováto analýza průzkumem v prostředí, kde výsledek předem nemáme šanci uhodnout. Dopad na interpretaci konečného výsledku může být zásadní.

Snad jsme zde vypracovanými příklady přispěli k propagaci rozborů korelací více proměnných. Interpretační přínos těchto analýz je zřejmý a silně přispívá i ke studiu kauzálních vztahů. Jak jsme již rozebírali dříve (díl LIX seriálu), samotné statistické prokázání vztahu, např. korelací, neznamená průkaz kauzality. Pokud ale vztah dvou proměnných potvrdíme i parciálními korelacemi s vyloučením jiných ovlivňujících proměnných, jde o krok, který průkaz kauzality přibližuje.