Analýza dat v neurologii

LXXV. Příklady chybné korelační analýzy

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková

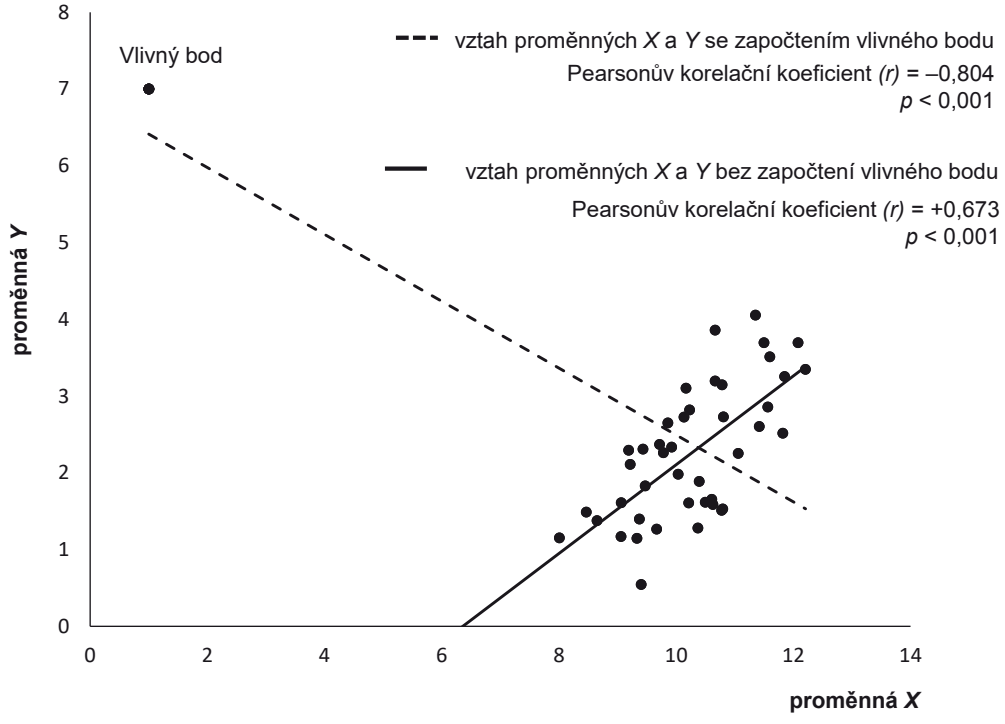
Institut biostatistiky a analýz,
LF MU, Brno

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

Tento díl seriálu věnujeme reakci na podnět jednoho z čtenářů. V předchozích dílech jsme u analýzy korelace často varovali před chybným hodnocením či interpretací korelace. Zdůrazňovali jsme, že průkaz korelace sám o sobě není důkazem příčinné závislosti. A naopak, že nekorelovanost neznamená nezávislost, neboť vztah dvou proměnných může být komplikovaný, nelineární, a korelační koeficient ho nemusí vždy podchytit. Avšak nejzávažnější chyby vznikají, pokud je hodnota Pearsonova koeficientu korelace odhadována na datech, která nejsou pro výpočet této parametrické lineární korelace vhodná. Právě na tyto situace mířil dotaz čtenáře, který se ptal, zda může v určitých situacích dostat při výpočtu korelace zcela opačný výsledek než je realita naměřená v datech. Odpověď na tuto otázku je bohužel kladná a tyto situace se zde pokusíme ukázat na třech modelových příkladech.

Velkou nevýhodou Pearsonovy korelace je totiž její vysoká citlivost na odchylky od normálního rozdělení korelovaných proměnných. Skutečně, jediná odlehlá hodnota může doslova otočit výsledek analýzy a místo reálně existující kladné korelace vypočítáme korelaci zápornou. A aby toho nebylo málo, tak při dostatečně velkém vzorku dat vyjde tento zcela nesprávný výsledek jako statisticky významný. Příklad 1 ukazuje přesně takovou situaci. Dopad jedné odlehlé hodnoty, které se také někdy říká vlivný bod, je zásadní a zcela mění výsledek analýzy. Přitom taková hodnota může být i výsledkem nesprávného opsání naměřených dat do ta-

Příklad znázorňuje soubor dat dvou proměnných, ve kterém je viditelná jedna extrémně odlehlá hodnota. Pokud bychom tento tzv. vlivný bod zahrnuli do korelační analýzy, získali bychom statisticky významný záporný korelační koeficient. Pokud odlehlou hodnotu z výpočtu vyloučíme, existuje mezi proměnnými *X* a *Y* statisticky významná, avšak kladná korelace ($r = 0{,}673$). Je-li takový vlivný bod výsledkem chybného měření či chybného zaznamenání dat, pak je výpočet Pearsonovy korelace na celém souboru nesprávný. Odlehlá hodnota nadto porušuje předpoklad normálního rozdělení hodnot *X* a *Y*, který musí být u parametrické korelace vždy splněn.



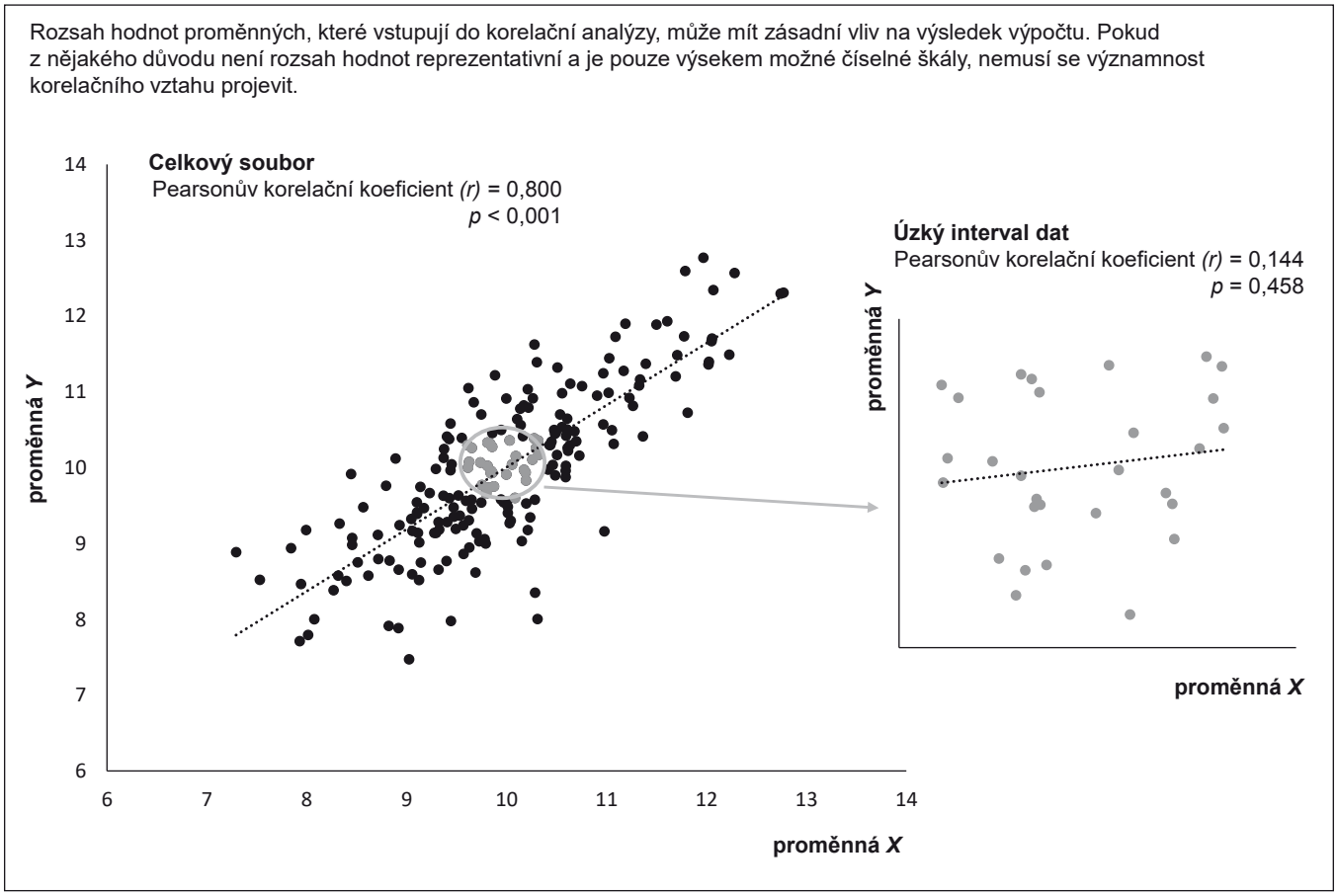
Příklad 1. Ukázka dopadu tzv. vlivného bodu na výpočet Pearsonova korelačního koeficientu.

bulky. Ve velkém souboru si autor analýzy nemusí odlehlé hodnoty mezi mnoha čísly všimnout. Proto je zásadní před korelační analýzou vždy ověřit normalitu rozložení proměnných *X* a *Y*. Rovněž je nutné prohlédnout si vztah proměnných v grafu.

K příkladu 1 je nutné doplnit jednu velmi podstatnou poznámku. Oba prezentované výsledky, tedy výpočet korelace se započítáním anebo naopak s vyloučením odlehlé hodnoty, jsou numericky proveditelné a čistě jako výpočet je lze matematicky provést. V tom právě spočívá záludnost vlivu odlehlých měření či chyb v datech. Software provede jakýkoli výpočet, který mu je zadán, a pro nezkušeného zpracovatele může statistická významnost hodnoty korelačního koeficientu vypadat jako potvrzení správnosti výsledku. Pro konečnou interpretaci výsledku je ovšem taková chyba naprosto fatální. Nelze se tedy divit editorům významných vědeckých časopisů, že si někdy žádají zdrojová data publikujících týmů, zejména jsou-li prezentované výsledky z nějakého pohledu překvapivé či nečekané.

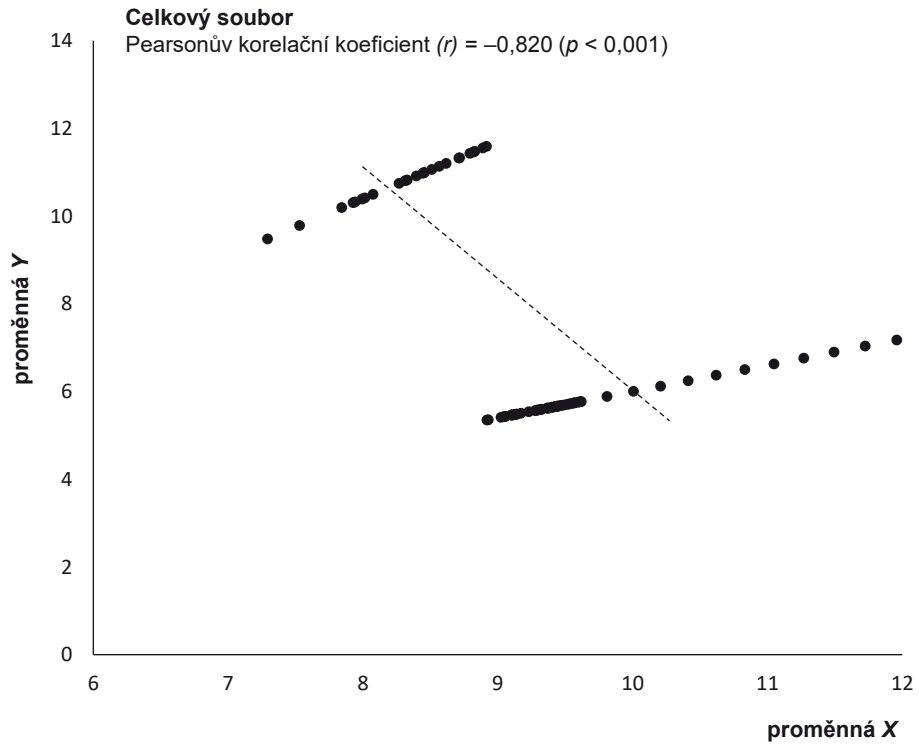
Příklad 2 ukazuje na další z možných komplikací korelační analýzy, tentokrát citlivost odhadu korelačního koeficientu k číselnému rozsahu korelovaných hodnot proměnných *X* a *Y*. Je jistě žádoucí, aby do korelační analýzy proměnné vstupovaly s reprezentativní škálou svých číselných hodnot. Pokud z nějakého důvodu zúžíme analýzu na omezený interval možných hodnot *X* a *Y*, nemusí vztah proměnných na tomto intervalu odpovídat vztahu na celé škále možných hodnot. Zde samozřejmě nemůžeme paušálně mluvit o chybě. Pokud je omezení analyzovaných hodnot řádně zdůvodněno a popsáno, pak je takový postup jistě legitimní. Problém nastává ve chvíli, kdy je analyzován nereprezentativní rozsah hodnot například v důsledku zkresleného výběru vzorku k analýze. V takovém případě nemá hodnota korelačního koeficientu smysluplnou interpretaci.

Poslední příklad ukazuje poněkud extrémní situaci, kdy mezi proměnnými *X* a *Y* existují různé (dílčí) vztahy, například v závislosti na hodnotách jedné z proměnných. Na grafu v příkladu 3 vidíme dvě jasné přímkové závislosti mezi *X* a *Y*, obě s různými sklony. Hodnoty *Y* větší než 8 vykazují vůči *X* jinou závislost než hodnoty menší než 8. Celková záporná korelace počítaná na všech hodnotách spojených dohromady nemá žádnou smysluplnou interpretaci a fakticky jen maskuje skutečný průběh vztahu *X* a *Y*. Příklad pracuje s modelovými daty, která bychom v takto jasné podobě asi v reálné klinické praxi nenaměřili. V obou kategoriích hodnot *Y* je vztah s *X* čistě přímkový, tedy bez rozptylu hodnot, a připomíná tak spíše kalibrační měření v laboratoři. Avšak rozdílný tvar závislosti proměnných při různých hodnotách jedné nebo obou z nich není v přírodě nic výjimečného. Učebnicovým příkladem může být stimulace odpovědi nějakého biologického systému při rostoucích dávkách nějaké látky, např. mikroživiny. Od určité úrovně dávky se ale růst odpovědi systému zastaví anebo může začít klesat, protože vysoké koncentrace látky budou naopak působit toxicky. V takových případech je nutné korelační analýzu provádět odděleně pro různé kategorie hodnot proměnných. Spojení dat do jednoho souboru nepovede k relevantnímu popisu takto složitých vztahů.



Příklad 2. Vliv hodnoceného intervalu dat na výpočet Pearsonova korelačního koeficientu.

Graf znázorňuje extrémní situaci, kdy mezi proměnnými *X* a *Y* existují velmi silné vztahy v závislosti na hodnotách jedné z proměnných. Je zřejmé, že pro hodnoty *Y* větší a menší než 8 existují silné lineární závislosti s proměnnou *X*, každá s jiným sklonem. Taková situace by v praxi samozřejmě vyžadovala další detailní vysvětlení a studium. Je však jisté, že aplikace korelační analýzy na celý soubor je nesmyslná a vede k zavádějící záporné, vysoce významné, hodnotě korelačního koeficientu.



Příklad 3. Ukázka komplikované závislosti proměnných *X* a *Y*.

Závěrem lze shrnout, že korelační analýza je skutečně velmi citlivá na podobu analyzovaných dat. Neopatrným postupem můžeme snadno dospět k velmi zavádějícímu výsledku. Naštěstí máme k dispozici hned několik postupů, jak můžeme velmi snadno a bez složitých matematických postupů nesprávnému výsledku zabránit. Zmiňme se zejména o následujících třech:

- **Grafické znázornění vztahu *X* a *Y*.** Tento postup není jistě třeba květnatě zdůvodňovat. Korelační analýzu by vždy měla doprovázet grafická vizualizace vztahu obou proměnných. Grafy samozřejmě není nutné vždy publikovat, ale jako pracovní nástroj odhalující většinu potenciálních problémů v datech jsou nepostradatelné.
- **Současný výpočet parametrické (Pearsonovy) a neparametrické (Spearmanovy) korelace.** Ačkoli to na první pohled nevypadá koncepčně, současný výpočet těmito dvěma postupy není nic špatného. Není-li v datech nějaký závažný problém, odchylky od normality, odlehlé hodnoty apod., měly by oba korelační koeficienty vyjít přibližně stejné. Jako vážné varování je třeba vnímat výsledek, kdy se tyto typy koeficientů numericky zásadně liší anebo dokonce jeden vyjde kladný a druhý záporný.
- **Ověření vlivu odlehlých hodnot či skupin odlehlých hodnot.** Výpočet korelace není při současném výkonu výpočetní techniky nijak zatěžující, a lze jej tedy opakovat s vyloučením podezřelých či odlehlých hodnot. Pokud se vyloučením jediného bodu výsledek korelace zásadně změní, je třeba tyto hodnoty dále ověřit. Je jistě správné, aby výsledek výpočtu nebyl závislý na jediné hodnotě v datovém souboru. Podobně lze přistupovat i ke skupině hodnot.