Korelace a kauzalita: polemika s jednou reakcí na blog „Corona matematicky i lidsky“ od Karla Janečka

Arnošt Komárek

Matematicko-fyzikální fakulta Univerzity Karlovy, Katedra pravděpodobnosti a matematické statistiky

Jedna část blogu „Corona matematicky i lidsky“ od Karla Janečka (https://blog.aktualne.cz/blogy/karel-janecek.php?itemid=39249#more), který jsem si v nedávné době se zájmem přečetl, udává jako příklad možného zneužití vědy použití „Kulveitova modelu“ (Brauner et al., *Science*, 2021, doi: 10.1126/science.abd9338). Karel Janeček uvádí:

> *„Například Kulveitův model pro hodnocení účinků restriktivních opatření (Brauner et al., 2021) je velmi nespolehlivý a může přinášet velmi nepřesné a zavádějící výsledky z důvodu multikolinearity jednotlivých faktorů. Model navíc implicitně a neoprávněně předpokládá, že veškeré pozorované efekty jsou jen díky vládním opatřením, což může způsobit chybu záměny kauzality a korelace.“*

Jeden z autorů zmíněného modelu, Jan Kulveit, následně zareagoval na svém FB profilu následovně:

> *„Nerad bych, aby se opakované nepravdy staly součástí common knowledge, takže mi přijde potřeba na to reagovat.*
>
> *Takže ne, uvedené není pravda. Technicky jde o to, že něco model vysvětluje jako efekt opatření, něco pomocí tzv. „noise terms“. Podrobně je to v appendixu článku, rovnice A5 A6 a sekce B3 a v separátním článku (https://arxiv.org/abs/2007.13454). Jde o relativně běžný postup Bayesovského modelování popsaný např. ve standardní učebnici oboru "Bayesian Data Analysis", A Gelman et al.*
>
> *Tvrzení Karla Janečka je nepravdivé na úrovni matematiky, podobně jako třeba není pravda že* $2 + 2 = 21$. *Kdokoli uvedené tvrdí buď použitému modelu nerozumí, nebo se v úvahách spletl, a nebo věci rozumí, nespletl, a neříká pravdu. Je přitom celkem jedno jestli jde o tvrzení nositele Fieldsovy medaile, Nobelovy ceny, guru originální mystické školy, předsedy odborné společnosti, myče oken nebo náhodného komentátora.“*

Považuji za nutné upozornit, že „opakovaná nepravda“ rozhodně není nepravdou a hluboce se mýlí naopak Jan Kulveit, který dle mého názoru naprosto netuší, jaký je rozdíl mezi kauzalitou a korelací (na což upozorňuje Karel Janeček). Ke správné *interpretaci* výsledků jakékoliv statistické analýzy (nebo modelu) je totiž zapotřebí disponovat nejenom „správným“ modelem (resp. dle terminologie G. E. P. Boxe modelem, který je sice nejenom špatný, neboť všechny modely jsou špatné, ale též užitečný), ale je též nutné vědět, jakým způsobem byla získána data, pomocí nichž je model odhadován a tomuto následně přizpůsobit interpretace. Ano, „Kulveitův model“ obsahuje „noise terms“. Zde bych připomenul, že zahrnutí „chybových členů“ pokrývajících jak nevysvětlené efekty, tak třeba chyby měření (např. v případě hodnocení fyzikálních experimentů) je zcela běžné pro prakticky všechny statistické modely, nejenom Bayesovské. Nebudu se na tomto místě nikterak vyjadřovat k tomu, zda je „Kulveitův model“ pouze špatný nebo špatný, ale též užitečný, neboť pro potřeby této krátké polemiky je to úplně jedno. Hlavním problémem je totiž určení toho, k čemu daný model může být potenciálně užitečný. S ohledem na fakt, že model je odhadován na základě „observačních“ dat, tj. dat, jejichž sběr není žádným způsobem řízen, resp. „pod kontrolou“ (jako by tomu bylo v případě fyzikálního experimentu nebo řádné klinické studie), a s ohledem na fakt, že model

explicitně uvažuje jako možné efekty pouze efekty „vládních“ opatření (*non-pharmaceutical interventions*), nelze z modelu v žádném případě usuzovat na kauzální (*příčinné*) efekty „vládních“ opatření na cokoliv! Jsou-li k dispozici pouze observační data, je pro účely zjišťování kauzálních vlivů zapotřebí (kromě mnoha dalších kroků) zahrnout do modelu *všechny* faktory, jež mohou potenciálně ovlivňovat odezvu (zde počty nakažených či zemřelých) a jež mohou nějakým způsobem souviset s faktory zájmu („vládními“ opatřeními). Na tomto základním stavebním kamenu statistické inference, srovnatelnému s tvrzením, že $2 + 2 = 4$) nic nemění to, že v modelu jsou zahrnuty „noise/error/. . . terms“. Podrobně se této problematice věnuje celá jedna oblast statistiky zvaná „kauzální analýza“ (*causal analysis*).

V souhrnu, z „Kulveitova modelu“ (a mnohých jiných založených na „observačních“ datech bez pořádně provedené kauzální analýzy) lze možná usuzovat na korelace (asociace), v žádném případě však na kauzality. Vzhledem k tomu, že ani typický politik nechápe rozdíl mezi kauzalitou a korelací, je základní povinností vědce na tato omezení upozorňovat a to opakovaně. Z vývoje posledních měsíců se totiž zdá, že celá řada „modelářů“ nejenom že na tento rozdíl neupozorňuje, ale dokonce aktivně podsouvá kauzální interpretace na místa, kde jsou zcela neopodstatněná, srovnatelná s věštbou z křišťálové koule a v důsledku nebezpečná. Pro ilustraci uveďme jeden aktuální příklad. Mezi okresy ČR s nejnižší incidencí nákazy SARS-CoV-2 nyní patří tři nejdéle „uzavřené“ okresy Trutnov, Cheb a Sokolov. Vhodný model by tedy patrně ukázal (statisticky) významný vliv uzávěry okresů na incidenci nákazy. Odsud lze bezpochyby usuzovat, že uzávěra okresů *koreluje* se snížením incidence nákazy. Lze ale též usuzovat na *příčinný* vliv uzávěry okresů? Nikoliv, resp. rozhodně ne na základě analýzy, jež jiné vlivy vůbec neuvažuje (bez ohledu na to, jaké „noise terms“ v modelu jsou). *Možná* příčinná trajektorie by zde totiž *mohla*[1] být např. následující: (1) vysoká incidence nákazy *způsobuje* vyhlášení uzávěry okresů, ale současně též vysoká incidence *způsobuje* zrychlené promoření (a imunizaci), (2) zrychlené promoření (které ale probíhá spolu s uzávěrou okresů) *způsobuje* pozdější nízkou incidenci (díky vyšší imunizaci obyvatelstva předmětných oblastí). Vzhledem k tomu, že zrychlené promoření jde ruku v ruce spolu s uzávěrou okresů, nalézáme následně (pomocí „modelu“) též korelaci nízké incidence s uzávěrou okresů. Bez hlubší analýzy nelze nicméně v žádném případě tvrdit, že uzávěra okresů *způsobila* nízkou incidenci. Vzhledem k tomu, že velká část „vládních“ opatření (nejenom v České republice) pouze reaguje na (do jisté míry přirozený) vývoj epidemie (tedy opatření jsou spíše *zapříčiněna* vývojem epidemie a nikoliv naopak), je značně nezodpovědné na základě běžně dostupných dat podsouvat komukoliv (a politikům neznalým souvislostí tím spíš) *kauzální* interpretace z modelů, jejichž jedinou ambicí může být zjistit, jak spolu jednotlivé faktory *korelují*. Naopak tam, kde byl v nějaké míře přirozenou formou simulován „experiment“ mající za cíl srovnat efektivitu dvou druhů léčby (Švédsko vs. zbytek Evropy, Florida vs. Kalifornie, dvě oblasti v Dánsku) a kde tedy lze usuzovat i na *kauzality* se ukazuje, že efekt (minimálně některých) „vládních“ opatření na vývoj epidemie je přinejmenším sporný. Současně je zde neoddiskutovatelný *příčinný* devastující efekt těchto opatření na nejrůznější části ekonomiky i společenského života. I nadále nezodpovězenou otázkou tedy zůstává, proč je i nadále používána „léčba“, jejíž škodlivé vedlejší efekty jsou všem dobře známy, ale skutečný efekt ani po roce nebyl řádně prokázán. Pravdou zůstává, že taková léčba pouštění žilou byla po staletí používána bez řádného prokázání jejího efektu na stav pacienta. . .

Autor je docentem statistiky na MFF UK, pravděpodobnost a statistiku tamtéž vystudoval (Mgr.) v roce 2000. Následně (2001) získal titul MSc. v oboru biostatistika na *Universiteit Hasselt* v Belgii a Ph.D. (2006) na *Leuven Biostatistics and statistical Bioinformatics Centre Katholieke Universiteit Leuven* v Belgii. Autor je bývalým předsedou mezinárodní *Statistical Modelling Society* a editorem odborného časopisu *Statistical Modelling* (nadmediánový IF v sekci *Statistics and Probability*). Autor publikoval 19 metodologických a 36 aplikačních odborných článků v časopisech s IF, jež byly více jak 1500-krát citovány a dosáhl H-indexu 20.

[1] Nikterak netvrdím, že tomu tak je, ale bez řádné analýzy toto vyloučit rozhodně nelze.